Proceedings of the 11th International Conference
on Electronic Publishing

# **Openness in Digital Publishing:** Awareness, Discovery and Access

Organised by the Vienna University of Technology (Vienna, Austria)

Vienna
June 13-15, 2007

Editors:

**Leslie Chan**
University of Toronto at Scarborough (Canada)

**Bob Martens**
Vienna University of Technology (Austria)

IRIS-ISIS Publications, 2007

**Proceedings of the 11th International Conference on Electronic Publishing Vienna, 2007**

Vienna University of Technology (Vienna, Austria)
*http://www.tuwien.ac.at*

Edited by:
Leslie Chan, University of Toronto at Scarborough (Canada)
Bob Martens, Vienna University of Technology (Austria)

Published by:
ÖKK-Editions, Vienna

ISBN: 978-3-85437-292-9
First edition

Disclaimer
Any views or opinions expressed in any of the papers in this collection are those of their respective authors. They do not represent the view or opinions of the Vienna University of Technology, University of Toronto at Scarborough, the editors and members of the Programme Committee, nor of the publisher IRIS-ISIS Publications and conference sponsors.

Any products or services that are referred to in this book may be either trademarks and/or registered trademarks of their respective owners. The Publisher, editors and authors make no claim to those trademarks.

# Members of the ELPUB2007 Programme Committee

Apps, Ann, University of Manchester (UK)
Baptista, Ana Alice, University of Minho (Portugal)
Bentum, Maarten van, University of Twente (The Netherlands)
Borbinha, Jose Luis, INEC-ID (Portugal)
**Chan, Leslie, University of Toronto at Scarborough (Canada)**
Cetto, Ana Maria, IAEA (Austria)
Cooper, Graham, University of Salford (UK)
Costa, Sely M.S., University of Brasilia (Brazil)
Delgado, Jaime, Universitat Pompeu Fabra (Spain)
Diocaretz, Myriam, MDD Consultancy (The Netherlands)
Dobreva, Milena, Inst. of Mathematics and Informatics, Acad. Sciences (Bulgaria)
Engelen, Jan, Katholieke Universiteit Leuven (Belgium)
Gradmann, Stefan, University of Hamburg (Germany)
Guentner, Georg, Salzburg Research (Austria)
Hedlund, Turid, Swedish School of Economics and Business Administration, Helsinki (Finland)
Heinrich, Klausjuergen, Donau-Universitaet Krems (Austria)
Ikonomov, Nikola, Institute for Bulgarian Language (Bulgaria)
Iyengar, Arun, IBM Research (USA)
Jezek, Karel, University of West Bohemia in Pilsen (Czech Republic)
Knoll, Adolf, Czech National Library (Czech Republic)
Krottmaier, Harald, Graz University of Technology (Austria)
Kreulich, Klaus, Munich University of Applied Sciences (Germany)
Linde, Peter, Blekinge Institute of Technology (Sweden)
Martens, Bob, Vienna University of Technology (Austria) – CHAIR
Moens, Marie-Francine, Katholieke Universiteit Leuven (Belgium)
Moore, Gale, University of Toronto (Canada)
Mornati, Susanna, CILEA (Italy)
Nisheva-Pavlova, Maria, Sofia University (Bulgaria)
Paepen, Bert, Katholieke Universiteit Leuven (Belgium)
Perantonis, Stavros, NCSR - Demokritos (Greece)
Savenije, Bas, Utrecht University Library (The Netherlands)
Schranz, Markus, Pressetext Austria (Austria)
Smith, John, University of Kent at Canterbury (UK)
Tonta, Yaşar, Hacettepe University (Turkey)

## Keynotes

## Open Access and New Publishing Models

## Emerging Business Models for e-Content Delivery

## Enabling Accessibility

## Development of Repository and Publishing Tools

## Assessment of Open Access and Enabling Frameworks

## Open Access Journals and Regional Perspectives

## Ontology and Meaning Extraction

## Digital Heritage and Access

## Impact Analysis and Open Access Journals

## Web 2.0 and Social Media

## Interoperability and Open Repositories

## Workshop 1

## Workshop 2

## Workshop 3

## Demonstrations

## Posters

# Preface

Dear readers and delegates at ELPUB 2007,

It is a pleasure for us to present you with this volume of proceedings, consisting of scientific contributions accepted for presentation at the 11the ELPUB conference, organised by the Vienna University of Technology, Austria.

The 11th ELPUB conference keeps alive the mission of the ten previous international conferences on electronic publishing - held in the United Kingdom (in 1997 and 2001), Hungary (1998), Sweden (1999), Russia (2000), the Czech Republic (2002), Portugal (2003), Brazil (2004) Belgium (2005) and Bulgaria (2006) - which is to bring together researchers, lecturers, developers, entrepreneurs, managers, users and all those interested in issues regarding Electronic Publishing in widely differing contexts.

The theme for this year's conference, "Openness in Digital Publishing: Awareness, Discovery, and Access", is devoted to exploring the full spectrum of "openness" in digital publishing, from open source applications for content creation to open distribution of content, and open standards to facilitate sharing and open access to scientific publications. In addition to technical papers, we also encouraged submissions reporting on research on economics of openness, public policy implications, and institutional support and collaboration on digital publishing and knowledge dissemination. The goal is to encourage research and dialogues on the changing nature of scholarly communications enabled by open peer-to-peer production and new modes of sharing and creating knowledge.

In order to guarantee the high quality of papers presented at ELPUB 2007, all submissions were peer–reviewed by the Programme Committee (PC), whose thirty-four highly qualified and specialised experts represent many different countries and cover a wide variety of institutional and knowledge domains. The PC did a great job in selecting the best submissions for ELPUB 2007, and we would like to thank them sincerely for the valuable time and expertise they put into the peer review process.

At the conclusion of the peer review procedure, all selected and confirmed entries for this conference, including full papers for scientific presentations, and shorter papers for workshops and demonstrations, were pre–published in the ELPUB Digital Library at http://elpub.scix.net. Potential delegates could therefore see, in advance, what could be expected at the meeting. The same system – SOPS, or SciX Open Publishing System – was also used to set up the submission and review of abstracts.

To assist with the assignment of reviewers, submitters were asked to characterise their entries by selecting 3–5 key areas out of a larger list of subject descriptors. In a similar way, reviewers identified their 3–5 fields of expertise and this allowed the Programme team to map papers to reviewers. Finally, the same system supported the Programme Committee with the scheduling of the sessions and with grouping papers according to common and over-lapping themes. The Table of Content of this volume follows both the themes and the order of the sessions in which they were scheduled during the conference.

As with all previous ELPUB conferences, the collection of papers and their metadata are made available through several channels of the Open Archives Initiative, including Dublin Core metadata distribution and full archives at http://elpub.scix.net. It may appear ironic to have a printed proceedings for a conference dedicated to Electronic Publishing. However, the "need" for printed publications is an old and continuing one. It seems that it is still essential for a significant number of delegates to have "something tangible" in their hands and their respective university administrations.

We hope you enjoy reading the proceedings. It is also our pleasure to invite delegates and readers to ELPUB 2008, taking place in Toronto, Canada. The 12th ELPUB conference will be organised by the University of Toronto, and this marks the first time the conference series will be held in North America. Details of the conference will be forthcoming at the ELPUB web site.

Finally, we would like to thank our Keynote Speakers, Keith Jeffery and Norbert Kroó, for their insightful and timely contributions to the conference. Thanks also go to Grace Samuels for checking the references, to Tomo

Cerovsek for maintaining the submission- and review-interface. We would also like to thank the sponsors for their generous contributions.

With our best wishes for a very successful conference,

| | |
|---|---|
| Leslie Chan<br>Programme Committee Chair<br>University of Toronto Scarborough | Bob Martens<br>General Chair<br>Vienna University of Technology |

# Technical Infrastructure and Policy Framework for Maximising the Benefits from Research Output

*Keith G. Jeffery*

Science and Technology Facilities Council, Rutherford Appleton Laboratory, OX11 0QX UK
e-mail: keith.g.jeffery@rl.ac.uk

## Abstract

Electronic publishing is one part of a much larger process. There is a research lifecycle from creation of a programme for funded research through research proposals, projects, outputs (including publications), exploitation (both for further scholarly work and for commercial or quality of life benefits) and creation of the next programme. Throughout this lifecycle information is the lifeblood; publications are used and created at all stages. The vision proposed brings together electronic research publications with associated datasets and software all contextualised by a CRIS (Current Research Information System) which provides information on projects, persons, organisational units, outputs (products, patents, publications), events, facilities, equipment and much more. Via the CRIS, research output can be linked to financial, project management and human resource data: indeed finally the cost of production of a publication can be compared against its benefit. Realising the vision requires advanced IT architectures including GRIDs and ambient computing. Against this vision current debates about subscription-based publishing and gold author-pays open access publishing, about grey literature and green open access self-archiving can be regarded with clarity and objectivity. The way ahead is clear: funders of research should mandate green self-archiving for the benefit of research and of the twin beneficial consequences: wealth creation and improvement in the quality of life. These benefits far outweigh any short-term benefits from the publishing industry in profits or tax-take. There is still plenty of market opportunity for publishers and their doomsday predictions are unsustainable.

**Keywords:** open access; CRIS; e-infrastructure; repository

# 1 Maximising the Benefits from Research Output

## 1.1 The Requirement

### 1.1.1 Introduction

Let us start with that which is required. This is detailed below by type of user (actor) and role but we can surely agree that the overall aim must be that research output causes wealth creation and / or improvement in the quality of life. It follows therefore that maximising these desirable properties requires maximum access to research output. Provision of maximum access has technical, legalistic and economic implications. It also requires a broader context to ensure the research output material is understood and used appropriately.

### 1.1.2 The Actors

The researcher requires access to find relevant pre-existing research output and to find possible research collaborators. The research manager requires access to check completeness of recorded outputs from her institution, to compare with that of other institutions and thus to develop strategy for her institution. The funding agency requires access to ensure defined outputs from the funded research proposal are delivered, to compare outputs with those from other funding agencies and to find appropriate referees. The policymaker requires access to compare outputs produced by different continents, countries, institutions and research teams. The innovator requires access to find new ideas which are exploitable for wealth creation or improvement in the quality of life. The educator requires access to obtain teaching material. The student requires access to use learning material. The media require access to obtain information that can be recast as ‘stories’ which popularise research or raise social, ethical, political or economic issues concerning the research for the public interest.

### 1.1.3 The Roles

Any competent researcher before starting a new research idea will review the existing research output. The more complete and accessible this is to her, the better the review will be, nugatory effort will be avoided and a better (novel) idea will be formulated. A researcher working in one topic area may find an applicable and appropriate technique –such as an experimental protocol, or a computer program for simulation or statistical reduction - from another topic area. As a result of one of the above, or by an independent search, a researcher may find a potential collaborator or complementary co-worker for a research idea.

One measure of a researcher capability is evaluation of produced output. The more complete and accessible outputs are, the better the quality of the evaluation. The metrics imposed on the raw data (i.e. how one ranks different publication channels such as journals) are a separate issue, but without complete and verifiable raw data evaluations are worthless. Similarly the performance of an organisational unit can be evaluated based on its outputs. Indeed, one could compare inputs (funding) with outputs as evaluated to obtain some idea of effectiveness and efficiency.

One may wish to evaluate the literature in different topic areas of fields of research. This may inform strategic decisions on research funding, or areas of priority in a research institution. The literature provides a source of ideas, usually with associated research to demonstrate their potential use. This is a mine of information for the entrepreneur or innovator who wishes to invest venture capital to create products or services with associated wealth creation (jobs, profits for shareholders).

Today's teaching material is the research output of years ago. As the pace of learning increases, and the volume of research output increases, there is a need for faster and easier access to appropriate research literature by educators. Modern learning is more project-based and less 'chalk and talk'. Students are encouraged to utilise technology to find relevant information.

Journalists and other media professionals need easy access to research outputs in order to find interesting 'stories' for popularising, to research (verify) the background to 'urban myths' about research and to find researchers suitable for appearing on TV programmes or writing articles.

### 1.1.4 Conclusion

We can conclude that all these actors, in the various example roles discussed, require easy (fast, efficient) access to research output material. Technically this implies the need for excellent descriptive metadata, fast searching of metadata, fast searching of text and multimedia and well-structured results. Furthermore access to heterogeneous distributed repositories should appear homogeneous and local to the end-user. This implies reconciliation to a canonical syntax (structure) and semantics (meaning) which in turn is likely to involve translation of character sets, language and ontological terms. Legalistically it requires unfettered access although restrictive metadata may document - for software to enforce - claimed rights which should be respected (like attribution) and even may define a price for access. Economically it requires a business model where costs are minimised (ideally zero as seen by the end-user), any income lies where the work is done and costs are borne where benefit is obtained. Furthermore, ideally the actors require the research output material in the context of research project, researchers, organizations involved, facilities and equipment, funding etc.

## 1.2 A Scenario

All the actors require access anytime, anyplace, anywhere (so-called martini computing) via any appropriate device. The access should be not only to local (job, role or personal) information but, with minimal effort, to the whole world of research information.

A researcher should be able remotely to set up and control experiments (physical experimentation), take and visualise results, access relevant research literature, access datasets and analytical or simulation software (in-silico experimentation) and create new publications (whether academic or project management reports or deliverables) with automated assistance. She should be able to complete research proposals with intelligent assistant software to fill in the standard form fields. She should be able to find suitable research partners in academia or industry. She should be able to utilise computation power, storage and network resources without knowing where they are – only knowing their capabilities are suitable for her task and respecting any restrictions concerning rights acknowledgement or payment. She should also be able to do all the management/administrative tasks comfortably and efficiently within the same environment: completing time sheets, expense claims, purchase requisitions, travel plans etc. The management of research publications must lie comfortably within this environment.

Similarly research managers in research institutions or funding agencies should be able to gain quickly the 'state of the world' in any research area to compare their own organisation with others and thence plan appropriate strategies. This implies knowing what other funding agencies or research institutions have currently in terms of projects, persons, organisation units (e.g. research teams, departments), funding programmes, research outputs (products, patents, publications), events, facilities and equipment and so on.

The point to be stressed is that research outputs are part of a much larger environment, all of which must be recorded and accessible for the end-user to appreciate the research output material.

# 2 Technical Infrastructure

## 2.1 Introduction

The solution comes in several components: the e-Infrastructure provides the connectivity, computing power and software engineering environment for ease of access and ease of use. CRIS (Current Research Information Systems) provide structured information documenting the context of the research and providing structured metadata. OA repositories (of publications) provide the scholarly research output. e-Research repositories (research datasets and software) provide the detailed underpinning material of the research.

## 2.2 e-Infrastructure

Over the last few years it has become apparent that the e-infrastructure solution is based on GRIDs and SOA (service oriented architecture) [1, 2]. The original GRID idea provides metacomputing (linked supercomputers) [3]. The original WWW idea provides access to information but without computation. Bringing them together provides a user-invisible platform [4]. Adding self-* properties (self-management, self-composition, self-repairing, self-tuning) [5] makes the platform effective and efficient. Utilising a SOA (Service-Oriented Architecture) based on discoverable reliable services (pieces of software that execute some function and can be composed into larger software structures to perform human-recognisable tasks) increases the reliability and decreases the software cost. The SOKU (Service-Oriented Knowledge Utility) concept [6] shows much promise: each SOKU would be wrapped in metadata to allow its discovery (descriptive metadata) and to control (parameterise) its execution in both functional (how it does what it does) and non-functional (under what conditions e.g. rights attribution, price) it does it, modes (restrictive metadata) (Figure 1).

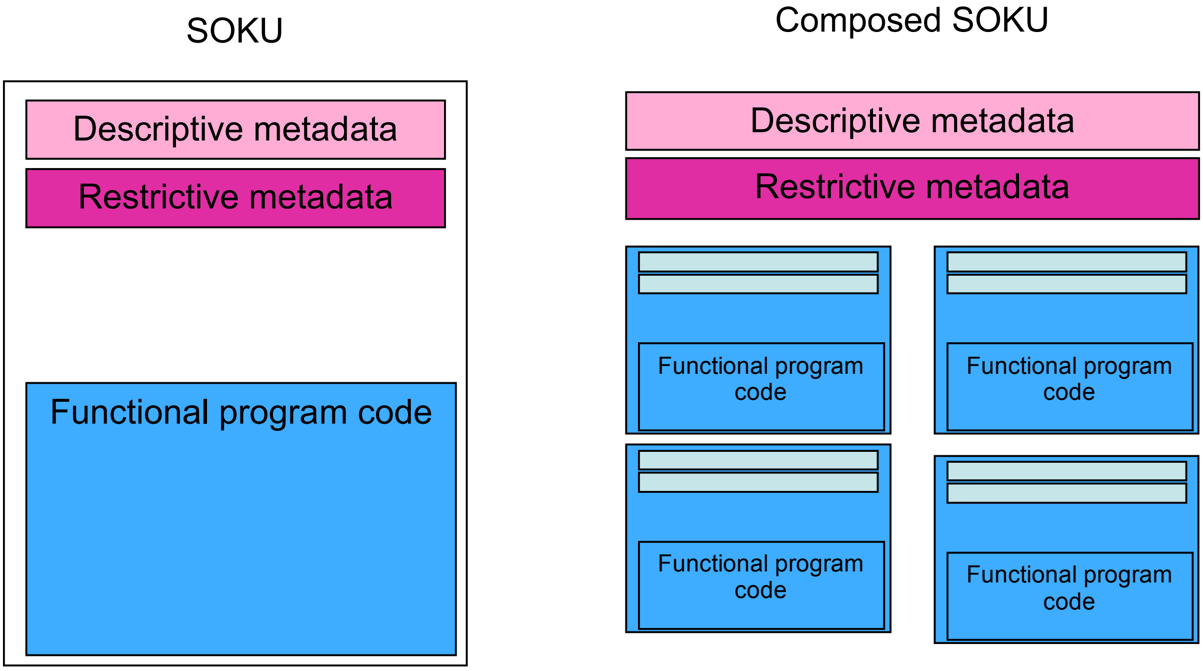


**Figure 1: SOKU**

The critical requirement for effective e-infrastructure has been recognised internationally. The ‘cyberinfrastructure’ [7] in North America follows the work on e-infrastructure based on GEANT in Europe [8] itself partly stimulated by the requirements of research facilities [9]. Individual European countries, too, have invested in e-infrastructure; an example is e-Science (applications) utilising the national GRID service (middleware) itself based on JANET (network) in UK [10]. Similar initiatives have been taken elsewhere notably in Australasia, Japan, China, Singapore, India and also in South America.

These e-infrastructures provide fast networking linking supercomputers, repositories and access to experimental facilities. They have schemes for identification, authentication and authorisation of usage. They have middleware to make the base resources invisible to the end-user and to optimize resource allocation. They are developing methods for homogeneous access to heterogeneous resources.

To date the work has largely been academic. The IT companies have been involved in producing components of the solution; e.g. IBM has an autonomic computing product, ORACLE has a clustered database product. Univa [11] offers a commercialised version of the popular open source GLOBUS middleware. However, there are extensive developments underway in many IT companies to produce GRID/SOA-compatible products and some are even basing their future architectures on SOKU.

The challenge posed is how to utilise this emerging e-infrastructure for benefit and specifically how to use it to make more accessible and available the research literature in a form appropriate for the actors performing their roles as outlined above.

## 2.3 CRIS

CRIS (Current Research Information Systems) have been developed over the last 40 years. Currently an EU Recommendation to member states, CERIF (Common European Research Information Format) is being adopted quite widely and it allows interoperation. A CRIS typically has information on projects, persons, organisational units, funding programmes, research outputs (products, patents, publications), facilities and equipment and events. The novelty of CERIF is its formal data structure, its use of linking relations to allow n:m relationships with role and temporal duration, its use of multiple character sets and provision of multilinguality.

Consider the following case illustrated in (Figure 2)**Fehler! Verweisquelle konnte nicht gefunden werden.** : A person A is an employee of organisation O and a member of organisations M and N both of which are parts of O. She is author of X in which O claims the IPR (intellectual property right) and project leader of P. In CERIF the following records would be in base tables: Person: A; OrgUnit: O,M,N; Publication: X; Project: P. The link tables would be: Person-OrgUnit: A-employee-O, A-member-M, A-member-N; OrgUnit-OrgUnit: M-partof-O; N-partof-O; Person-Publication: A-author-X; OrgUnit-Publication: O-IPR-X; Person-Project: A-projectleader-P. In fact, the link tables include, as well as role, the temporal information concerning start and end date-time. In this example it may be that when A authored X she was no longer a member of M. This, relatively simple, example illustrates the power of CERIF as a data model.

CERIF is maintained by the not-for-profit organisation euroCRIS (www.eurocris.org) from whence details are available. Commercial CRIS offerings are available from uniCRIS [12] which is fully CERIF-compatible, Atira [13], and Avedas [14]. Many funding agencies and research institutions have some form of ‘home-brew’ CRIS, the majority are more-or-less CERIF-compatible. The provision of CRIS in a modern e-infrastructure environment has been discussed in [15].

## 2.4 Repositories

Repositories store and provide access to the detailed information. It is usual to separate repositories of research publications from repositories of research datasets and software (e-Science or, better, e-Research repositories) because of their different access patterns and different metadata requirements. The e-Research repositories require much more detailed metadata to control utilisation of the software and datasets in addition to metadata to allow discovery of the resources. At present they tend to be specific to an individual organisation because of their novelty and the differing requirements on metadata imposed by different (commonly international) communities e.g. in space science, atmospheric physics, materials science, particle physics, humanities or social science. Publication repositories typically use some form of Dublin Core Metadata [16] and most are OAI-PMH (Open Access Initiative – Protocol for Metadata Harvesting) [17] compliant for interoperation and are indexed by Google Scholar. Example software systems are ePrints [18], Dspace [19], Fedora [20] and ePubs [21].

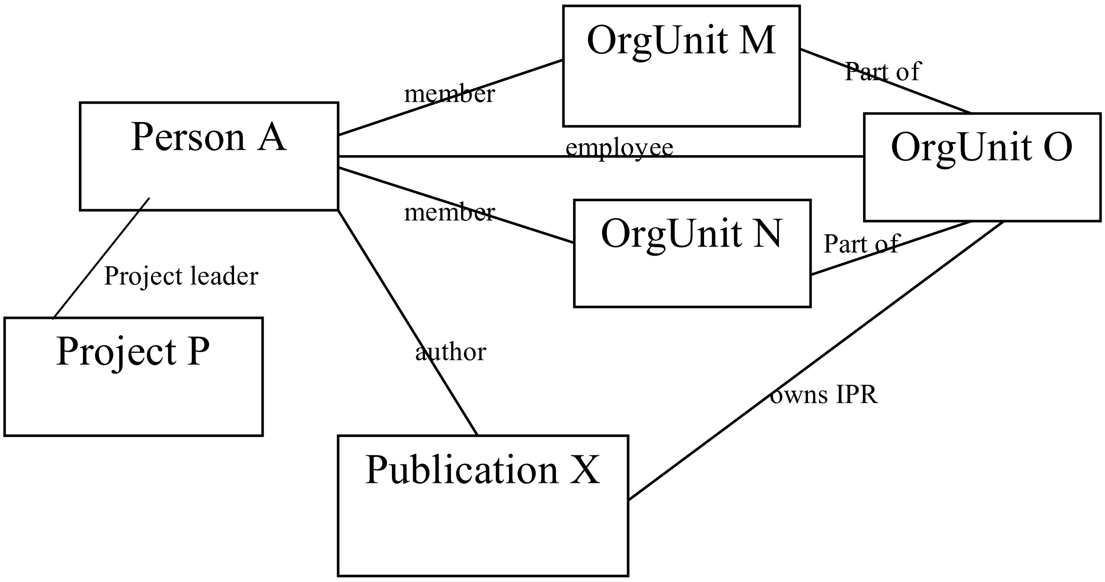


**Figure 2: Example of CERIF**

## 2.5 Metadata and Curation

Digitally-created articles rely heavily on both the metadata record and the articles themselves being deposited. International metadata standards and protocols must be applied to repositories so that retrieval may be consistent with appropriate recall (precision) and relevance so that harvesting (or homogeneous retrieval access) across repositories can take place. A model for formalising metadata [22] is required.

The current DC metadata standards DC [16] and OAI-PMH [17] for interoperability are insufficient for scalable, automated retrieval with appropriate relevance (precision) and recall. DC is machine-readable but not machine-understandable. One basic problem is that a formalised syntax and semantics (vocabulary) for each relevant DC element was not specified in 'simple DC' and has only partially been overcome by the use of namespaces in 'qualified DC'. A second problem concerns the element set tags 'contributor', 'creator' and 'publisher' which are actually roles of a person or organisational unit and should be represented by a relationship (between the article and the person or organisational unit) where the role value belongs to a namespace and is temporally limited . A third problem is the tag 'relation' which is extremely general; the real world is much better modelled through typed relations with role and temporal validity. Other problems include the tag 'coverage' which only recently has been separated into temporal and spatial aspects yet these are fundamental retrieval criteria for much material. A formalised version of DC overcoming these limitations has been suggested [23] and defined [24] to form also part of the CERIF model allowing tight integration with CRIS. Recently the DC community has recognised these problems and with more recent work [25, 26] is attempting to address them.

To ensure that research output material is available for future generations, curation and preservation issues must be addressed. There is current work to define metadata standards to achieve this [27] but a major problem concerns maintaining the articles on current (i.e. usable) media.

## 2.6 Integration

The linking together at an institution of a 'green' OA repository of articles, a CRIS (to provide contextual information) and an OA repository of research datasets and software [28] (Figure 3) ensures that an institution can manage its IP for benefit whether that benefit is in innovation and investment, in educational resources, in stimulation of future research or in publicity. Furthermore, the formalised structure of the CRIS allows a reliable workflow to be engineered which in turn encourages deposit of research outputs. Such a system is being implemented progressively at STFC Rutherford Appleton Laboratory where the CERIF-CRIS is named the Corporate Data Repository, the OA repository is ePubs and the e-research repository is the e-Science repository.

Linking together these institutional CRIS systems - which have a formal structure and hence can be interoperated reliably and in a scalable way [29] - provides a network of access to institutional OA repositories (of articles) or e-research repositories via the CERIF-CRIS gateways enhancing and controlling the access using the CERIF-CRIS information as formalised, structured and contextual metadata which is more detailed than DC and suitable for intelligent (machine-understandable) interoperation (Figure 4). Interoperation of CERIF-CRIS has been demonstrated, most recently for euroHORCS (European Heads of Research Councils) in October 2006. However, as yet, the whole architecture has not been demonstrated.

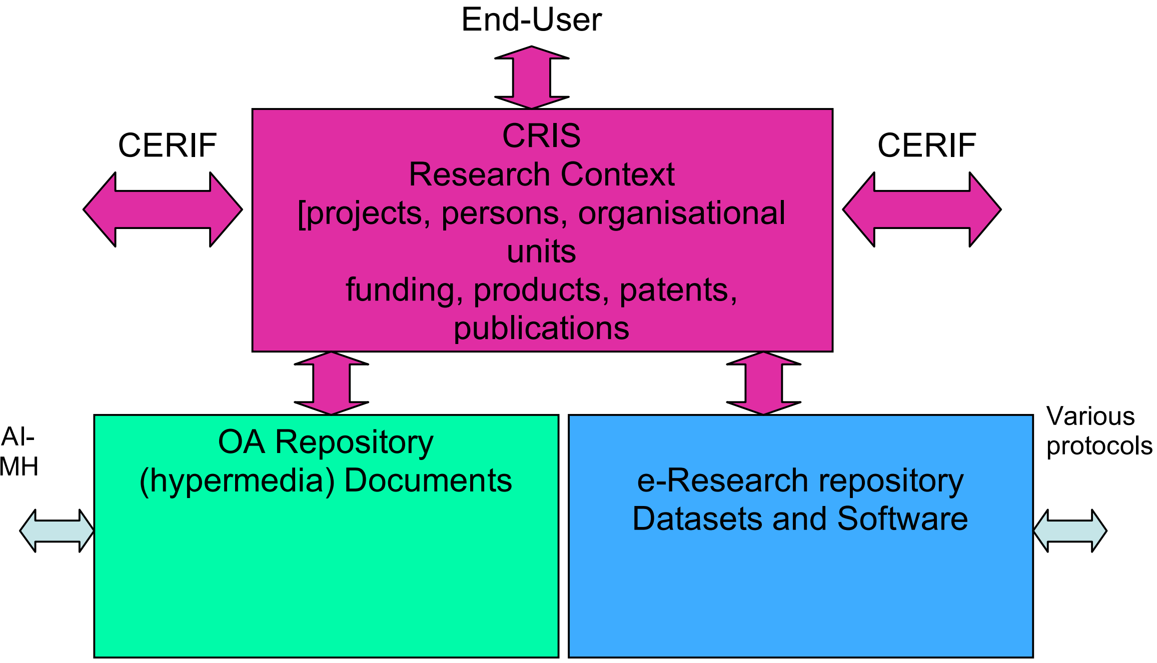


**Figure 3: Architecture for an Institution**

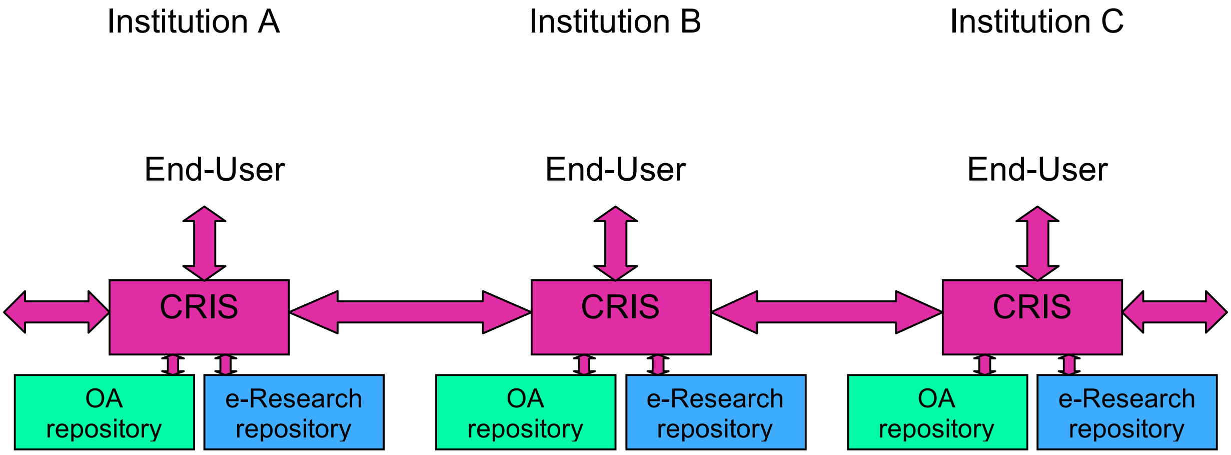


**Figure 4: Architecture for OA**

# 3 Policy Framework

## 3.1 Introduction

The contentious area of discussion in this subject area is open access to research output publications. A brief overview is at [30]. We assume the conference attendees have a good knowledge of OA to distinguish the dimensions of the topic: ‘green’ and ‘gold’; thematic (central) and institutional (distributed); peer-reviewed or not. Furthermore we assume a general knowledge of the differences between white and grey literature. This section discusses motivations and barriers and then concentrates on the two major topics to overcome the barriers: metadata and mandates and finally concludes with speculations on the future.

## 3.2 Motivations

Open Access (OA) means that electronic scholarly articles are available freely at the point of use. The subject has been discussed for over 10 years [31], but has reached a crescendo of discussion over the last few years with various declarations in favour of OA from groups of researchers or their representatives [32-35]. The UK House of Commons Science and Technology Committee [36] considered the issue in 2004, reporting in the summer in favour of OA. This indicates the importance of the issue, and led to statements from large research funding bodies such as the Wellcome Trust [37] and the Research Councils UK [38]. More recently the USA has attempted to move in this direction [39]. What has motivated this interest?

Ethics: There is an ethical argument that research funded by the public should be available to the public. Since research is an international activity, this crosses national boundaries.

Research Impact: The internet provides an opportunity. Modern harvesting techniques and search engines make it possible to discover publications of relevance if they are deposited in an OA repository with a particular metadata standard. If all authors did this then the world of research would be available ‘at the fingertips’. There is evidence that articles available in an OA repository have more accesses / downloads (readers), citations and therefore impact [40, 41]

Costs and economic benefit: There is concern over the hindrance to research caused by the cost of journal subscriptions, whether electronic or paper. These costs run well above the rate of inflation with the result that libraries with restricted budgets (i.e. all of them!) are no longer providing many journals needed by researchers [42-44]. Estimates of the costs of ‘gold’ OA publishing indicate that for a productive institution these costs could exceed by a factor of 3 current subscription costs. Economic benefit of improved (open) access has been studied and ‘green’ OA is regarded as beneficial [45, 46].

Metrics: measures of utilisation of research publications are used for various statistical purposes – usually to indicate quality which may be used in evaluation to allocate research funding. Articles in an OA repository allow automation of such metrics including measures of impact and – most importantly – at the level of the article (not the channel) and evenly across all disciplines and language encodings in contradistinction to how ISI manages ranking. This aspect links with those of research impact and costs and economic benefit.

Added value: articles in an OA repository can easily be linked to structured data contextualising the research (CRIS) [47, 48] and thence to repositories of research datasets and software [28].

Just reward: There is also concern that in traditional scholarly publishing, most of the work (authoring, reviewing, editing) is done freely by the community and that the publishers make excessive profits from the actual publishing (making available) process.

## 3.3 Barriers to OA

Despite the positive motivations there are barriers to OA.

Loss of publisher income: The major objection to ‘green’ self-archiving comes from publishers and learned societies in publisher role (many of which depend on subscriptions to their publications) who fear that ‘green’ OA threatens their business viability. To date there is no evidence that ‘green’ archiving harms the business model of publishing [49, 50]. There is evidence that ‘green’ archiving increases utilisation, citation and impact of a publication [51, 52] and has economic benefits [45, 46]. Whilst the major commercial publishers could provide additional value-added services to offset the impact of OA on current business models, the impact on learned societies may require new business models to be developed.

Copyright: Copyright agreements between authors and publishers may inhibit the ‘green’ route. However, to date, over 90% of publication channels (the variability depends on exactly what is counted) allow ‘green’ author deposit although some insist on an embargo period before the publication is available for OA [53]. In contrast some publishers of journals – of which Nature is the most well-known – do not demand copyright from the author but merely a licence to publish, leaving copyright with the author or their institution.

Difficulties in access and utilization: despite the Dublin Core metadata standard [16] and an interoperation protocol [17] there are difficulties in an end-user obtaining appropriate relevance (precision) and recall in retrieval – certainly when compared with a well-structured library catalog system using e.g. [54]. This indicates that the metadata is insufficient for the purpose. Similarly, if the end-user wishes easy access from the article to research context or associated research datasets and software this is currently extremely difficult. However, linking a repository of articles to a CRIS provides structured metadata which improves greatly relevance (precision) and recall and also provides a link through to e-research repository information.

Completeness: there is great difficulty in persuading researchers to deposit their material in OA repositories. Estimates indicate an 8-15% fill of OA repositories [55] although when a funding organization or institution applies a mandate this rises rapidly to 60%-90% eventually approaching 100% [56]. Additionally, an institution may – following the mandate – assist in automating the process with a workflow such that there is minimum (re)keying of metadata [57]. Again this works best if there is a CRIS with structured metadata.

### 3.4 Mandates

Both the EU [58] and the USA (proposed US Federal Research Public Access Act [39]) have moved towards mandating that output of publicly funded research should be OA. Neither has (as yet) enacted the mandate. For a summary see [59]. The EU went against the results of its own commission study possibly as a result of the 'Brussels Declaration' from the STM (Science, Technology and Medicine) Publishing community [60] despite EURAB (EU Research Advisory Board) recommending green OA [61]. Various funding organisations have mandated open access for the outputs of research that they fund, based on the arguments in 3.2 above. The vast majority mandate 'green' OA (parallel self-archiving in an institutional repository) and some (Wellcome, Hughes) agree to fund in parallel 'gold' (author funder pays) with preferred publishers. More recently CERN (a research institution, not a funder) has proposed to go the 'preferred publisher gold' route [62]. This is surprising since CERN and the particle physics community pioneered 'green' OA with arXiv [63].

The preferred, optimal and recommended procedure is immediately upon acceptance for publication the metadata and full article are deposited in an institutional repository. If the publisher does not demand an embargo period both are set to open access; if an embargo period is demanded then only the metadata is made visible until the end of the embargo period. Of course, associated with the metadata record there can be (and ePrints [18] provides) a 'request button' so that the material can be sent automatically to any researcher who requests it under the usual 'fair use' conditions.

### 3.5 Integration

What is required now is for all funding agencies to mandate green OA in institutional repositories of research output articles, and for all research institutions to maintain such a repository linked to a CRIS and thence to a repository of research datasets and software. This would provide universal open access and allow researchers, research managers, innovators, policymakers, the media and others to access the research knowledge of the world easily, quickly and cheaply thus promoting wealth creation and improvement in the quality of life.

Such a move will be resisted by the Learned Societies (acting as publishers) and the publishing industry for business reasons. There are two possible ways forward: (1) press ahead with 'green' OA ignoring the opposing interests (2) while pressing ahead with 'green' OA also engage in debate with the opposing interests to reassure them that there are business models including OA that can work. Stevan Harnad takes the first view and refutes all needless speculation (a position we admire but with which we cannot agree wholeheartedly); we take the second view with a more pragmatic attitude to securing OA for the future.

Thus, there is a need for engagement with the Learned Societies to develop new methods of peer review which can be paid for in order to preserve those societies and the benefits they bring without requiring them to have a business model based on traditional publishing.

Finally there is a need for engagement with traditional publishers to explore what value-added products they could produce harvesting from a rich world of OA repositories of publications cross-linked via CRISs with associated research datasets and software.

## 4 The Way Forward

In the world of advanced e-infrastructures the progress of research, with its concomitant benefits in wealth creation and improvement in the quality of life, cannot be hindered by obsolete information availability (i.e. commercial publishing) channels.

### 4.1 Speculation: Future

Looking to the future speculatively, it is possible to imagine 'green' OA repositories becoming commonplace and used heavily. At that point, we argue, one could change the business model so that an author deposits in an open access 'green' repository but instead of submitting in parallel to a journal or conference peer-review process, the peer-review is done either by:

a) a learned society managing a 'college' of experts and the reviewing process – for a fee paid by the institution of the author or the author;

b) allowing annotation by any reader (with digital signature to ensure identification / authentication);

in both cases being alerted by 'push technology' that a new article matching their interest profile has been deposited.

The former peer-review mechanism would maintain learned societies in business, would still cost the institution of the author or the author but would probably be less expensive than publisher subscriptions or 'gold' (author or author institution pays) open access. The latter is much more adventurous and in the spirit of the internet; in a charming way it somehow recaptures the scholarly process of two centuries ago (initial draft, open discussion, revision and publication) in a modern world context. It is this possible future that is feared by commercial publishers.

## 5 Conclusion

Despite protests and obstacles to improved access to research material over the centuries from religious, commercial, professional or labour groups, none delayed for long progress to meet the requirement as defined by the research community. The advanced international e-infrastructure provides 'martini computing' and invisibility of resources to the end-user. It supports access to structured research information on projects, persons, organisational units, funding, research outputs (products, patents, publications), research facilities and equipment, events and more (CRIS). It supports repositories of articles and of research datasets and software. It supports access to experimental facilities and 'computational steering' of experiments whether physical or 'in silico'. There is a new world of research capability. Electronic research output publications must take their place in this new world of accessibility and utilisation unhindered by outdated prejudices. This will lead to maximum use of - and benefits from - the research output for quality evaluation, for innovation, for further research, for education, for research management and planning and for informing public debate on research issues.

## Acknowledgements

This article owes much to others. In the area of e-infrastructure I am indebted particularly to colleagues on the NGG expert groups and the UK e-infrastructure group. In the area of CRIS, members of euroCRIS have provided much insight. In the area of OA and repositories the luminaries include Stevan Harnad and Alma Swan to both of whom I owe much for deep discussions and shared thinking. Stevan Harnad, Leslie Chan and Anne Asserson kindly improved the article with their suggestions.

## References

[1] http://www.ercim.org/publication/Ercim_News/enw45/

[2] http://www.ercim.org/publication/Ercim_News/enw59

[3] FOSTER, I; KESSELMAN, C (Eds). The Grid: Blueprint for a New Computing Infrastructure. Morgan-Kauffman 1998

[4] ftp://ftp.cordis.europa.eu/pub/ist/docs/ngg_eg_final.pdf

[5] ftp://ftp.cordis.europa.eu/pub/ist/docs/ngg2_eg_final.pdf

[6] ftp://ftp.cordis.europa.eu/pub/ist/docs/grids/ngg3_eg_final.pdf

[7] http://www.adec.edu/nsf/nsfcyberinfrastructure.html

[8] EU reflection group roadmap: http://www.e-irg.org/roadmap/eIRG-roadmap.pdf

[9] ESFRI roadmap report: ftp://ftp.cordis.europa.eu/pub/esfri/docs/esfri-roadmap-report-26092006_en.pdf

[10] UK national report: http://www.nesc.ac.uk/documents/OSI/report.pdf

[11] http://www.univa.com/

[12] www.unicris.com

[13] www.atira.com

[14] www.avedas.com

[15] JEFFERY, K.G.; 'The New Technologies: can CRISs Benefit' in A Nase, G van Grootel (Eds) Proceedings CRIS2004 Conference, Leuven University Press ISBN 90 5867 3839 May 2004 pp 77-88

[16] http://dublincore.org/

[17] http://www.openarchives.org/OAI/openarchivesprotocol.html

[18] http://www.eprints.org/

[19] http://www.dspace.org/

[20] http://www.fedora.info/

[21] http://epubs.cclrc.ac.uk/

[22] JEFFERY, K G: 'Metadata': in Brinkkemper,J; Lindencrona,E; Solvberg,A (Eds): 'Information Systems Engineering' Springer Verlag, London 2000. ISBN 1-85233-317-0.

[23] JEFFERY, K G: 'An Architecture for Grey Literature in a R&D Context' Proceedings GL'99 (Grey Literature) Conference Washington DC October 1999 http://www.konbib.nl/greynet/frame4.htm

[24] ASSERSON, A; JEFFERY, K.G.; 'Research Output Publications and CRIS' The Grey Journal volume 1 number 1: Spring 2005 TextRelease/Greynet ISSN 1574-1796 pp5-8

[25] http://dublincore.org/documents/2007/04/02/abstract-model/

[26] http://dublincore.org/documents/2007/04/02/dc-rdf/

[27] http://ssdoo.gsfc.nasa.gov/nost/isoas/

[28] JEFFERY, K.G; ASSERSON, A: 'CRIS Central Relating Information System' in Anne Gams Steine Asserson, Eduard J Simons (Eds) 'Enabling Interaction and Quality: Beyond he Hanseatic League'; Proceedings 8th International Conference on Current Research Information Systems CRIS2006 Conference, Bergen, May 2006 pp109-120 Leuven University Press ISBN 978 90 5867 536 1

[29] JEFFERY, K G CRISs, Architectures and CERIF CCLRC-RAL Technical Report RAL-TR-2005-003 (2005) http://epubs.cclrc.ac.uk/work-details?w=33728

[30] http://www.ercim.org/publication/Ercim_News/enw64/jeffery.html

[31] HARNAD, S. (1995) A Subversive Proposal. In: Ann Okerson & James O'Donnell (Eds.) Scholarly Journals at the Crossroads; A Subversive Proposal for Electronic Publishing. Washington, DC., Association of Research Libraries, June 1995. http://www.arl.org/scomm/subversive/toc.html

[32] http://www.soros.org/openaccess/read.shtml

[33] http://www.earlham.edu/~peters/fos/bethesda.htm

[34] http://www.zim.mpg.de/openaccess-berlin/signatories.html

[35] http://www.oecd.org/document/15/0,2340,en_2649_201185_25998799_1_1_1_1,00.html

[36] UK House of Commons Science and Technology Select Committee http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/3990

[37] http://www.wellcome.ac.uk/doc_wtx026830.html

[38] http://www.rcuk.ac.uk/access/statement.pdf

[39] http://www.taxpayeraccess.org/frpaa/

[40] HARNAD, S.; CARR, L.; BRODY, T.; OPPENHEIM, C. (2003) Mandated online RAE CVs Linked to University Eprint Archives: Improving the UK Research Assessment Exercise whilst making it cheaper and easier. Ariadne 35. http://www.ecs.soton.ac.uk/~harnad/Temp/Ariadne-RAE.htm

[41] HARNAD, S. (2007) Open Access Scientometrics and the UK Research Assessment Exercise. Proceedings of the 11th Annual Meeting of the International Society for Scientometrics and Informetrics. Madrid, Spain, 25 June 2007 http://arxiv.org/abs/cs.IR/0703131

[42] ROWLANDS, I.; NICHOLAS, D. (2005) /New Journal Publishing

[43] SPARKS, S. (2005) /JISC Disciplinary Differences Report,/ Rightscom, London. http://www.jisc.ac.uk/uploaded_documents/Disciplinary%20Differences%20and%2=0Needs.doc.

[44] EPS (2006) UK scholarly journals: 2006 baseline report An evidence-based analysis of data concerning scholarly journal publishing, RIN, RCUK and DTI,. Available at http://www.rin.ac.uk/data-scholarly-journals.

[45] HOUGHTON, J., STEELE, C., SHEEHAN, P. (2006) Research Communication Costs in Australia: Emerging Opportunities and Benefits. A report to the Department of Education, Science and Training. http://www.dest.gov.au/NR/rdonlyres/0ACB271F-EA7D-4FAF-B3F7-0381F441B175/13935/DEST_Research_Communications_Cost_Report_Sept2006.pdf

[46] HOUGHTON, J.; SHEEHAN, P. (2006) The Economic Impact of Enhanced Access to Research Findings. Centre for Strategic Economic Studies Victoria University http://www.cfses.com/documents/wp23.pdf

[47] ASSERSON, A; JEFFERY, K.G.; 'Research Output Publications and CRIS' in A Nase, G van Grootel (Eds) Proceedings CRIS2004 Conference, Leuven University Press ISBN 90 5867 3839 May 2004 pp 29-40

[48] DIJK, ELLY; BAARS, CHRIS; HOGENAAR, ARJAN; VAN MEEL, MARGA (2006) NARCIS: The Gateway to Dutch Scientific Information ELPUB2006. Digital Spectrum: Integrating Technology and Culture - Proceedings of the 10th International Conference on Electronic Publishing held in Bansko, Bulgaria 14-16 June 2006 / Edited by: Bob Martens, Milena Dobreva. ISBN 978-954-16-0040-5, 2006, pp. 49-58

[49] SWAN, A;.BROWN, S (2005) Open access self-archiving: an author study. http://www.jisc.ac.uk/uploaded_documents/Open%20Access%20Self%20Archiving-an%20author%20study.pdf

[50] BERNERS-LEE, T.; DE ROURE, D.; HARNAD, S.; SHADBOLT, N. (2005) Journal publishing and author self-archiving: Peaceful Co-Existence and Fruitful Collaboration. http://eprints.ecs.soton.ac.uk/11160/

[51] HARNAD, S.; CARR, L.; BRODY, T.; OPPENHEIM, C. (2003) Mandated online RAE CVs Linked to University Eprint Archives: Improving the UK Research Assessment Exercise whilst making it cheaper and easier. Ariadne 35. http://www.ecs.soton.ac.uk/~harnad/Temp/Ariadne-RAE.htm

[52] HARNAD, S. (2007) Open Access Scientometrics and the UK Research Assessment Exercise. Proceedings of the 11th Annual Meeting of the International Society for Scientometrics and Informetrics. Madrid, Spain, 25 June 2007 http://arxiv.org/abs/cs.IR/0703131

[53] http://romeo.eprints.org/stats.php

[54] http://www.loc.gov/marc/

[55] http://www.crsc.uqam.ca/lab/chawki/graphes/EtudeImpact.htm

[56] SALE, A. (2007) The Patchwork Mandate D-Lib Magazine 13 1/2 January/February http://www.dlib.org/dlib/january07/sale/01sale.html doi:10.1045/january2007-sale.

[57] JEFFERY, K G; ASSERSON, A: 'Supporting the Research Process with a CRIS' in Anne Gams Steine Asserson, Eduard J Simons (Eds) 'Enabling Interaction and Quality: Beyond the Hanseatic League'; Proceedings 8th International Conference on Current Research Information Systems CRIS2006 Conference, Bergen, May 2006 pp 121-130 Leuven University Press.

[58] http://ec.europa.eu/information_society/activities/digital_libraries/doc/scientific_information/communication_en.pdf

[59] http://poynder.blogspot.com/2007/03/open-access-war-in-europe.html

[60] http://www.stm-assoc.org/documents-statements-public-co/2007%20BRUSSELS%20DECLARATION%20130207.pdf

[61] http://ec.europa.eu/research/eurab/pdf/eurab_scipub_report_recomm_dec06_en.pdf
965. Cambridge, Mass. : M.I.T. Press, 1965, p. 219.

[62] http://ec.europa.eu/research/science-society/document_library/pdf_06/aymar-022007_en.pdf

[63] http://arxiv.org/

# Scientific Publishing in the Digital Era

*Norbert Kroó*

Hungarian Academy of Sciences, Roosevelt tér 9, 1051 Budapest, Hungary
e-mail: kroo@office.mta.hu

## Keynote Abstract

The new information technology developments change drastically our life. The same applies to scientific research in general and the publication of findings in particular. It offers the chance for faster dissemination of results and broader access to date. The interests of scientists, financing organizations and libraries on one hand and publishers on the other do not overlap completely. Maximizing the speed of dissemination, broad access and securing quality and long time preservation are fields of overlapping interests. Mandatory deposit in open access repositories and pricing are still debated. The lecture discusses the above issues based partly on the basis of the author's motivation to maximize the benefits of public (and so EC) funded research in Europe, influenced by his experience both in European scientific organizations and advisory bodies of the EC.

**Keywords:** European Union; information technology; open access; research impact

# Open Access Publishing in High-Energy Physics

*Salvatore Mele*[1, 2]

[1] CH-1211, Genève 23, Switzerland
[2] On leave of absence from INFN, I-80126, Napoli, Italy
e-mail: Salvatore.Mele@cern.ch
On behalf of the SCOAP$^3$ Working Party

## Abstract

The goal of Open Access (OA) is to grant anyone, anywhere and anytime free access to the results of scientific research. The High-Energy Physics (HEP) community has pioneered OA with its "pre-print culture": the mass mailing, first, and the online posting, later, of preliminary versions of its articles. After almost half a century of widespread dissemination of pre-prints, the time is ripe for the HEP community to explore OA publishing. Among other possible models, a sponsoring consortium appears as the most viable option for a transition of HEP peer-reviewed literature to OA. A Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP$^3$) is proposed as a central body which would remunerate publishers for the peer-review service, effectively replacing the "reader-pays" model of traditional subscriptions with an "author-side" funding. Funding to SCOAP$^3$ would come from HEP funding agencies and library consortia through a re-direction of subscriptions. This model is discussed in details together with a quantitative description of the HEP publishing landscape leading to a practical proposal for a seamless transition of HEP peer-reviewed literature to OA publishing.

**Keywords:** open access publishing; high-energy physics; CERN; SCOAP$^3$

## 1 Introduction

The goal of *"Open Access"* (OA) is to grant anyone, anywhere and anytime, free access to the results of scientific research [1]. The OA debate has gained considerable momentum in recent years across many disciplines, both in the sciences and the humanities. In High Energy Physics (HEP) this debate is driven mostly by two factors:

- The "serials crisis" of ever-rising costs of journals, which has forced libraries to cancel a steadily increasing number of subscriptions, curtailing the access of researchers to scientific literature. This traditional business has become financially unsustainable for libraries and publishers alike;
- The increasing awareness that results of publicly funded research should be made generally available going past the availability of pre-prints, towards peer-reviewed literature.

HEP pioneered OA through its "pre-print culture": the mass mailing for four decades of preliminary versions of articles, so to ensure their largest diffusion. With the onset of the Internet, the HEP community spearheaded the culture of "repositories": online collections of freely accessible pre-prints. Thanks to the speed at which they make results available, repositories have become the lifeblood of HEP scientific information exchange. However, they usually contain the original version of articles *submitted* to journals, and not the final, peer-reviewed, *published* version. Notwithstanding the success of repositories, there is consensus in the HEP community that high-quality journals still play a pivot role, by providing [2]:

- quality control through the peer-review process;
- a platform for the evaluation and career evolution of scientists;
- a measure of the quality and productivity of research groups and institutes.

A powerful synergy can arise between the strong OA culture of the HEP community, which finds its roots in four decades of preprint circulation, and its continuing need for high-quality journals, leading to a unique opportunity for a possible transition to OA publishing of the HEP peer-reviewed literature. The community is now moving towards such groundbreaking transition through the establishment of a consortium, SCOAP$^3$ (Sponsoring Consortium for Open Access Publishing in Particle Physics). This consortium would engage publishers of high-quality peer-reviewed journals in order to cover the costs of the peer-review process with funds previously used for journal subscriptions. This idea is viable for the HEP community since the author and the reader communities largely overlap, and are mostly funded by the same actors. This article describes the SCOAP$^3$ initiative:

- section 2 puts the HEP publishing landscape into context, and describes the background to OA Publishing in HEP and the steps which led to the SCOAP$^3$ initiative;
- section 3 presents the SCOAP$^3$ model through the roles of the main stakeholders in HEP scientific publishing;
- section 4 illustrates the results of an analysis of the HEP publishing landscape and their consequences on the targets of the SCOAP$^3$ initiative;
- section 5 presents the financial aspects of the SCOAP$^3$ model together with a cost-sharing scenario based on an investigation of the author basis of HEP;
- section 6 concludes the article by presenting the status of the initiative at the time of writing.

## 2 Background

A recent study analyzed articles submitted in 2005 to the *arXiv.org* repository and classified in the *hep-ex, hep-lat, hep-ph* and *hep-th* categories and subsequently published. Out of a total of about 5'000 articles, more than 80% appeared in just six peer-reviewed journals from four publishers [3]: *Physical Review* and *Physical Review Letters* (published by the American Physical Society), *Physics Letters* and *Nuclear Physics B* (Elsevier), *Journal of High Energy Physics* (SISSA/IOP) and the *European Physical Journal* (Springer). Almost 90% of the articles were published by just four publishers, two out of which (American Physical Society and SISSA/IOP) are learned society.

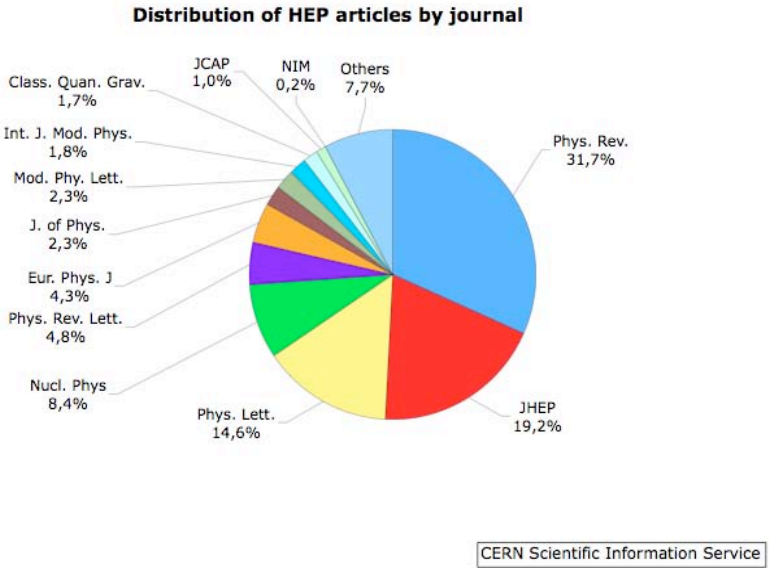


**Figure 1: Distribution of the HEP articles submitted in 2005 to the *arXiv.org* repository under the categories *hep-ex*, *hep-lat*, *hep-ph* and *hep-th* and subsequently published in peer-reviewed journals. A total sample of about 5'000 articles is considered. Only journal with a total share above 1% are considered, with the exception of *Nuclear Instrument and Methods in Physics Research* (NIM). The "Others" group comprises 77 remaining journals. From reference [3].**

These findings, summarised in figures 1 and 2, spotlight two fundamental points relevant for a possible transition of HEP publishing to OA: the volume of articles is small and these are concentrated in a few core titles, mostly published by learned societies. All HEP leading journals have recently taken a pro-active stance on OA. Journals from the American Physical Society, Elsevier and Springer offer authors an option to pay a fee to make their articles OA, while the *Journal of High Energy Physics* is recently experimenting with an institutional membership fee. The latter appears a more successful scheme, as funding mechanisms in HEP seldom include overhead for scientific publications to be directly used by authors. Moreover, the direct payment for the OA publication of an articles is perceived in very negative terms by the community, reminiscent of the unpopular "page charges" of some journals. This perception might have extended to other journals, such as the *New*

*Journal of Physics*, which are built on a "pay-per-article" Open Access model, but have so far attracted only a limited HEP content.

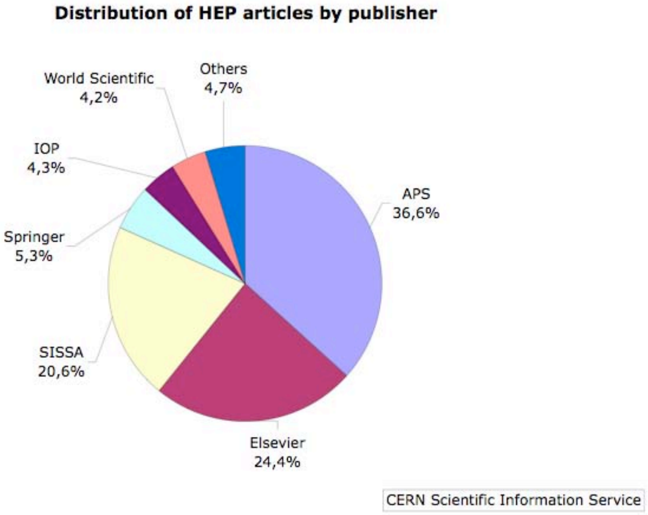


**Figure 2: Distribution by publisher of the HEP articles submitted in 2005 to the *arXiv.org* repository under the categories *hep-ex*, *hep-lat*, *hep-ph* and *hep-th* and subsequently published in peer-reviewed journals. A total sample of about 5'000 articles is considered. From reference [3].**

The debate on OA publishing in HEP was initiated by CERN. CERN is the leading HEP laboratory, with over half a century of history. Its flagship program, the LHC accelerator, will see four large experimental collaborations probe fundamental questions in our understanding of the Universe. CERN epitomizes cross-border collaboration in HEP: the LHC accelerator and detectors include components built in HEP laboratories and Universities around the world; the largest of the LHC experimental collaborations count as many as 2000 scientists, including about 400 students from 160 universities and laboratories spread over 35 countries. As part of its role to chart the future of HEP, in synergy with HEP funding agencies worldwide, CERN promoted several events focussed on OA publishing in HEP:

- September 2005. *Open meeting on the changing publishing model.* This event brought together representatives of authors, funding agencies and publishers with the aims of first discussing in HEP publishing issues such as publishing costs, competition, fair distribution of costs, opportunities for developing countries, alternative business models and the quality of peer-review [4];
- December 2005. *Colloquium on Open Access publishing in particle physics.* A representative group of authors, funding agencies and publishers indicated a possible way forward to OA publishing based on three pillars: asserting the complementary roles of repositories and peer-reviewed literature, decoupling preservation issues and the publication model, enshrining the importance of peer-review for evaluation and academic credibility [5];
- December 2005 to June 2006. *Task Force on Open Access Publishing in Particle Physics.* This tri-partite task force composed by authors, funding agencies and publishers was charged by the main stakeholders to "*study and develop sustainable business models for OA publishing for existing and new journals and publishers in particle physics*". In its report [2] it suggested to establish a sponsoring consortium, SCOAP3, as a central body which would remunerate publishers for the peer-review service, effectively replacing the "reader-pays" model of traditional subscriptions with an "author-side" funding;
- November 2006. *Establishing a sponsoring consortium for Open Access publishing in Particle Physics.* Following the task-force report and the acceptance of its model by representatives from major European stakeholders, a Working Party was established to develop a specific proposal for the creation of SCOAP$^3$, which is described in this article [6]. More information is contained in the SCOAP$^3$ Working Party report [7].

# 3 The SCOAP$^3$ Model

SCOAP$^3$ will act as a single interface between the main stakeholders of the HEP scientific information market: on one side the author and reader communities and on the other side the publishers of high-quality HEP journals. The aim of SCOAP$^3$ is to establish OA to HEP peer-reviewed articles along the lines of the Budapest Initiative [8], namely *"free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself"*. At the time of writing, SCOAP$^3$ is an initiative emanating from:

- several European funding agencies, among which IN2P3 and CEA (France), INFN (Italy), MPG (Germany), PPARC (U.K.), and other funding bodies from Greece, Norway, Sweden and Switzerland;
- the two largest European laboratories, CERN and DESY;
- national and international library consortia such as GASCO (Germany, Austria, Switzerland), INFER (Italy), COUPERIN (France), JISC (U.K.), ABM-uitvikling (Norway).

In the next months SCOAP$^3$ aims to federate similar agents worldwide: the SCOAP$^3$ model will only be successful if all countries contributing to the vast majority of the HEP literature become members of the consortium. Indeed, a pillar of the SCOAP$^3$ model is to ensure OA to all HEP articles appearing in high-quality journals, irrespective of the affiliation of their authors. Manuscripts from authors without academic affiliation or authors from less-privileged countries, which cannot be reasonably expected to contribute to the consortium at this time, will be treated like all other articles. The ethical reason of conserving the access of any author to peer review is obvious. At the same time, this choice has solid financial reasons: restricting OA privileges only to authors affiliated to some countries would simply replace the present toll-access barriers with different barriers, connected to the geographical origin of the articles. Moreover, if only a geographical subset of the HEP scientific literature were available OA, consortium members would still be required to purchase the remaining fraction, with no evident financial benefits from the OA transition.

SCOAP$^3$ will be financed with funds currently used for journal subscriptions by HEP funding agencies, laboratories and libraries. At the same time it will engage other bodies interested in the broad and free dissemination of scientific information. Each country will contribute to SCOAP$^3$ in a "fair" way, according to its share of the worldwide HEP scientific production, as discussed in Section 5. For the SCOAP$^3$ model to be successful, it should represent a stable, viable and sustainable alternative to subscriptions *vis-à-vis* its partners. It is therefore expected that the SCOAP$^3$ operation will follow the financial blueprint of large HEP scientific collaborations, which usually bind over one hundred laboratories and universities in Memoranda of Understanding spanning several decades.

The innovation of the SCOAP$^3$ model with respect to other OA options currently offered by most publishers is that it will centralize all OA expenses, which will not have to be borne by authors and research groups. These other "author-pays" options, of scarce success in HEP, are perceived as an even higher barrier than subscription charges, in particular for theoretical physicists from small institutions, whose articles account for the vast majority of HEP papers.

It is expected that SCOAP$^3$ will contribute to stabilize the rising cost of access to information in the HEP domain by virtue of increasing author awareness to costs and prices, and by fostering new competition in the market, linking quality and price.

A large fraction of the publications on core HEP subjects is published in a limited number of journals [3], as detailed in sections 2 and 4. Among those journals, some carry almost entirely HEP content. SCOAP$^3$ aims to assist publishers in converting these entire "core" journals to OA. It is expected that the vast majority of the SCOAP$^3$ budget will be spent for "core" journals with a "lump-sum" payment: SCOAP$^3$ pays a negotiated price for the peer-review of all articles processed by a journal. Many HEP articles appear in "broadband" journals, which carry just a small fraction of HEP content. It is expected that these articles will be sponsored by SCOAP$^3$ on a "pay-per-article" basis. Conference proceedings and monographs are not within the scope of SCOAP$^3$.

In the SCOAP$^3$ OA model, the publishers will have the prime responsibility of ensure quality of the highest standards through independent editorial boards and the peer review. They will ensure the dissemination of OA articles by posting them onto their web sites and, in addition, feeding them to a SCOAP$^3$ repository.

Publishers will benefit from a more sustainable business model than the traditional subscription scheme, becoming increasingly fragile. They will continue to meet the demand for print subscription, re-print of single articles, color plates in these printed versions, collections of articles in electronic or paper form, citation databases and other "premium" services, which are outside the scope of SCOAP$^3$.

# 4 The High Energy Physics Publishing Landscape

The definition of HEP is often linked to the theoretical and experimental study of particles produced at accelerators of ever-increasing energy. Both the field and its definition have evolved to include subjects naturally more close to the fields of nuclear physics, of astrophysics and of cosmology. Different authors, different journals and different funding agencies each focus on different parts of the HEP spectrum and therefore have a different definition of the field.

To be successful, SCOAP$^3$ should, at once, aim to convert to OA the subset of scientific literature of common interest to all players, while striving for as wide a scope as possible. A minimal set of common interest to the entire HEP community is constituted by a "core" set of topics such as the phenomenology and experimental investigations of elementary particles and their interactions, quantum-field theory and lattice-field theory. These topics are loosely related to the *hep-ex*, *hep-lat*, *hep-ph*, and *hep-th* areas of the *arXiv.org* repository, which often also carry content in cognate disciplines. Experimental techniques as well as mathematical and numerical methods are also included in this definition of HEP "core" articles. The definition of HEP article covers more loosely other fields of relevance to HEP, such as selected topics in nuclear physics, astrophysics, gravitation and cosmology.

It is important to note that the vast majority of HEP articles concern phenomenology and theory and have on average 2.6 authors [3]. On the other hand, publications on experimental results were often authored by up to 500 scientists in the last decade, while collaborations now publishing their analyses count up to 800 researchers and articles by LHC collaborations will have up to 2000 authors.

| Journal | Publisher | IF | $N_{tot}$ | $N_{HEP}$ | $N_{core}$ | $f_{HEP}$ | $f_{core}$ |
|---|---|---|---|---|---|---|---|
| *Phys. Rev. D* | APS | 4.8 | 2285 | 2101 | 1635 | 92% | 72% |
| *JHEP* | SISSA/IOP | 5.9 | 856 | 856 | 840 | 100% | 98% |
| *Phys. Lett. B* | Elsevier | 5.3 | 957 | 862 | 740 | 90% | 77% |
| *Nucl. Phys. B* | Elsevier | 5.5 | 522 | 481 | 465 | 92% | 89% |
| *Phys. Rev. Lett.* | APS | 7.5 | 3836 | 407 | 279 | 11% | 7% |
| *Eur. Phys. J. C* | Springer | 3.2 | 331 | 272 | 234 | 82% | 71% |
| *Mod. Phys. Lett. A* | World Scientific | 1.2 | 281 | 216 | 138 | 77% | 49% |
| *Phys. Rev. C* | APS | 3.6 | 853 | 298 | 136 | 35% | 16% |
| *Class. Quant. Grav.* | IOP | 2.9 | 491 | 255 | 89 | 52% | 18% |
| *Int. J. Mod. Phys. A* | World Scientific | 1.5 | 878 | 143 | 88 | 16% | 10% |
| *J. Math. Phys.* | AIP | 1.2 | 446 | 108 | 74 | 24% | 17% |
| *Phys. Atom. Nucl.* | Springer | 0.9 | 220 | 106 | 72 | 48% | 33% |
| *JCAP* | SISSA/IOP | 6.7 | 156 | 128 | 57 | 82% | 37% |
| *Gen. Rel. Grav.* | Springer | 1.6 | 190 | 103 | 20 | 54% | 11% |
| *Nucl. Instrum. Meth. A* | Elsevier | 1.2 | 1371 | 312 | 16 | 23% | 1% |

**Table 1: The most popular HEP journals and their publishers, together with their ISI impact factor, IF; the total number of articles published in 2005, $N_{tot}$; the number of HEP articles, $N_{HEP}$; and the number of articles in the HEP core subject, $N_{core}$. The journals are ordered in decreasing order of $N_{core}$. Only journals with $N_{HEP}$>100 are shown. The last two columns show the fractions $f_{HEP}$ and $f_{core}$ of HEP and core articles, respectively. From reference [7].**

In 2005, about 8'500 HEP articles were published in peer-reviewed journals, as included in the SPIRES database [9]. Of these, 5'200 articles are classified in the core HEP topics discussed above. Table 1 presents the most popular HEP journals and their corresponding publishers, together with their ISI impact factor, IF [10], and the total number of articles published in 2005, $N_{tot}$. The number of HEP articles, $N_{HEP}$, is also listed together with the number of articles in the core subject areas of phenomenology and experimental investigations of elementary particles and their interactions, quantum-field theory and lattice-field theory, $N_{core}$. The journals are ordered in decreasing order of $N_{core}$. Only journals with $N_{HEP}$>100 are shown. The last two columns show the fractions $f_{HEP}$ and $f_{core}$ of HEP and core articles, respectively [7].

As discussed in section 2, a recent study analyzed core HEP articles submitted in 2005 to the *arXiv.org* repository and classified in the *hep-ex, hep-lat, hep-ph* and *hep-th* categories and subsequently published. Out of a total of about 5'000 articles, more than 80% appeared in just six peer-reviewed journals from four publishers [3]. Five out of these six journals carry a majority of HEP content, as listed in table 1, these are:

- *Physical Review D* (published by the American Physical Society);
- *Physics Letters B* (Elsevier);
- *Nuclear Physics B* (Elsevier);
- *Journal of High Energy Physics* (SISSA/IOP);
- *European Physical Journal C* (Springer).

SCOAP$^3$ aims to assist publishers in converting these “core” journals entirely to OA. As described in the last column of table 1, these five “core” journals include up to 30% of articles beyond the core HEP topics, particularly in Nuclear Physics and Astroparticle Physics. These articles will also be included in the OA conversion of the journals. This is in the interest of the HEP readership and promotes the long-term goal of an extension of the SCOAP$^3$ model to these related disciplines.

The sixth journal, *Physical Review Letters* (American Physical Society), is a “broadband” journal, which carries only a small fraction (10%) of HEP content. SCOAP$^3$ aims to sponsor the conversion to OA of this fraction on an article-by-article basis. A similar approach holds for another popular “broadband” journal in instrumentation: *Nuclear Instruments and Methods in Physics Research A* (Elsevier), which carries about 25% of HEP content.

These seven journals covered, in 2005, around 4'200 core HEP articles and about 5'300 articles in the wider HEP definition, including all related subjects. The conversion to OA of these five “core” journals and the HEP part of these two “broadband” journals would cover over 80% of the core HEP subjects and over 60% of the entire HEP literature, including all related subjects. The remaining 3'300 HEP articles, not published in the journals mentioned above, are scattered over some 140 other journals. It is important to note that the SCOAP$^3$ model should not be limited to this set of journals but is open to all existing and future high-quality journals which carry HEP content, within budgetary limits.

## 5 Financial Aspects of the SCOAP$^3$ Model

The price of a journal is driven by the costs to run the peer-review system, by editorial costs for copy-editing and typesetting, by the cost for electronic publishing and access control, and by subscription administration. Some publishers today quote a cost, from reception to final publication, in the range of 1'000 –2'000 Euros per published article [11]. This includes the cost of processing articles which are eventually rejected, the fraction of which varies substantially from journal to journal.

The annual budget for a transition of HEP publishing to OA can be estimated from this figure and the fact that the five “core” journals, which cover a large fraction of the HEP literature, publish about 5'000 articles per year. Hence, we estimate that the annual budget for a transition of HEP publishing to OA would amount to a maximum of 10 Million Euros per year.

Another indication which corroborates this estimate is that the costs to run a “core” journal such as *Physical Review D,* amount to 2.7 Million Euros per year [11] and it covers about a third of the HEP publication landscape [3].

A “fair-share” scenario for the financing of SCOAP$^3$ is to distribute these costs among all countries active in HEP on a *pro-rata* basis, taking into account the size of the HEP author base of each country. To cover publications from scientists from developing countries, which cannot be reasonably expected to contribute to the consortium at this time, an allowance of not more than 10% of the SCOAP$^3$ budget is foreseen.

The size of the HEP author base in each country is estimated from a recent study [7] which considered all articles published in the years 2005 and 2006 in the five HEP “core” journals, *Physical Review D*, *Physics Letters B*, *Nuclear Physics B*, *Journal of High Energy Physics* and the *European Physical Journal C*, as well as those HEP articles published in the two “broadband” journals, *Physical Review Letters* and *Nuclear Instruments and Methods in Physics Research A*. A total sample of about 11'300 articles was considered and, in each of them, all authors were uniquely assigned to a given country. CERN was treated as an additional country.

In about 5% of the cases, authors were found to have multiple affiliations, often in different countries, reflecting the intense cross-border tradition of HEP. In these cases, the ambiguity in the assignment of authors to countries was solved as described in reference [7]. The results from this study are summarized in table 2 and figure 3.

| Country | Share of HEP Scientific Publishing |
|---|---:|
| United States | 24.3% |
| Germany | 9.1% |
| Japan | 7.1% |
| Italy | 6.9% |
| United Kingdom | 6.6% |
| China | 5.6% |
| France | 3.8% |
| Russia | 3.4% |
| Spain | 3.1% |
| Canada | 2.8% |
| Brazil | 2.7% |
| India | 2.7% |
| CERN | 2.1% |
| Korea | 1.8% |
| Switzerland | 1.3% |
| Poland | 1.3% |
| Israel | 1.0% |
| Iran | 0.9% |
| Netherlands | 0.9% |
| Portugal | 0.9% |
| Taiwan | 0.8% |
| Mexico | 0.8% |
| Sweden | 0.8% |
| Belgium | 0.7% |
| Greece | 0.7% |
| Denmark | 0.6% |
| Australia | 0.6% |
| Argentina | 0.6% |
| Turkey | 0.6% |
| Chile | 0.6% |
| Austria | 0.5% |
| Finland | 0.5% |
| Hungary | 0.4% |
| Remaining countries | 3.7% |

**Table 2: Contribution to the HEP scientific publishing of several countries. Co-authorship is taken into account on a pro-rata basis, assigning fractions of each article to the countries in which the authors are affiliated. The last cell aggregates contributions from countries with a share below 0.4%. This study is based on all articles published in the years 2005 and 2006 in the five HEP "core" journals, *Physical Review D, Physics Letters B, Nuclear Physics B, Journal of High Energy Physics* and the *European Physical Journal C* and the HEP articles published in two "broadband" journals, *Physical Review Letters and Nuclear Instruments and Methods in Physics Research A.* A total sample of about 11'300 articles is considered. From reference [7].**

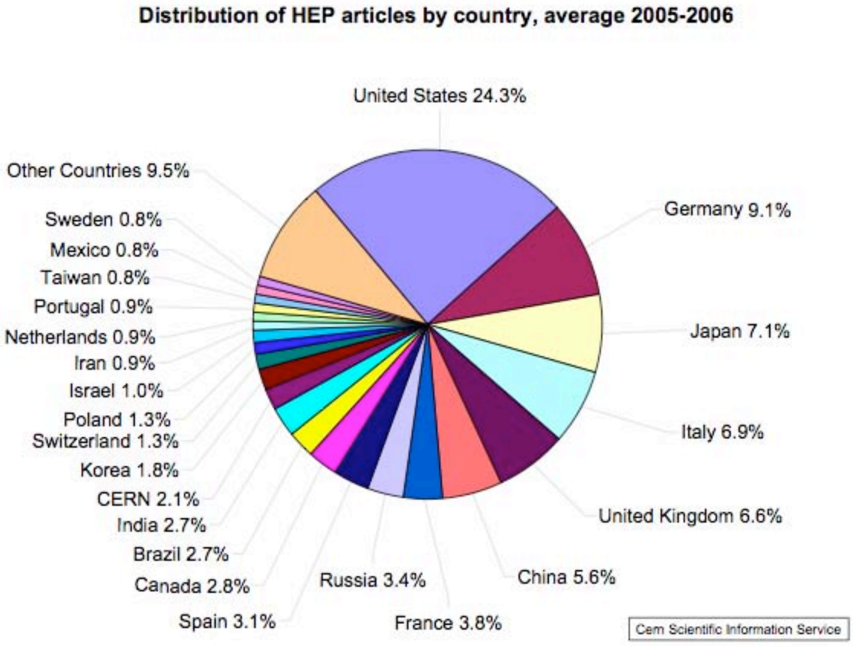


**Figure 3: Contribution to the HEP scientific publishing of several countries. Co-authorship is taken into account on a pro-rata basis, assigning fractions of each article to the countries in which the authors are affiliated. The last cell aggregates contributions from countries with a share below 0.4%. This study is based on all articles published in the years 2005 and 2006 in the five HEP "core" journals, *Physical Review D, Physics Letters B, Nuclear Physics B, Journal of High Energy Physics* and the *European Physical Journal C* and the HEP articles published in two "broadband" journals, *Physical Review Letters and Nuclear Instruments and Methods in Physics Research A*. A total sample of about 11'300 articles is considered. Contributions from countries with a share below 0.8% are summed in the slice denoted as "Other Countries". From reference [7].**

## 6 Conclusions and Outlook

At the time of writing, SCOAP$^3$ is an initiative emanating from leading European funding agencies, the two largest HEP European laboratories, and national and international library consortia. The fundamental pillar of the SCOAP$^3$ model is the federation of HEP funding agencies and library consortia worldwide. HEP is the most global of the scientific enterprises and the conversion to OA of its literature, with all the ethical, scientific and financial benefits it implies can only be achieved in a global co-ordinated process. A crucial step towards OA publishing in HEP is therefore the search for a world-wide consensus around the SCOAP$^3$ initiative, aiming to expressions of interest from HEP funding agencies and library consortia in Europe, the United States and beyond.

Once sufficient funds will have been pledged towards the establishment and the operation of SCOAP$^3$, a tendering process involving publishers of high-quality HEP journals will take place. Provided that the SCOAP$^3$ funding partners are ready to engage into long-term commitments, most publishers are expected to be ready to enter into negotiations along the lines presented in this article.

The outcome of the tendering process will allow the complete SCOAP$^3$ budget envelope to be precisely known and therefore the precise contribution expected from each country. A Memorandum of Understanding for the governance of SCOAP$^3$ will then be signed by funding agencies and leading national and international library consortia. Contracts with publishers will be established in order to make Open Access publishing in High Energy Physics a reality at the beginning of 2008, when the first experimental and theoretical publications of the CERN LHC program will appear.

The conversion of the HEP scientific publishing to the OA paradigm, along the lines presented in this article, will be an important milestone in the history of scientific publishing. The SCOAP$^3$ model could be rapidly generalized to other disciplines and, in particular, to related fields such as Nuclear Physics or Astroparticle Physics.

## Acknowledgements

This work summarises the report of the SCOAP$^3$ Working Party which, between December 2006 and April 2007, drafted the blueprint for the establishment, financing and operation of the consortium, with vision and dedication. I am grateful to all members of the Working Party for having shared their unique experiences towards making OA in HEP a success: S. Bianco, O.-H. Ellestad, P. Ferreira, F. Friend, P. Gargiulo, R. Hanania, S. Henrot-Versille, A. Holtkamp, P. Igo-Kemenes, D. Jarroux-Declais, M. Jordão, B.-C. Kämper, J. Krause, T. Lagrange, F. Le Diberder, A. Lemasurier, A. Lengenfelder, C. Lindqvist, S. Plaszczynski, R. Schimmer, J. Vigen, R. Voss, M. Wilbers, J. Yeomas, K. Zioutas. We have been constantly inspired by the OA vision of CERN Director General, R. Aymar.

## Notes and References

[1] http://oa.mpg.de/openaccess-berlin/berlindeclaration.html [Last visited 4 April 2007].

[2] VOSS, R., *et al.*, *Report of the Task Force on Open Access Publishing in Particle Physics,* CERN, 2006; http://library.cern.ch/OATaskForce_public.pdf.

[3] MELE, S., *et al.,* JHEP 12(2006)S01; arXiv:cs.DL/0611130.

[4] http://open-access.web.cern.ch/Open-Access/20050916.html [Last visited 4 April 2007].

[5] http://indico.cern.ch/conferenceDisplay.py?confId=482 [Last visited 4 April 2007].

[6] http://indico.cern.ch/conferenceDisplay.py?confId=7168 [Last visited 4 April 2007].

[7] BIANCO, S., *et al.*, *Report of the SCOAP$^3$ Working Party,* CERN, 2007; in preparation. To obtain a copy please contact Salvatore.Mele@cern.ch.

[8] http://www.soros.org/openaccess/read.shtml [Last visited 4 April 2007].

[9] http://slac.stanford.edu/spires [Last visited 4 April 2007].

[10] http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor/ [Last visited 4 April 2007].

[11] BLUME, M. Round table discussion: *Policy Options for the Scientific Publishing System in FP7 and the European Research Area*. Conference on Scientific Publishing in the European Research Area: Access, Dissemination and Preservation in the Digital Age, Brussels 15-16 February 2007.

# Importance of Access to Biomedical Information for Researchers in Molecular Medicine

*Annikki Roos; Turid Hedlund*

Information Systems Science, Department of Management and Organization, Swedish School of Economics and Business Administration, Arkadiankatu 22, 00100 Helsinki, Finland
e-mail: annikki.roos@ktl.fi; turid.hedlund@hanken.fi

## Abstract

In this paper, we analyze and describe the information environment of biomedicine from the point of view of the researchers in molecular medicine, which is a sub branch of biomedicine. We shall describe the nature of the discipline and its reflections to the information environment. A survey concerning the most important information resources in one molecular medicine research unit was conducted, and in this paper the main results of the survey is reported. The role of scholarly journals in the research process will also be analyzed. Special attention will be given to the possibilities of open access to the research process.

**Keywords:** information environment; information resources; databases; research process; molecular medicine

## 1 Introduction

The aim of this paper is to analyze and describe the information environment of biomedicine from the point of view of the researcher in molecular medicine (MM), a sub branch of biomedicine. Our target group is a research group containing researchers at different stages of their research career and the focus of study is on their daily work using information resources as part of the research process. The discipline is a rapidly growing and developing new research methods and processes which can be observed by the fact that pure laboratory work is to a growing degree transformed to computerized techniques. We argue that the change of the discipline from mainly laboratory based work to data based work has thoroughly changed the research processes. This has natural implications also to the information environment, as well as information retrieval, sharing practices and usage of information.

In this study the focus of research and our main research questions deal with the information environment of molecular medicine and firstly what are the main changes it has undergone. Secondly we investigate by conducting a survey, which are the most important information resources for researchers at different stages of their research career and thirdly what is the role of scholarly journals in the research process? For example, what is the publishing strategy and the criteria for choosing a journal to publish in.

We selected one research unit working in MM in Finland as a case. A web survey was conducted and qualitative information about researchers, their current work tasks, used information resources, publishing strategies and practices were gathered. A presentation and a feedback session concerning the results of the enquiry were given to the researchers. In this session important and explaining comments were given by the researchers in the target group about the use of information resources which have been taken into account when analysing and reporting the results of this study.

The outline of the paper is as follows: In Section 2, we describe the nature of the discipline and its reflections to the information environment. In Section 3, the effects of the changes in the environment will be analyzed against research process and scholarly communication practices. Special attention will be given to the experienced possible effects of open access in its different forms to the process. In Section 4, the results of the study are reported and in Section 5 we come to the conclusions and discussion.

## 2 Molecular Medicine as a Discipline

The discipline of biomedicine is growing exponentially. There are many factors behind the growth, of which the most important might be substantial increase in government support, the continued development of biotechnology industry, and the increasing adoption of molecular-based medicine. [1]. It has been pointed out in many sources that the nature of biomedicine has changed. It has transformed from laboratory based science to an information science, science “in silico”. [e.g. 2, 3, 4], which means mainly the computerization of the research process.

Specialization to different research domains, fields and sub-disciplines qualifies biomedicine. As Buetow felicitously remarks each of these “speak its own scientific dialect”. Like in many other scientific fields, “big science” (i.e. big budget, big staff, big machines etc.) is a growing challenge to the discipline. Research equipment and technology are extremely expensive and these are factors which have been leading researchers to work on teams. Biomedicine, according to Buetow is a “team science”. It is typical of biomedical research teams that many research problems in order to be solved have to cross traditional discipline boundaries. [1].

Molecular medicine, a sub-discipline of biomedicine is a practice oriented, applied science and utilizes molecular and genetic techniques in the study of the biological processes and mechanisms of diseases. It is highly reliant upon the development of techniques and technology for acquiring data. [5]. Its final, practical task is to provide new and more efficient approaches to the diagnosis, prevention, and treatment of a wide spectrum of congenital and acquired disorders [6]. The nature of MM, like biomedicine in general is interdisciplinary, it could also be seen as a hybrid of biomedicine and molecular biology. Molecular biology in turn is based on the combination of biochemistry, cell biology, virology and genetics [7].

## 3 Information Environment and the Changing Research Process

We define information environment in this study as the entity of information objects as well as the tools and services needed to retrieve, manage and analyze them.

A large volume of data in combination with the diversity of data types is typical for MM information environment. The characteristic of the data is that it is rapidly expanding and ever-changing. [1]. Most of the research databases, like genomic and proteomic databases are commonly updated and globally shared. A yearly updated list of online molecular biology databases is found in the website of Nucleic Acid Research [8].The January 2007 edition contained almost 1000 databases [9]. The amount of data growth could be described by for example the situation of the GenBank, a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation. It doubles in size about every 18 months. At the beginning of 2007, it contained over 65 billion nucleotide bases from more than 61 million individual sequences. [10].

What is even more challenging is that there is a need to integrate different kinds of data, e.g. to move between the biological and chemical processes, organelle, cell, organ, organ system, disease specific, individual, family, community and population. [1]. Like Butler notices, there are some disciplines which already have software that allows data from different sources to be combined seamlessly. For example, a gene sequence can be retrieved from the GenBank database, its homologues using the BLAST alignment service, and the resulting protein structures from the Swiss-Model site in one step. [11]

In parallel with the growth of data, the number of different tools, developed for data retrieval and analysis is growing. An actively maintained directory of bioinformatic links lists over 1000 web servers and other useful tools, databases and resources for bioinformatics and molecular biology research in 2006 [12, 13].

PubMed, the most important bibliographic database in biomedicine consisted in 2006 of 16 million references. The growth rate of the database is about 12 000 references every week, which means yearly over 600 000 new references. The growth curve of Medline, the main database in PubMed is illustrated in Figure 1. These lines describe the growth of traditional, published material, mainly in article format in biomedicine in a condensed way. It seems that inside the growing domain, there are some really “hot topics” where the amount of literature increase is extreme.

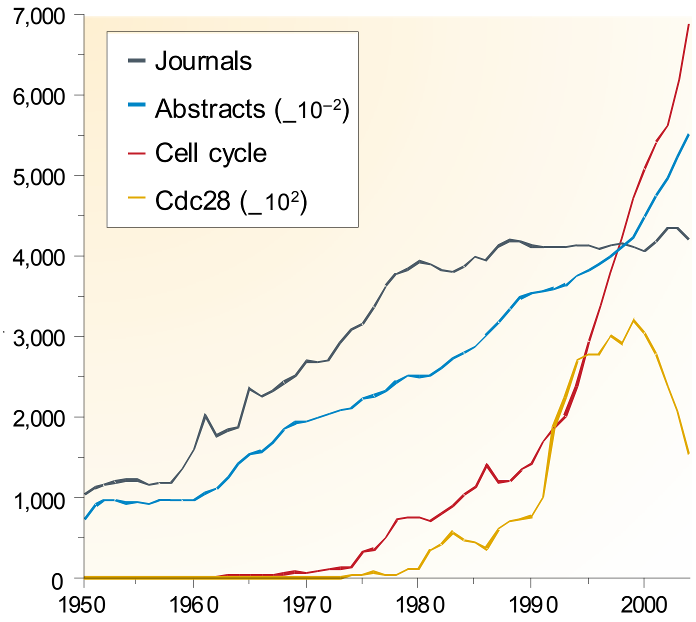


**Figure 1: Growth of Medline: the number of journals, abstracts, papers on the cell cycle and papers on Cdc28 [14 published in Nature Reviews Genetics]**

The typical features of MM information environment could be concluded as large volume and constantly growing number of data and published material, diversity of data types, great number of retrieval, analysis and other tools, interdisciplinary and globally shared and updated environment and team work. This is a fertile ground for the creation of new knowledge and inventions, but the lack of integration constitutes an increasing challenge to the development.

Cannata et al. have urged the organization of bioinformatics resources; data, knowledge, computational resources and services as a solution to the disintegration. They talk about "bioinformatics resourceome projcet" which would mean a process of creating a distributed system for describing resources, announcing their availability, and presenting this to the research community in an easy-to-navigate manner. The first step would be creation of an overall, distributed and collaboratively expandable ontology. [15, 16]. Mukherjea [17] has described the possibilities of using the semantic web in integrating the information resources. Grid technology has also been seen as a technical solution to the disintegration of data, information and tools. [1]

# 4 Results of the Survey

## 4.1 About the Research Unit and the Current Tasks

The research unit chosen as the case is situated in a Finnish research institute. As their aim, the unit declares to produce top level research in the molecular background of cardiovascular, immunological and neuropsychiatric diseases. At the moment of enquiry (February 2007), the unit consisted of 10 research groups with 83 researchers. From these 58 were PhD students and the rest were graduate students, group leaders and senior researchers We received totally 63 answers (75.9 %) to our web survey. 43 (68%) of those who responded were students and 20 (32%) were senior researchers, post docs and group leaders.

The research subjects of the groups were quite different, some of the groups concentrating on the genetic background of common diseases ("complex diseases"), some mainly to molecular genetics of monogenic diseases. There was also one bioinformatics group and one which specialized mainly in systems biology, one to quantitative genetics and a couple of groups mainly to the cell and molecular biology of certain diseases. We assume that the diversity of the research subjects caused some variety to reported work tasks between groups.

In the survey, all researchers were asked about their current work tasks and about information resources related to their current project or tasks and some information about usage of resources in general were asked. Respondents did get free spaces to write about their information resources, we gave only some examples for possible answers. We tried to get as broad a spectrum of possible resources, and did not want to limit or direct answers more than necessary. For current work tasks, we gave nine alternatives, from which it was possible to choose as many as were needed. Researchers were also able to add new tasks when necessary.

From the following figure (Figure 2.) the distribution of current tasks and their frequency among researchers is shown. The most common task among researchers was writing a report or an article, about totally 67 % of the researchers were doing it currently, the distribution among seniors and students is 70 % (seniors) and 67 % (students). Two-thirds of researchers were reading, 76 % of them were students. Of those working in the laboratory 74 % were students. It was more common (43 % of the respondents) to search information about literature from databases than data from data collections (25 %). Over one-third of the researchers were doing scientific computing. The researchers, who were studying the genetic background of “complex diseases” were practicing more scientific computing than most of the other groups. In two research groups where two-thirds (over 70 %) of all respondents answered that they were doing scientific computing.

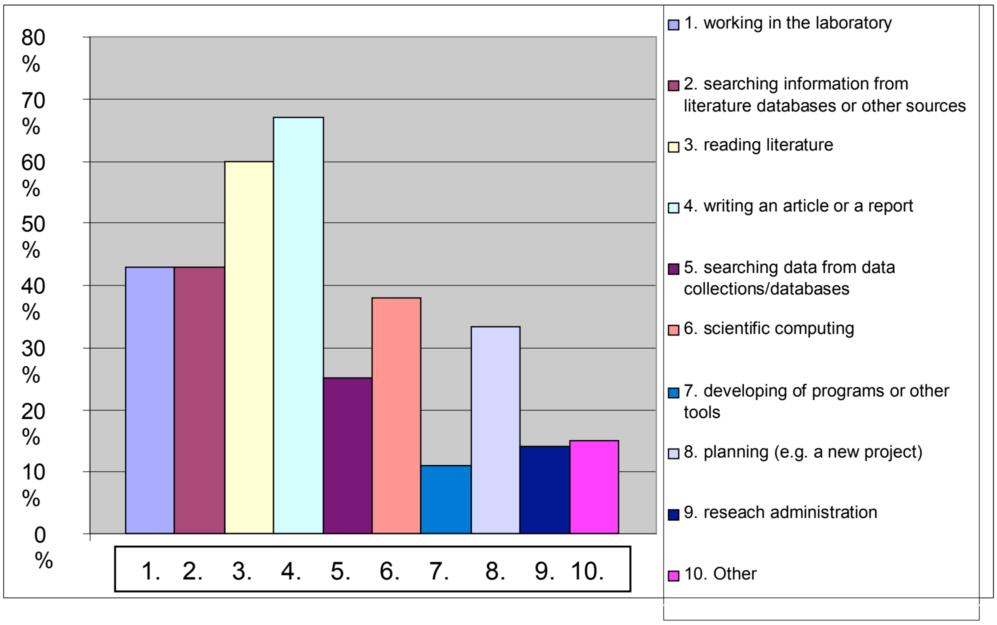


**Figure 2: Current work tasks of researchers**

## 4.2 Most Important Resources

When asked to choose at least the three most useful resources for their current research projects, doctoral and graduate (n=43) students named more resources than seniors and group leaders (n=20). PubMed got most references as the most useful resource in both groups. In the student’s group UCSC Genome Browser was second and Google third as the number of references are concerned. In the seniors’ group the ranking was contrary.

As their first information source 68% of the respondents named intranet/internet and in practice according to their answers, this means mainly PubMed and Google. 27 % of all researchers did prefer to contact a colleague or a supervisor. There seems to be no difference between students and seniors. In the feedback session researchers commented that the first information source depends on the nature of the issue: in practical questions and problems a colleague is preferred. It might also be possible, that some personal characters of the group leaders might at least partially explain the difference. The results indicate that in certain groups more researchers than on average in the groups favoured contacting a colleague in the first place. However, this is a speculation and needs to be observed more thoroughly.

When asked about which published material they use, the majority of respondents (53 %) answered that they use only or dominantly articles. 35 % of the researchers responded that they used articles and books equally and the rest 12 % named articles, databases and also some books.

When asked to name journals that researchers follow regularly, 23 % of the respondents reported that they do not follow any particular titles, rather their own topic from the literature databases. All of these respondents were graduate and doctoral students. Almost all graduate students belonged to this group.

91 % of the researchers said that they had used data collections during their current project. Those who did not use were juniors, who had recently started research work or researchers who were at the moment mainly working in the laboratory and writing articles. The problem with the reported data resources was that, because the question was open, researchers' answers were at very different levels. Some of them named quite general data collections, like "protein databases", or merely services or portals, like Entrez, while there were also respondents who used the detailed names of the databases or services. Totally 43 different data resources or services were named. The most common were NCBI and Entrez databases from National Centre for Biotechnology Information (by NIH and NLM) and UCSC Genome Bioinformatics resources, especially one tool, namely UCSC Genome Browser.

53 % or the researchers replied that they had used some research tools during their current project. The selection of tools and programs was also very diverse, from programs developed in their own laboratory to the commercial products. Totally 67 different tools were named. Students were naming more tools than seniors. The most often mentioned tool was Primer3, which is a PCR (Polymerase chain reaction) primer designer tool. The largest group in our survey as a whole was proteomics and sequence databases and analyzes tools.

It was noteworthy that tools for data mining seem to be common, but none mentioned text mining tools or tools for hypothesis creation. A tool called iHOP was familiar to the researchers. It's interesting, because it integrates gene and protein data from different collections with scientific literature.

Social bookmarking tools like Nature's Connotea were not named, neither any blogs. When asked why not, the answer in the feedback session was that they did not find those useful because their research problems were so specific: "they are only a waste of time". According to some opinions published in Nature researchers in general have not been eager to accept these tools because they might have been afraid of the poor image of the new tools and might have suspected the tools might damage their career [ see 18].

## 4.3 Role of Scholarly Journals in the Research Process

Writing and publishing articles in scientific journals are seen as an important part of the research process in biomedical sciences and molecular medicine. This is shown among others in [14] but also in this present case study of the research group on MM in Finland. When asked about their current work tasks about 67 % of the researchers in the case group answered that they were writing an article or a report.

Since the research group constitutes of senior researchers as well as doctoral and graduate students this can be seen as a high percentage. The amount of work and the importance of article writing is also to be seen in the results presented in Table 1., where we were asking the researchers questions about their publishing strategy for the coming year. All of the senior researchers and group leaders are going to publish at least 1 article, most of them (87.5%) are going to publish at least two articles and 75% of the group leaders and 43% of the senior researchers are planning to publish at least three articles. We have counted as main authors, the first and second author and the last author. In this case study most of the senior researchers and group leaders are acting as supervisors to younger researchers, why it seems appropriate that the last author is counted as important.

| | |
|---|---|
| Group leaders | 100% minimum 1 article as main author (1 & 2 or last) |
| Senior Researchers | 100% minimum 1 article as main author |
| Post doc | 83,3% minimun 2 articles as main author |
| Doct.students | 88% minimum 1 article as main author |
| Graduate stud. | 73,3% minimum 1 article as main author |

**Table 1: Publishing strategy regarding scientific articles of researchers for the coming year of the researchers in MM**

When looking at realized results (from 2006) for publications from the research group, 71 research articles in A-class journals and a total of 79 scientific articles were published. Of these 13 articles were in open access hybrid journals (applying some type of embargo) and 2 articles were in purely open access journals.

Regarding the choice of where to publish the researchers were presented the following criteria: impact, the speed of publishing, scope, open access or some other criteria, of which they were asked to name the one they regarded as most important.Impact was named as the most important by 58% of the researchers and scope by 39%. A few of the researchers named a combination of scope and impact. Open access as the main criteria was named by only 3% of the researchers.

The researchers were also asked to name journals with a suitable scope for publishing. On the top of the list of journals with suitable scope (Table 2.) was Nature genetics (named by 15). The impact factor for Nature genetics is also very high (25.797).

| Journal title | Number of nominations | Impact factor the journal |
|---|---|---|
| Nature genetics | 15 | 25.797 |
| Human molecular genetics | 11 | 7.764 |
| Molecular psychiatry | 10 | 9.335 |
| American journal of human genetics | 9 | 12.649 |
| European journal of human genetics | 6 | 3.251 |
| Nature | 6 | 29.273 |

**Table 2: Top listing of journals with suitable scope for publishing**

However, even though journals hold an established position in scholarly communication, there has appeared comments and viewpoints which have suggested that because scientific publications are slow and access to them is limited they act more as barriers to the development of new knowledge and science. [19].

In fact, traditional journals have very seldom made it possible to attach data files containing research data to the article. However, digital publishing and open access initiatives have opened up new possibilitities for scientific publishing (Björk 2007). In a study by Hedlund and Roos (2007) on publishing practices among biomedical researchers, the authors found that there is a growing rate of research publications in BioMed Central by Finnish researchers during the years 2003-2004. Cockerill & Tracz (2006) name fileds like bioinformatics, genomics and systems biology as possible success fields for open access.The initiative from the open access journal publishers BioMed Central is to put up a structured XML version of each full text article for data mining. There is also an increasing number of institutional repositories that allow researchers to upload data files linked to their published articles, which then serve as a possible source for data mining. Cockerill and Tracz (2006) argue that in the future the potential reader of a research article may not be only human beings but instead software agents looking for data to extracted and processed for a knowledge base. Therefore open access is important for work that involves multiple disciplines, as for example computer scientists, mathematicians and biologists collaborating in the areas of systems biology and bioinformatics.

## 5 Conclusions and Discussion

The information environment of researchers in MM could be summarized in the following diagram (Figure 3.)

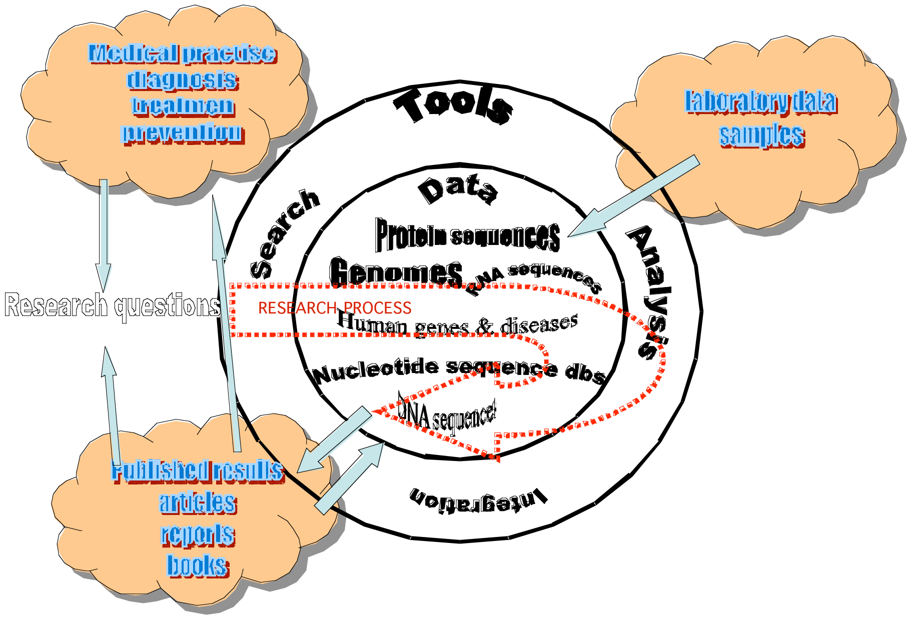


**Figure 3: The research process and the information environment of molecular medicine**

It can be concluded that access and use of data resources is an important and integral part of the research process in MM. The amount of different data collections, searching and analysis tools is huge. The disintegration of the environment seems also to be quite problematic.

We noticed that a more thorough analysis would be needed to make any conclusions about the relationship between the different work tasks in the research process and the used resources. We assume that many of the tasks might consist of several levels all of which might be worked out via different resources. The reason for this being for example in the varied complexity of the research problems.

The number of published articles is growing exponentially, especially in the “hot topics” of the domain. Researchers might find it difficult to follow even the development in their own research area. Maybe this is the reason why students do not follow particular journals, rather topics. The amount of literature is growing so fast that they are not able to do anything else than to follow the most recent and important articles from reference databases like PubMed. The disinterest to follow particular journals might also be due to the fact that they are not so well integrated into the domain yet, or it could be possible that their research subjects are so interdisciplinary that at least at the beginning of their career they are not able to follow any particular titles.

Journal publishing is still seen as the prominent way of distributing research results in molecular medicine. It has been shown in the case study that writing articles and reports is occupying the researchers as an important part of the research process. Even though many attempts to introduce open access, e.g. by providing institutional and national licences to cover authorship fees in BioMedCentral journals there still seems to be a strong reliance on traditional journals and especially journals with high impact.factors. Publishing in journals with high impact factor and the right scope is a strong base in the prevailing publishing strategy. However, it could be possible that the importance of traditional publishing channels and particularly articles might be on their way to change in the future if the text mining and hypothesis creation tools will be developed, and if the technical cyberinfrastructure with semantic web tools will be developed to integrate the environment. Open access will be helpful and a natural part of this development.

## Notes and References

[1] BUETOW, KH. Cyberinfrastructure: Empowering a "Third Way" in Biomedical Research. 2005:821-4.

[2] LENOIR, T. Shaping Biomedicine as an Information Science. *Conference on the History and Heritage of Science Information Systems*: Information Today 1999.

[3] LENOIR, T; ALT, C. Flow, Process, Fold: Intersections IN. In: Picon A, Ponte A, eds. *Science, Metaphor, and Architecture*. Princeton: Princeton University Press 2003:314-53.

[4] HINE, C. Databases as Scientific Instruments and Their Role in the Ordering of Scientific Work. *Social Studies of Science* 2006:269-98.

[5] MACMULLEN, W. Information problems in molecular biology and bioinformatics. Journal of the American Society for Information Science and Technology. 2005;56(5):447-56.

[6] GOOSSENS; M. Principles of molecular medicine. New England Journal of Medicine. 1999;340(20):1601-2.

[7] MIETTINEN, R; TUUNAINEN, J; KNUUTTILA, T; MATTILA, E. Tieteestä tuotteeksi? Yliopistotutkimus muutosten ristipaineessa. Helsinki: Yliopistopaino 2006.

[8] Nucleic Acids Research. Oxford Journals | Life Sciences | Nucleic Acids Research | Database Summary Paper Alpha List. 2007 [cited 10 April 2007]; Available from: http://www.oxfordjournals.org/nar/database/a/

[9] GALPERIN, M Y. The Molecular Biology Database Collection: 2007 update. Nucleic Acids Research. 2007;35(Database issue):D3.

[10] BENSON, DA; KARSCH-MIZRACHI I; LIPMAN D J; OSTELL J; WHEELER D L. GenBank. *Nucleic Acids Research* 2007:D21-5.

[11] BUTLER, D. Mashups mix data into global service. Nature. 2006;439(7072):6-7.

[12] UBiC. NAR Web Server Issue (July 1, 2006) - UBC Bioinformatics Centre. 2007 [cited 10 April 2007]; Available from: http://bioinformatics.ubc.ca/resources/links_directory/narweb2006/

[13] FOX, J A; MCMILLAN, S; OUELLETTE, B F. A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. *Nucleic Acids Research*: Oxford Univ Press 2006:W3.

[14] JENSEN, L J; SARIC, J; BORK, P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet. 2006 Feb;7(2):119-29.

[15] CANNATA, N; CORRADINI, F; MERELLI, E. A Resourceomic Grid for bioinformatics. Future Generation Computer Systems. 2007;23(3):510-6.

[16] CANNATA, N; MERELLI, E; ALTMAN, RB. Time to Organize the Bioinformatics Resourceome. *PLoS Computational Biology* 2005:e76.

[17] MUKHERJEA, S. Information retrieval and knowledge discovery utilising a biomedical Semantic Web. Brief Bioinform. 2005 Sep;6(3):252-62.

[18] BUTLER, D. Science in the web age: Joint efforts. Nature. 2005;438(7068):548-9.

[19] INSEL, T R; VOLKOW, N D; LI, T K; BATTEY, J F; LANDIS, S C. Neuroscience Networks. PLoS Biology. 2003;1(1):e17.

# Representing and Coding the Knowledge Embedded in Texts of Health Science Web Published Articles

*Carlos Henrique Marcondes* [1]*; Marília Alvarenga Rocha Mendonça* [1]*; Luciana Reis Malheiros* [2]*; Leonardo Cruz da Costa* [3]*; Tatiana Christina Paredes Santos*[4]*; Luciana Guimarães Pereira*[5]

[1] Department of Information Science
e-mail: marcon@vm.uff.br; e-mail: mariliaalvarenga@terra.com.br;
[2] *Department of Physiology and Pharmacology*
e-mail: malheiro@vm.uff.br
[3] *Department of Computer Science*
e-mail: leo@dcc.ic.uff.br
[4] *Biomedicine student*
e-mail: tatianacps.uff@gmail.com
[5] *Library and Information Science student*
e-mail: lucianaguipe@yahoo.com.br
Universidade Federal Fluminense
R. Miguel de Frias, 9 – Icaraí
24220-008 - Niterói – RJ Brazil

## Abstract

Despite the fact that electronic publishing is a common activity to scholars, electronic journals are still based in the print model and do not take full advantage of the facilities offered by the Semantic Web environment. This is a report of the results of a research project with the aim of investigating the possibilities of electronic publishing of journal articles both as text for human reading and in machine readable format recording the new knowledge contained in the article. This knowledge is identified with the scientific methodology elements such as problem, methodology, hypothesis, results, and conclusions. A model integrating all those elements is proposed which makes explicit and records the knowledge embedded in the text of scientific articles as an ontology. Knowledge thus represented enables its processing by intelligent software agents The proposed model aims to take advantage of these facilities enabling semantic retrieval and validation of the knowledge contained in articles. To validate and enhance the model a set of electronic journal articles were analyzed.

**Keywords:** electronic publishing; scientific communication; semantic web; knowledge representation; ontologies

## 1 Introduction

Nowadays, electronic Web publishing is a common activity to scholars and researchers. However, scientific communication is still a slow social process which largely depends on discourse, text producing, reading/interpreting/inquiring and peer-reviewing by scholars until new knowledge is incorporated into the corpus of Science. The potential of new information technology (IT) has been applied to modern bibliographic information systems to improve scientific communication, providing fast notification and immediate access to full-text scientific documents. But IT is not yet used to directly process the knowledge embedded in the text of scientific articles.

Semantic Web Initiative is a future vision of the Internet which aims to structure today's vast Web content, adding semantic to this content [1]. The technologies and methodologies that have been developed in the context of Semantic Web will enable this content to be understandable not only by people but also by software agents, enabling them to *reason* on this content in achieving different intelligent and useful tasks. In the Semantic Web context, electronic publishing can be a cognitive tool with potential that is far from being explored. Today electronic journals are still based on the print mode. Electronic Web published articles are knowledge bases, but for human reading.

Before the rise of the Web, what constitutes the accented scientific knowledge of humanity was fuzzy, lacks formalization, and was scattered across journals collections throughout libraries. Today there are two main barriers to a large scale use of this knowledge: the amount of information available throughout the Web and the fact that knowledge is embedded in the text of scientific articles in an unstructured way, not adequate for program processing.

Today, different scientific communities are developing Web ontologies which formally record the knowledge in a domain. W3C [2] defines ontology as "*a knowledge representation*". According to Jacob [3 p. 200] an ontology is "*a partial conceptualization of a given knowledge domain, shared by a community of users, that has been defined in a formal, machine-processable language for the explicit purpose of sharing semantic information across automated system*". In a near future, formal ontologies will be developed and recorded in program readable format, containing the accented knowledge in specific domains. Applying Semantic Web technologies to identify and record the knowledge embedded in the text of scientific articles in program-understandable format and compare it to the knowledge recorded in Web ontologies may be a key feature to the development of a future e-Science environment. Both these knowledge resources may be accessed by software agents on behalf of their owners, thus providing scientists with new tools to information and knowledge retrieval, to identify, evaluate and validate new contributions to Science.

The present research is looking for a new paradigm in scientific Web publishing: to publish not only text, for human reading, but also knowledge, formalized as ontologies, able to be processed by software agents. The objective of this research is to develop a Web publishing model which will be the basis for the future development of enhanced scientific authoring, publishing, retrieval and validating tools. These tools will enable the electronic publishing of scientific articles not only as texts for human reading, but also as a knowledge base in program-understandable format. The model aims to identify and record the semantic elements which constitute the knowledge embedded in the text of a scientific article.

What is the nature of scientific knowledge? This knowledge today, although recorded in digital format as Web published articles, are unstructured and not in adequate format for processing by software agents. According to Brookes [4 p. 131]: "*knowledge is a structure of concepts linked by their relations and information is a small part of such a structure*". Sheth [5 p. 1] states that "*Relationships are fundamental to semantics – to associate meaning to words, items and entities. They are a key to new insights. Knowledge discovery is about discovery of new relationships*". Miller [6 p. 306] answer these questions as: "*The above remarks imply-that science is a search after internal relations between phenomena".* Here scientific knowledge is considered as discovering relations between phenomena.

By the 16th century, a mark in the institutionalization of Science is the establishment of the scientific method as a procedure to achieve and communicate true statements in Science. A special element of scientific method is the hypothesis. As Scientific Methodologies handbooks emphasize, the role of hypotheses are central to Science in providing a provisory explanation to a phenomena and thus guiding the scientific inquiry. In the scientific method the hypothesis is the element which expresses a relation between phenomena.

Although a complex phenomena, scientific reasoning as expressed in the text of scientific articles must serve to an essential communicational role to Science as an institution: to validate the knowledge contained in the article, enabling any scientist to reproduce the steps taken by the author in his/her experiment. The need of this rigid protocol when communicating research results is stated by The International Committee of Medical Journals Editors, http://www.icmje.org:

> "*The text of observational and experimental articles is usually (but not necessarily) divided into sections with the headings Introduction, Methods, Results, and Discussion. This so-called "IMRAD" structure is not simply an arbitrary publication format, but rather a direct reflection of the process of scientific discovery*"

It is assumed here that knowledge in the text of articles – scientific methodology elements as the Problem, Hypothesis, Results and Conclusions – are all interrelated, constituting the content of the reasoning process developed by the author through which he/she communicates a new discovery. With the support of a Web authoring/publishing tool these semantic elements – the knowledge contained in the article -, can be identified, extracted and recorded in machine-understandable format, as an ontology. Knowledge thus recorded can be processed by software agents thus enabling semantic retrieval, consistence and validate checking. The ontology representing the knowledge extracted from the article can also be compared, matched and aligned to public Web ontologies which more and more represent the corpus of public knowledge in specific domains, thus enabling the establishment of formal relationship between both ontologies. Fails to establish these relationships may be evidences of new discoveries, since it can indicate that the knowledge in the article is not yet represented in the ontology which stores the accented knowledge in a specific domain.

## 2 Methodology

Building models is an important tool in Science. It enables Science to cope with complex phenomena such as scientific reasoning in communicating new discoveries through the text of scientific articles. An initial semantic model was developed, based on literature on Scientific Methodology, Philosophy and Epistemology of Science. Using the initial framework 53 articles on Health Science were analyzed with the aim of enhancing and validating the model. Articles were choose from two outstanding Brazilian research journals, 20 articles from the Memórias do Instituto Oswaldo Cruz, which scope is mainly Microbiology, http://www.scielo.br/revistas/mioc, 20 articles from the Brazilian Journal of Medical and Biological Research, http://www.scielo.br/revistas/bjmbr. Both are international journals using English as primary language. These journals were selected because initially we intended to interview authors personally. 14 additional articles about stem cells were analyzed too. Stem cells as an emerging research area in rapid development, was chosen expecting to find articles reporting important discoveries. Articles analyzed were selected from three recent reviews which present the stem cells research development in a historical perspective, promoting the advances in research, which was of special interest to this research. These reviews are "The Human Embryonic Stem Cell and the Human Embryonic Germ Cell", the official National Institute of Health (USA) resource for stem cells research, http://stemcells.nih.gov/, the article by Bongso et al. [7] and the article by Friel et al. [8].

The analysis simulates the tasks to be performed by an authoring/publishing tool when interacting with an author to identify and record the knowledge embedded in the text of an article. Scientific articles are highly conventional text types, with clear goal shared by authors and readers. Articles in Health Science are chosen for analysis due to their highly standardized structured, the so-called IMRAD – Introduction, Material and Methods, Results and Discussion - structure.

In order to explore the possibilities of using the model to identify new discoveries in Science, it is also verified if concepts found in the knowledge extracted from each article's text exist in a public knowledge base. DECS – Descritores em Ciência da Saúde - http://www.bireme.br/php/decsws.php, a Portuguese version of MeSH – Medical Subject Headings – http://www.nlm.nih.gov/MeSH/, and MeSH itself were both used in this experience in the role of a public knowledge base, with which subject headings found in the article's corresponding Lilacs (Latin America and Caribbean Literature on Health Science) or Medline database records are compared. MeSH is a component of UMLS - Unified Medical Language System -, http://www.nlm.nih.gov/pubs/factsheet/umls.html. It is a project of National Library of Medicine, USA, which aims to unify and encompass different medical specialized terminologies, thesaurus and classification schemas. UMLS evolves towards an ontology – the UMSL Semantic Network - in which concepts are organized in 134 classes or "semantic types" and 53 "types of relations".

The article analysis used the following form:

| | |
|---|---|
| **ARTICLE ANALYSIS FORM** | |
| **Journal: Memórias do Instituto Oswaldo Cruz** | **URL**: http://www.scielo.br/revistas/mioc |
| **Reference**<br>CAMARA, Geni NL, CERQUEIRA, Daniela M, OLIVEIRA, Ana PG *et al.* **Prevalence of human papillomavirus types in women with pre-neoplastic and neoplastic cervical lesions in the Federal District of Brazil.** *Mem. Inst. Oswaldo Cruz.* [online]. Oct. 2003, vol.98, no.7 [cited 10 March 2005], p.879-883. Available from World Wide Web: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0074-02762003000700003&lng=en&nrm=iso>. ISSN 0074-0276 | |
| **METHOD OF REASONING** | |
| **Deductive: X Inductive: Abductive:** | |
| **PROBLEM** (extracted from the text) | |
| As a contribution to the public health authorities in planning prophylactic and therapeutic vaccine strategies, we describe the prevalence of human papillomavirus (HPV) types in women presenting abnormal cytological results in Pap smear screening tests in the Federal District, Central Brazil.(Abstract)<br>In contrast to what is observed in developed countries, cervical cancer mortality in Brazil is still high. (Introduction) | |

| |
|---|
| **HIPOTHESIS – previous** (extracted form the text) |
| The chronic infection by certain types of human papillomavirus (HPV) is definitely related to the incidence of cervical cancer **(Lorincz et al. 1992, IARC 1995)** and the HPVs –16, -18, -31, -33, -35, -45, -51, -52, and -58 can now be considered as cervical carcinogenic agents **(Muñoz 2000)**. Squamous carcinomas and adenocarcinomas are the most frequent cervical neoplasias, and may develop from intraepithelial lesions, easily detected in preventive cytological exams **(Sherman et al. 1994)**. |
| **Normalized Relation**<br>HPV infection is related to the incidence of cervical pre-neoplasic and neoplasic lesions |
| **Antecedent:** HPV, different types / Papillomavirus Humano, |
| **Type fo relation:** causes / T147 UMLS Semantic Network |
| **Consequent**: cervical pre-neoplasic and neoplasic lesions / Infecções Tumorais por Vírus, Neoplasias do Colo |
| **Mapping to DECS: M (mapped)** |
| **DECS Subject Headings**<br>Papillomavirus Humano/classificação, Infecções Tumorais por vírus/epidemiologia, Neoplasias do Colo Uterino/virologia,<br>Papillomavirus Humano/genética, Infecções Tumorais por Vírus/patologia<br>Infecções Tumorais por Vírus/virologia, Neoplasias do Colo Uterino/diagnóstico<br>Doenças do Colo Uterino/patologia, Doenças do Colo Uterino/virologia<br>DNA Viral/genética, Esfregaço Vaginal, Reação em Cadeia da Polimerase<br>Polimorfismo de Fragmento de Restrição, Genótipo, Fatores de Risco<br>Prevalência |
| **Citations:** (Lorincz et al. 1992, IARC 1995), (Muñoz 2000), (Sherman et al. 1994). |
| **EXPERIENCE** |
| **Results** |
| **Measure:** prevalence |
| **Context:**<br>Environment:<br>Place: Distrito Federal, Brazil / Brasil/epidemiologia<br>Time:<br>Group: women / Humano, Feminino, Adulto, Meia-Idade |
| **Methodology:** |
| **Conclusions** |
| **Observations:** |

**Figure 1: Article Analysis Form**

## 3 Results

We envisage an authoring/publishing software tool which will be available to the author during the process of Web publishing his/her article, and interactively will capture the articles knowledge, recording it in a standard program readable format. This knowledge can then be retrieved and processed by semantic retrieval tools. Validation tools or software agents could also compare the knowledge extracted from articles with that held in public ontologies like the UMLS and thus indicate inconsistencies, faults and even new discoveries. The overall authoring/publishing environment is discussed in Marcondes [9] and illustrated in Figure 1. The authoring/publishing software tool development and how to identify new discoveries using the model proposed are in our agenda and will be object of future research. The present research is conceived only with proposing, testing and validating a model to the knowledge extracted from the article's text by a future authoring/publishing tool to be developed.

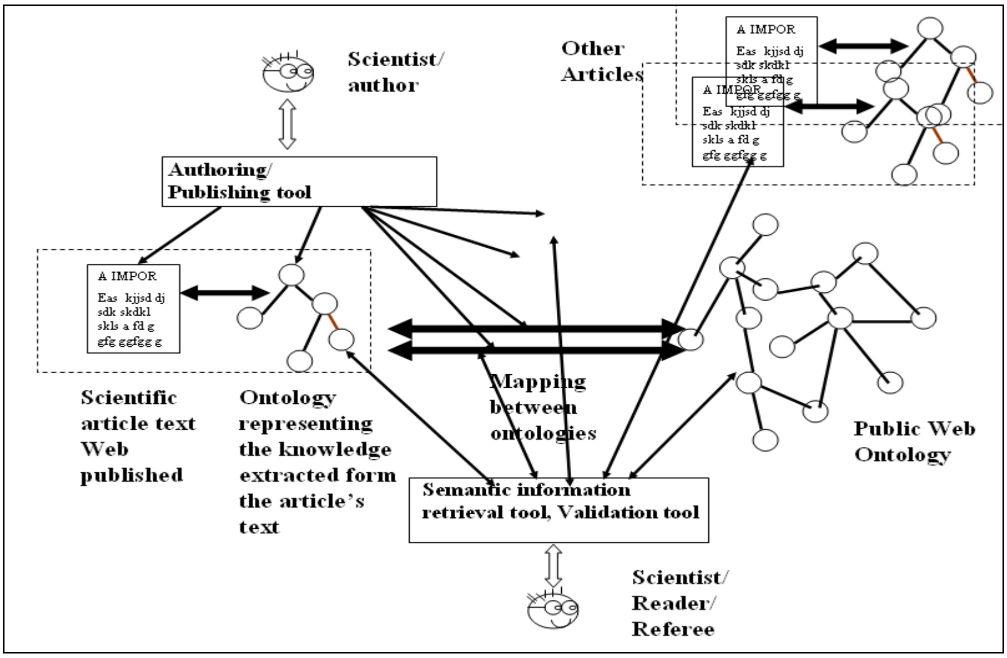

Scientist/ author
Other Articles
A IMPOR
Eas kjjsd dj sdk skdkl skls a fd g gfg ggfgg g
Authoring/ Publishing tool
A IMPOR
Eas kjjsd dj sdk skdkl skls a fd g gfg ggfgg g
Scientific article text Web published
Ontology representing the knowledge extracted form the article's text
Mapping between ontologies
Public Web Ontology
Semantic information retrieval tool, Validation tool
Scientist/ Reader/ Referee

**Figure 2: Author's editing/Web publishing environment**

What are the methods to achieve the truth in Science? These questions date back to Greek Philosophy with Epistemology, Rhetoric, Dialectics and Sophistic. Aristotle proposed patterns of reasoning from which true statements could be achieved from previous statements. He invented the reasoning method called *deduction,* through which particular statements can be derived from general statements. These patterns were systematized by Medieval Scholastics.

A branch of this discussion with important contributions came at the Modern Age, with the establishment of the scientific method by Francis Bacon [10]. In opposition to Medieval Scholastics, Bacon emphasized the importance of observational experiments to achieve general laws in Science. His reasoning method of deriving general statements from a particular number of observational cases was called *induction*. Besides all criticisms to the bases of the scientific method induction reasoning is still a strong basis to experimental Science.

Pierce adds to deduction and induction the abduction method of reasoning. According to him abduction is essentially the creative process of generation new explanatory hypotheses from apparently unstructured observational data. Pierce also integrated abduction with deduction and induction, proposing a whole method to scientific inquiry: a new hypothesis is abductively generated; its consequences are deductively inferred and inductively tested.

Abduction is considered as the logic of discovery by many researchers as Hoffmann [11], Magnani [12] and Paavola [13]. Pierces' example of abductive reasoning is Kepler discovery that planet orbits are not circles, as believed Copernic, but ellipsis. Abduction has always been associated with new discoveries both by Pierce himself and by researchers working on his legacy. Induction and Deduction are always associated with hypotheses testing and their ratification or refusal, an incremental increase to knowledge stock.

An article's knowledge - or semantic elements - appears according to the reasoning procedure employed by the author. It is important to identify these semantic elements to the development of an ontology which will guide a future authoring/publishing software tool while interacting with the author during knowledge extracting from article's text as a by product of the writing/publishing activity.

The article analysis showed three patterns of reasoning procedures. According to the reasoning procedure employed scientific articles can be classified as *theoretical articles*, which employ abductive reasoning and *experimental articles* which employ inductive or deductive reasoning. The elements complaining the structure of knowledge contained in the text of the article differs depending on the type of reasoning procedure used by the author.

These elements are: the PROBLEM the article is trying to address, the HYPOTHESIS, where the author states a RELATION between phenomena, a possible empirical controlled EXPERIMENT with the aim of observing the phenomena described, specific of experimental articles, divided in RESULTS – tables, figures, numeric data, reporting the observations made -, MEASURE used, a specific CONTEXT where the empirical observations take place, subdivided in ENVIRONMENT – a hospital, a crèche, a high school -, a geographical PLACE where the empirical observations take place, TIME when the empirical observations occurs, a specific GROUP – pregnant women, early born babies, mice - in which the phenomena occurs, and CONCLUSION – a set of propositions made by the author as a result of his/her findings.

Although all these elements are important to reasoning procedure, the hypothesis is the element which has the potential to hold new knowledge. The hypothesis has the form of a RELATION formed by two or more ARGUMENTS linked by a TYPE_OF_RELATION. In every article analyzed concepts found in the ARGUMENTS were tentatively mapped to concepts taken from the UMLS verifying if these concepts correspond to DECS/MeSH subject heading extracted from the article's record in Medline or Lilacs databases.

Theoretical-abductive model of articles are based on synthesis of Gross [14] and Hutchins [15] proposals. *Theoretical-abductive* articles analysis different previous hypotheses, show their faults and limitations and propose a new hypothesis; the reasoning is as follows:

> *a PROBLEM is identified, with the following aspects and data;*
> *the previous authors/HYPOTHESES are not satisfactory to solve the PROBLEM due to the following criticism;*
> *so, we propose this new HYPOTHESIS which we consider as a new pathway to solve the PROBLEM.*

*Experimental-inductive* articles propose a hypothesis and develop experiments to test and validate it; reasoning is as follows:

> *a PROBLEM is identified, with the following aspects and data;*
> *a possible solution to this PROBLEM can be based on the following new HYPOTHESIS;*
> *we developed an EXPERIMENT to test this HYPOTHESIS and it comes at the following RESULTS.*

In experimental-inductive articles, a CONCLUSION is one of the following types: or it corroborates the hypothesis, or it refuses the hypothesis or it partially corroborates the hypothesis. However in some cases, the CONCLUSION is neither the former, it just reports intermediate, not conclusive results toward the hypotheses corroboration.

*Experimental-deductive* articles use hypothesis proposed by other researchers cited by the article's author and apply it to a slightly different context; reasoning is as follows:

> *a PROBLEM is identified, with the following aspects and data;*
> *in literature the previous authors/HYPOTHESIS are proposed;*
> *we choose the following previous HYPOTHESIS;*
> *we enlarge and re-contextualize this HYPOTHESIS; we developed a EXPERIMENT to test it in this new context;*
> *the EXPERIMENT shows the following RESULTS in this new CONTEXT.*

Experimental articles also can compare various phenomena or hypotheses, as in a comparative study, a very usual type of article in Health Sciences. The different reasoning procedures can be formalized in an Ontology for Scientific Knowledge in Articles, as illustrated in Figure 2. This ontology has the following Classes and Properties:

```
Classes: THEORETICAL reasoning and
         EXPERIMENTAL reasoning
              Subclasses: INDUCTIVE reasoning and
                          DEDUCTIVE reasoning
Properties: PROBLEM
         HYPOTHESIS  (previous or new)
              Sub-properties: ANTECEDENT
                              TYPE-OF-RELATION
                              CONSEQUENT
```

REFERENCES (just in previous HYPOTHESIS)
EXPERIMENT
Sub-properties: RESULTS (quantitative data, tables, etc.)
MEASURE
CONTEXT
Sub-properties: SPACE
TIME
GROUP

Two Classes of articles were identified: Theoretical and Experimental. Experimental articles in turn have two Subclasses, Inductives and Deductives. The Properties of articles are the following: Theoretical-abductive articles have a PROBLEM, one or more previous HYPOTHESIS, that are discussed, criticized and rejected as solutions to the PROBLEM posed. So, the author proposes a new HYPOTHESIS which may be a solution to the PROBLEM. Theoretical-abductive articles do not present experimental results.

Experimental articles in turn always present experimental results. Experimental-deductive articles have the following Properties: a PROBLEM, one or more previous HYPOTHESIS, by different authors, that are adopted to guide an experiment. Previous HYPOTHESIS are extended, restricted or inserted in a new CONTEXT. An experiment is developed bases in the previous HYPOTHESIS applied to the new CONTEXT and the results of the EXPERIMENT are reported.

Experimental-inductive proposes an original new HYPOTHESIS to address a PROBLEM, develop an experiment to test this HYPOTHESIS and the results of the EXPERIMENT are reported.

HYPOTHESES have an ANTECEDENT, a TYPE-OF-RELATION and a CONSEQUENT. HYPOTHESES hold the knowledge embedded in the article as it proposes a relation between phenomena.

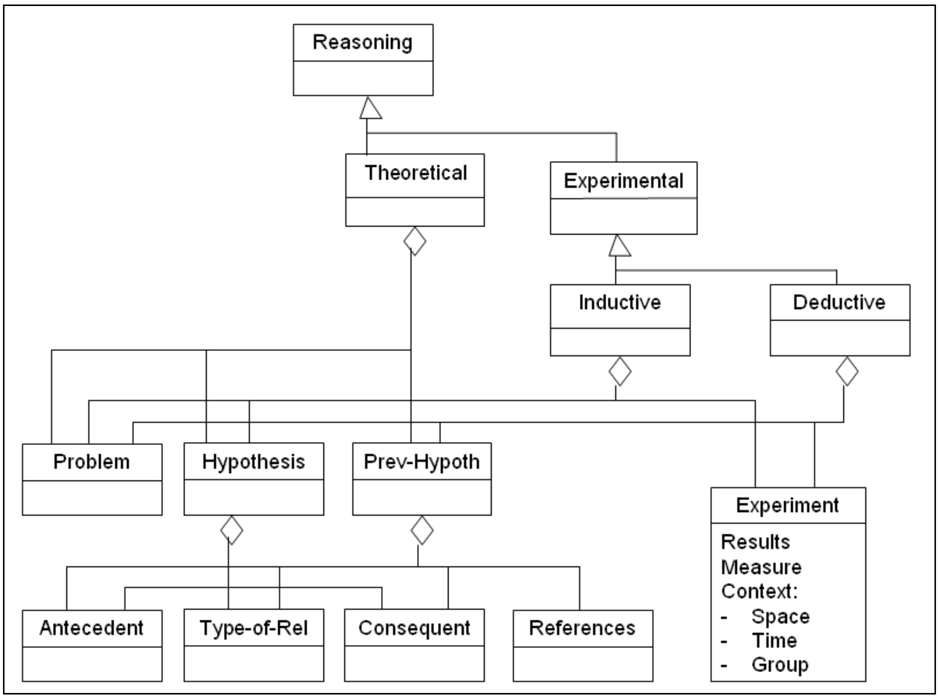


**Figure 3: Class diagram of the Ontology for Scientific Knowledge in Articles**

We plan to implement the Ontology for Scientific Knowledge in Articles in OWL [16]. The ontology will guide a future authoring/publishing tool in its interaction with an author to extract and record the knowledge embedded in the text of an article. Quantitative results of the analysis done on 53 articles are showed in Table 1. According to the classification proposed the majority of articles are experimental articles, 50 out of 53. Just 3 are theoretical-abductive articles.

| Articles analyzed | Exp-inductives | Exp-deductives | Theor-abductives | TOTAL |
|---|---|---|---|---|
| MIOC | 4 | 15 | 1 | 20 |
| BJMBR | 4 | 13 | 2 | 19 |
| STEM CELLS | 10 | 4 | 0 | 14 |
| TOTAL | 17 | 33 | 3 | 53 |

**Table 1: Results of the articles analysis**

In all articles the HYPOTHESIS was generally found in the Introduction section, in the Title or in the Abstract. Articles were considered Fully Mapped when concepts in both ARGUMENTs and the TYPE OF RELATION where fully mapped to one or more DECS/MeSH concepts that index the record in databases as Medline and Lilacs and there is a UMLS Semantic Network Relation corresponding to the TYPE OF RELATION. Articles were considered Partially Mapped when concepts in at least one of the ARGUMENTs or in the TYPE OF RELATION where fully mapped to one or more DECS/MeSH concepts and UMLS Semantic Network Relations. Articles were considered Not Mapped when any concept in neither the ARGUMENTs nor in the TYPE OF RELATION were fully mapped to DECS/MeSH concepts and UMLS Semantic Network Relations. The mapping of concepts to the DECS/MeSH is lower - which may be an indicative of new discoveries -, in a research area as stem cells in comparison to the two Brazilian journal. Table 2 shows these results.

| Articles analyzed | MIOC | BJMBR | STEM CELLS |
|---|---|---|---|
| Total of articles | 20 | 19 | 14 |
| Fully mapped | 11 | 4 | 0 |
| Partially mapped | 9 | 10 | 11 |
| Not mapped | 0 (0%) | 5 (25%) | 2 (7%) |

**Table 2: results of the mapping of concepts found in hypotheses to DECS/MeSH**

## 4 Discussion

The majority of articles found are experimental, 50 out of 53. The experimental articles all fit in the IMRAD model, with definite textual parts while the theoretic-abductive articles not. This fact may indicate a pattern of research characterized as “normal Science” according to Kuhn’s [17] theory.

Although foreseen in the literature only three theoretical-abductive articles were found among the articles analyzed. As this is the type of article which reports expressive paradigm changes in a scientific area it is expected that they are not very usual. But their existence is certain. For example, Watson and Crick article proposing a model to the DNA molecule is a typical theoretical-abductive article. All three articles found do not fit into the IMRAD structure. They do not have sections such as *Material and Method* and *Results*. Some review articles and letters to the editor have some traces of theoretical-abductive articles and must be object of future research.

Stem Cells potentialities constitute a new paradigm in cell biology. “*A new era in stem cell biology began in 1998 with the derivation of cells from human blastocysts and fetal tissue with the unique ability of differentiating into cells of all tissues in the body, i.e., the cells are pluripoten.*” (http://stemcells.nih.gov/). Since then two problems face the researches in the area: how to maintain stem cells cultures indefinitely undifferentiated in specialized cell types as bone, skin, liver, etc., and how to start and control differentiation into specific cells types. In the Stem Cells articles group there is a predominance of experimental articles reporting culture or control methods, in all of which the TYPE OF RELATION was mapped to relation “method” (UMLS Semantic Network T183). All articles of this group seem to report incremental advances in knowledge. Nome theoretical-abductive article was found in this group.

Few articles are totally mapped to DECS/MeSH concepts and to UMLS Semantic Network Relations. The process of mapping the concepts found in the ARGUMENTs and in the TYPE OF RELATION of each HYPOTHESIS is just a by-product of the data generated by the analysis process, just an explorative pathway to generate data for future research. In the majority of cases concepts in the ARGUMENTs were too specific in comparison to DECS/MESH concepts used to index the record. On the other hand the majority of TYPE OF

RELATIONs identified was satisfactorily mapped to UMLS "relations". This fact may be due to the difference in numbers: there are more than 730.000 concepts in UMLS and just 53 "relations". Relations are more stable across the time and more generic in comparison to concepts in a scientific area. Another explanation to this fact is that there is always a delay to these concepts be incorporated in the UMLS, so it is in dead an indicative of new discoveries. Anyway, operational results enabling software agents to compare the knowledge extracted from the text of articles to the knowledge record in Web ontologies according to the model proposed deserves more research.

The analysis performed shows that the scientific reasoning elements, according to the type of reasoning employed, are structured, forming an ontology, in the sense used in knowledge engineering, as in Sowa [18]. This enables a software agent to perform *inferences* on this structure. Based on the example analysis presented in Figure 1 knowledge extracted from articles, marked up and recorded as described would enable the following queries by a semantic information retrieval system:

> *- which other articles have hypotheses suggesting HPV as the cause of cervical neoplasias in women?*
> *- which articles have hypotheses suggesting other causes to cervical neoplasias different from HPV in women?*
> *- which articles have hypotheses suggesting HPV as the cause of cervical neoplasias in groups different from women?*
> *- which articles have hypotheses suggesting HPV as the cause of other pathologies different from neoplasias?*
> *- which articles have hypotheses suggesting HPV as the cause of cervical neoplasias in different contexts? (not in women from Federal District, Brazil).*

To publish scientific articles both as text and as machine readable knowledge bases seems to be a promising approach. It will enable the processing of this knowledge by software agents, thus improving critical inquiry, semantic querying and validation of scientific contributions to Science. Experimental Science, as Health Science, offer a solid basis to the development of the model, due to its formalism, derived from the use of the Scientific Method as an reasoning strategy in the text of scientific articles. The model outlined is a semantic model which aims to identify the semantic content of scientific reasoning. It is intended to be the basis to the development of a Web authoring/publishing tool. To reach this objective new research on computational techniques must be developed. We envisage an authoring/publishing tool that offers researchers/authors an interactive Web environment which, through a rich dialogue and using text extraction techniques, interactively identify and extract relevant contents of the article been written/published. This content will then be represented in machine-understandable format as an ontology, using OWL. Scientific articles so published throughout the Web can then be interlinked and linked to the increase number of Web ontologies, forming a rich knowledge network, thus enabling software agents to help scientist identify and validate new discoveries to Science. As the model proposed became more robust, there are plans to test it in other empirical science areas and even in areas as social sciences.

## 5 Conclusion

In all articles analyzed a relation expressing the mainly findings reported in the article was identified. This seems to indicate that scientific knowledge as expressed in the text of scientific articles can be represented as relations between phenomena. The amount of scientific knowledge now available throughout the Internet is so vast that it can only be processed with the aid of computer power. Here is proposed a standard representation to this knowledge feasible to be processed by software agents. This is essential if the intention is to use software agents to large scale processing of this knowledge in tasks as knowledge validation, semantic retrieval, identification and evaluation of discoveries.

Articles analyzed are very few and restricted to a single scientific area. If we are going to establish a new paradigm in electronic scientific publishing in which articles are published not only to human reading but also to be processed by software agents, this deserves more research. The model proposed is just a starting point to be discussed and enhanced by the scientific community.

Indexing language, as different Thesaurus largely used in information systems, select a set of concepts to describe a document. All knowledge organization effort is oriented toward the organization of systems of concepts. Generally all these concepts play an identical role when representing and retrieving a document. Although relations play a key role in scientific knowledge conventional indexing languages play no attention to them. Indexing language to no express the relations held between the subject headings indexing a document.

Indexing language must include relations between subject headings. There is also a need of the development of a taxonomy of relations used in Science to help indexing/retrieval scientific articles.

The model proposed also points to the standardization of a SkML - Scientific Knowledge Mark up Language - encompassing the semantic content of scientific articles Web published. This article highlights the benefits of a semantically richer format to represent the knowledge in scientific articles. With the aid of adequate software tools, this knowledge can be extracted as a by-product of authoring/publishing an article by the author. This opens an all new perspective in scientific knowledge acquisition, storage, processing and sharing.

## Notes and References

[1] BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. *Scientific American*, May, 2001. Available from Internet: <http://www.scian.com/2001/0501issue/0501berners-lee.html>.

[2] *OWL Web Ontology Language Guide.* Available form Internet: <http://www.w3c.org/TR/2004/REC-owl-guide-20040210/>.

[3] JACOB, E. K. Ontologies and the Semantic Web. *Bulletin of the American Society for Information Science and Technology*, Abril/May, 2003.

[4] BROOKES, B. The foundations of Information Science. Part I. Philosophical aspects. *Journal of Information Science*, vol. l2, 1980, pp. 125-133.

[5] SHETH, A; ARPINAR, I. B.; KASHYAP, V. Relationships at the heart of semantic web: modeling, discovering and exploiting complex semantic relationships. In: NIKRAVESH, M. Et al. *Enhancing the power of the internet studies in fuzziness and soft computing*. Springer-Verlag, 2002. Available from Internet: <http://cgsb2.nlm.nih.gov/~kashyap/publications/relations.pdf>.

[6] MILLER, D. L. Explanation Versus Description. Philosophical Review, vol.. 56, no. 3, May, 1947. pp. 306-312. doi:10.2307/2181936.

[7] BONGSO, A; RICHARDS, M. History and perspective of stem cell research. Best Practice & Research Clinical Obstetrics & Gynaecology, vol. 18, no. 6, 2004. pp. 827-842.

[8] FRIEL, R; SAR, S; MEE, P. Embryonic stem cells: Under standing their history, cell biology and signalling. Advanced Drug Delivery Reviews, vol.57, no.13, 2005. pp. 1894-1903.

[9] MARCONDES, C. H. From scientific communication to public knowledge: the scientific article Web published as a knowledge base. In: Proc. 9th ICCC ElPub - International Conference on Electronic Publishing, Leuven, Belgium, 2005, 9, p.119-27. Available from Internet: <http://elpub.scix.net> .

[10] BACON, F. *Novum Organum*. São Paulo : Abril Cultural, 1973.

[11] HOFFMANN, M. Is there a "Logic" of Abduction? In: Proc. 6th. Congress of the IASS– AIS International Association for Semiotics Studies, Guadalajara, Mexico, 1997. Available from Internet: <http://www.unibielefeld.de/idm/personen/mhoffman/papers/abduction-logic.html>.

[12] MAGNANI, L. *Abduction, Reason, andScience: Processes of Discovery and Explanation*. New York : Kluwer Academic; Plenun Publishers, 2001.

[13] PAAVOLA, S. Abduction as a Logic and Methodology of Discovery: the Importance of Strategies. Foundations of Science, Vol.9, No. 3. November, 2004. p. 267-283. doi: 10.1023/B:FODA.0000042843.48932.25.

[14] GROSS, A. G. The Rhetoric of Science. Cambridge, Massachusetts; London: Harvard University Press, 1990.

[15] HUTCHINS, J. On the structure of scientific texts. In: Proc. 5th. UEA Papers in Linguistics, Norwich.. Norwich, UK: University of East Anglia, 1977. p.18-39.(Conference Proceedings). Available from Internet: <http://ourworld.compuserve.com/homepages/wjhutchins/UEAP/L-1977.pdf>.

[16] OWL- Ontology Web Language, a W3C standard language to represent ontologies. Available form Internet: <http://www.w3.org/2004/OWL/>

[17] KUHN, T. The structure of scientific revolutions. In: Foundations of the unity of Science, vol. 2. Chicago : the University of Chicago Press , 1970.

[18] SOWA, J. *Knowledge representation: logical, philosophical and computational foundations*. Pacific Grove, CA : Brooks/Cole, 2000.

# Automatic Content Syndication in Information Science:
# A Brazilian Experience in the Creation of RSS Feeds to e-journals

*Robson Lopes de Almeida*

Department of Information Science, Universidade de Brasília
Campus Universitário Darcy Ribeiro, Brasília, DF - Brasil
e-mail: rlalmeida@unb.br

## Abstract

This paper reports the partial results of an exploratory study which intends to develop a methodology for a Web feed-based aggregation content service to electronic journals in Information Science. Ten scientific e-journals were chosen as sample to demonstrate the potential of the Web syndication technology. These e-journals are supported by the Brazilian Electronic Journal Publishing System (SEER), adapted from the Open Journal Systems (OJS), an open source software for the management of peer-review journals, developed by the Public Knowledge Project (PKP). In this context, the present study describes the concepts of aggregation and content syndication in Web environments. Moreover, it discusses the possibilities, advantages and eventual barriers to the implementation of RSS applications concerned with electronic journals in Information Science, specially the ones supported by the OJS Systems.

**Keywords:** metadata aggregation; content syndication; electronic journal; RSS; web syndication

## 1 Introduction

With the advent of the so-called Technologies of Information and Communication (TICs), particularly the Internet, which stands out as its main exponent, a significant raise in the amount of information can be observed, and we are exposed to them in our daily life. These pieces of information, when they are not useless, end up leading to a real overload, which is harmful to the absorption of the contents that really interest us. This also causes a sense of discomfort to the majority of people.

In the early 90's, with the advent of World Wide Web – the graphic and multimedia part of the Internet – information started to be even more easily disseminated. Because of that, new contents have been added to the web disorderly. Nowadays, the raising amount of Internet-generated information is not an object for information scientists only, but also for researchers from several different areas of study. They have also been attentive to the effects caused by every kind of information overload.

It is true that the simplicity of the existing web publishing tools has been useful not only to those who produce but also to the ones who acquire information, offering dynamic and low-cost mechanisms in order to communicate new ideas. The expansion of the blogosphere phenomenon [1] is a proof of that. On the other hand, the fast dissemination of digital information has demanded close attention in relation to the quality of the content which is about to be published, and also discernment concerning its use; otherwise, we run the risk of having our precious time drastically wasted.

Although this problem can be considered a natural consequence of our "Information Society", historically, in the 50's, the first systems able to select relevant information to a certain user, considering his/her profile of interest, appeared. This concept is called Selective Dissemination of Information (SDI), created by Hans Peter Luhn, from IBM Corporation, in order to improve the alert services offered by libraries, documentation centers and specialized centers of documental information.

From this perspective, and considering the current Web chaos, we intend to deal with the concepts of "content syndication" and "content aggregation", which has became popular from an Internet technology standard that allows users to receive updates to Web-based content of interest, simply called RSS.

The use of RSS began about ten years ago, meeting Internet user's information needs, keeping people up to date with new and revised information without making them feel lost facing Web content overload.

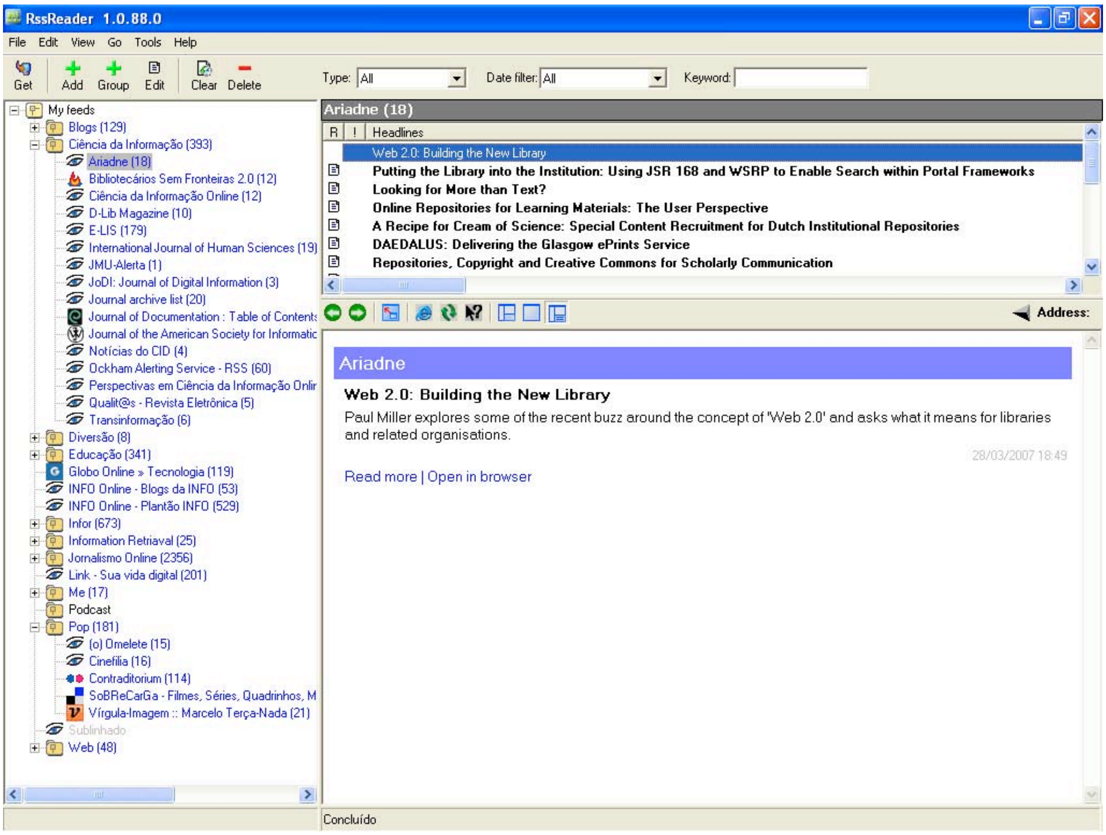
The present study intends to comment on the possibilities of how this resource can be used by electronic journals, especially the ones which already count on resources which make the implementation of RSS feeds [2] easier. In addition, it discusses the advantages to content publishers and to readers/users, and also the possible barriers to this implementation. Finally, this paper describes the progress of the first RSS feeds created by the author from a sample of Information Science e-journals supported by the Brazilian Electronic Journal Publishing System (SEER/OJS).

## 2 What is RSS?

Basically, the RSS format can be understood as a dialect or part of vocabulary from the XML family [3], meant for automatic capturing and website content distribution, used to publish frequently updated digital content, such as blogs, news feeds or podcasts. RSS allows Internet users to subscribe to websites that provide RSS feeds; these are typically sites that change or add content regularly. However, its applicability is not strict to these domains, once everything which is possible to be described by means of <tags> can be integrated by RSS.

The popularity of RSS technology is due to the agility which this format provides to the reading of new contents, since it does not need any access to websites where the information was originally published. In fact, the main feature of the RSS pattern is to allow a website´s frequent reader to track updates on the new issues of an e-journal, for example. Moreover, another advantage to the user is the facility of finding, in one single place, the current summaries of the main publications in the particular area. These characteristics called our attention to a deeper investigation concerning its use in Brazilian scientific journals on Information Science.

The most practical way of benefiting from this technology is having a news aggregator software, a type of application that retrieves syndicated Web content that is supplied in the form of Web feed. Such softwares are generally free, easy to install, and the great majority resembles an e-mail reader. Figure 1 presents a typical screen with one of these applications. It is possible to see in the left column all the chosen and added feeds, which can be read in the right column. On top of the right column, there is a list of headlines; while at the bottom, we can observe part or the complete post text. When double clicked in the headline title, the full content will appear in the inferior window, exactly as it was originally published in the Web. These headlines may be stored or deleted. There is also the possibility of filtering them by subject or date.



Type: All

Date filter: All

My feeds
Blogs (129)
Ciência da Informação (393)
Ariadne (18)
Bibliotecários Sem Fronteiras 2.0 (12)
Ciência da Informação Online (12)
D-Lib Magazine (10)
E-LIS (179)
International Journal of Human Sciences (19)
JMU-Alerta (1)
JoDI: Journal of Digital Information (3)
Journal archive list (20)
Journal of Documentation : Table of Contents
Journal of the American Society for Informatic
Notícias do CID (4)
Ockham Alerting Service - RSS (60)
Perspectivas em Ciência da Informação Onlir
Qualit@s - Revista Eletrônica (5)
Transinformação (6)
Diversão (8)
Educação (341)
Globo Online » Tecnologia (119)
INFO Online - Blogs da INFO (53)
INFO Online - Plantão INFO (529)
Infor (673)
Information Retriaval (25)
Jornalismo Online (2356)
Link - Sua vida digital (201)
Me (17)
Podcast
Pop (181)
(o) Omelete (15)
Cinefilia (16)
Contraditorium (114)
SoBReCarGa - Filmes, Séries, Quadrinhos, M
Vírgula-Imagem :: Marcelo Terça-Nada (21)
Sublinhado
Web (48)

Ariadne (18)

Web 2.0: Building the New Library
Putting the Library into the Institution: Using JSR 168 and WSRP to Enable Search within Portal Frameworks
Looking for More than Text?
Online Repositories for Learning Materials: The User Perspective
A Recipe for Cream of Science: Special Content Recruitment for Dutch Institutional Repositories
DAEDALUS: Delivering the Glasgow ePrints Service
Repositories, Copyright and Creative Commons for Scholarly Communication

Ariadne

Web 2.0: Building the New Library

Paul Miller explores some of the recent buzz around the concept of 'Web 2.0' and asks what it means for libraries and related organisations.

28/03/2007 18:49

**Figure 1: screenshot of RSS Reader, an stand-alone client aggregator**

Another way of getting the same function without having to install or configuring any kind of software is by using Web-based aggregator – a remote-hosted service offered by a third party that allows you to subscribe to and read feeds. To use this kind of free service, the user creates an account and then logs in to perform all feed-related activities, like reading a set of news sources in several XML-based formats. The user can find the news bits and display them in reverse-chronological order on a single page.

By means of these RSS "readers", it is possible to make a kind of subscription of the contents of different sources by themes, and quickly examine the title of the news articles and the summaries of a new issue in a condensed way. When the user finds some information which arouses his/her interest, the only thing he/she needs to do is to click on the title of the article to read its full content.

The name "RSS" is an umbrella term for a format that spans at least two different (but parallel) formats. Then, RSS could stand for "Rich Site Summary", "RDF Site Summary" or "Really Simple Syndication", depending on which version you are using. Regardless of what they are called or the version number, feeds are all XML-based languages. That is to say they are written to conform to the XML rules. For those who are familiar with HTML (Hypertext Markup Language) code, the structure of feeds will look familiar as we can see in figure 2. However, differently from HTML, which is limited to provide a universal format to represent information, without making reference concerning the structure and semantics of the data, RSS, as an XML-based language, is able to represent information about resources in the Web. It is intended to represent metadata about Web resources, such as the title, author, date of a webpage, copyright and licensing information about a Web document.

```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0">
  <channel>
    <title>Example e-Journal </title>
          <description>Short description of the journal </description>
          <link>http://www.cid.unb.br/</link>
          <language>pt-br</language>
          <lastBuildDate>Wed, 27 Mar 2007 14:02:11 -0300</lastBuildDate>
          <managingEditor>rlalmeida@unb.br</managingEditor>
          <pubDate>Wed, 17 Jan 2007 15:47:50 -0200</pubDate>
     <item>
       <title>Title of Article 1</title>
       <link>http://www.cid.unb.br/ e-journal/article_1</link>
      <description>
        Abstract of article 1
       </description>
    </item>
    <item>
      <title> Title of Article 1</title>
      <link>http://www.cid.unb.br/e-journal/article_2</link>
      <description>
        Abstract of article 2
      </description>
     </item>
</channel>
</rss>
```

**Figure 2: Example of a RSS feed (2.0 version)**

The use of specific tags, such as <title>, <link> and <description> allows us to treat each information unit (the title, the place where the information can be found and the summary) as a distinct object. This enables us to structure information so that it is interpreted and treated by means of computer resources, such as scripts or special softwares. This procedure turns data into qualified objects, such as attributes. This way, there is a possibility of automatic reusing of the information, making easier to share it with users and with other information systems (interoperability), and also organize it into database and do automatic research.

This way, it is simple to create a website reply to the content of another one which has an RSS feed. To do this, it is necessary to insert, in a page of the intended site, a script which points to an original website XML archive. These codes can be easily found in the Internet itself.

## 3 A Brief History of Web Syndication

In the late 90's, some experiences began, intending to provide required information to the Internet user in the Web context, such as Crayon [4]. Another initiative which was successful at that time was the resource called "Active Channel", developed by Microsoft to its browser Internet Explorer 4.0.

These two projects had in common the mission of joining in one place (aggregating) varied and scattered contents in the Web, by means of a technology called "push", since the idea was to send customized information to their users, instead of waiting for them to visit the websites to "pull" the desired contents. Specialized companies, such as the North American PointCast were pioneers in this type of syndication format, but the problem is that they were not interoperating; in fact they worked independently. Moreover, the softwares were too complex to users who were still getting familiar with the recently-created Web environment. At that time, Ramanathan Guha and other researchers from Apple Computer developed the Meta Content Framework (MCF), a specification of a format for structuring metadata about websites and other data.

At the beginning of March, 1999, when the research project was discontinued, Guha left Apple for Netscape Corporation, where he adapted MCF to use XML. He created the first version of the Resource Description Framework (RDF), which turned to be the basis for the creation of the first Web syndication format, called RSS (RDF Site Summary), 0.90 version.

In July, 1999, Dan Libby, also from Netscape, improved the format and produced a prototype tentatively named RSS 0.91 (RSS standing for Rich Site Summary) to be used in the "My Netscape" portal, as a pattern to the construction of headlines publishing systems in webpages, working as a summary of the content of a site with its respective links to the original information sources.

In the following year, as the group of developers from Netscape decided to leave the portal business, a lower-sized company called UserLand decided to keep on developing the RSS in order to apply it to their virtual electronic diaries tools, which, later, would become popular as weblogs or just blogs.

In August, 2000, another group of independent programmers (RSS – DEV Working Group), led by Rael Dornfest from computer book Publisher O´Reilly and Associates, proposed a new specification named RSS 1.0, according to the RDF metadata format. This one joined most of the preceding versions of RSS.

However, the group from UserLand, led by Dave Winer, continued their work, developing other versions of RSS, such as the 0.92 and 0.93, until they reached the version 2.0, in September, 2002. The abbreviation RSS had, then, another meaning: Really Simple Syndication, once its focus was on the simplicity in content syndication. Nowadays, this version is widely used by thousands of websites, including blogs and podcasts.

As Winer left UserLand, Berkman Center to Internet and Society, from Harvard University, was, then, in charge of the development of RSS 2.0, making this technology available to public domain, under a Creative Commons license, in 2003. In this same year, a group of leadering service providers, tool vendors and independent developers, worried about this problem of the development of the RSS specifications, decided to create a new format to content distribution: Atom [5] (originally called Echo). Its aim was to be 100% neutral, open and easily implemented by any developer. Atom is also based on the XML format, but its development is considered more sophisticated. According to specialists, it consists of a proposal of unification of RSS 1.0 with RSS 2.0, and it might be its natural substitute, since it counts on the support of great corporations, such as Google, which has adopted this format to its blog service. The final draft of Atom 1.0 syndication format was published in July, 2005, and was accepted by the IETF (Internet Engineering Task Force) as a "proposed standard" in August of 2005. The work, then, continued on the further development of the publishing protocol and various extensions to the syndication format.

In December, 2005, the Microsoft Internet Explorer team and Outlook team announced that they would be adopting the feed icon first used in the Mozilla Firefox browser, effectively making the orange square with white radio waves the industry standard for both RSS and related formats such as Atom.

In February, 2006, Opera Software announced they would also add the orange square to their Opera 9 release. Also in 2006, Microsoft decided to incorporate the RSS 2.0 extensions in its operational system Windows Vista, while Google announced the launch of a new content syndicated reader tool – the Google Reader – with support to RSS. Seven years later, the technical developments related to Web syndication seemed to be just beginning, with companies investing in new applications. But what stands out in the moment is the fact that the content providers and readers have found in Web syndication technology a fast and practical way of distributing and receiving updated contents through the Web [6].

# 4 Methodology

The sample chosen to carry out this research was the collection of Brazilian electronic journals in the area of Information Science which use the Brazilian Electronic Journal Publishing System (SEER), a tool applied to the administration of the editorial process of electronic journals, adapted from Open Journal System (OJS) to Portuguese language by the Brazilian Institute of Information in Science and Technology (IBICT/MCT), in 2004.

The preference for journals based on OJS is due to the fact that this is a consolidated system for publication and managing of peer-reviewed publishing. In March, 2007, over 900 titles were using OJS [7], including the main titles in Brazil, thanks to the effort of IBICT in offering specialized training to the editors. Another important factor determined the choice for these journals based on OJS: the system already has an RSS/Atom plug-in that produces Web feeds from articles that have been published in journals since the 1.x version. However, this feature is already included in recent releases of OJS 2.x.

Ten of these journals, having met the required criteria, were the object of several analyses, during the months of January and February, 2007. These analyses intended to investigate the main characteristics of the journals, according to the way of divulging its content, and, mainly, if they made Web feeds available to their clients by means of incorporating the RSS format in the publishing OJS tool, or even if it had at least an alert service, through electronic mail to notify the updating of the current edition.

Once the journals selected in this study did not present any kind of feeds, they were handed-created by means of an authorship tool called FeedForAll (http://www.feedforall.com) and were, later, hosted by the author´s webserver. The initial idea was to create a basis with the contents of those feeds, so that they could easily syndicate.

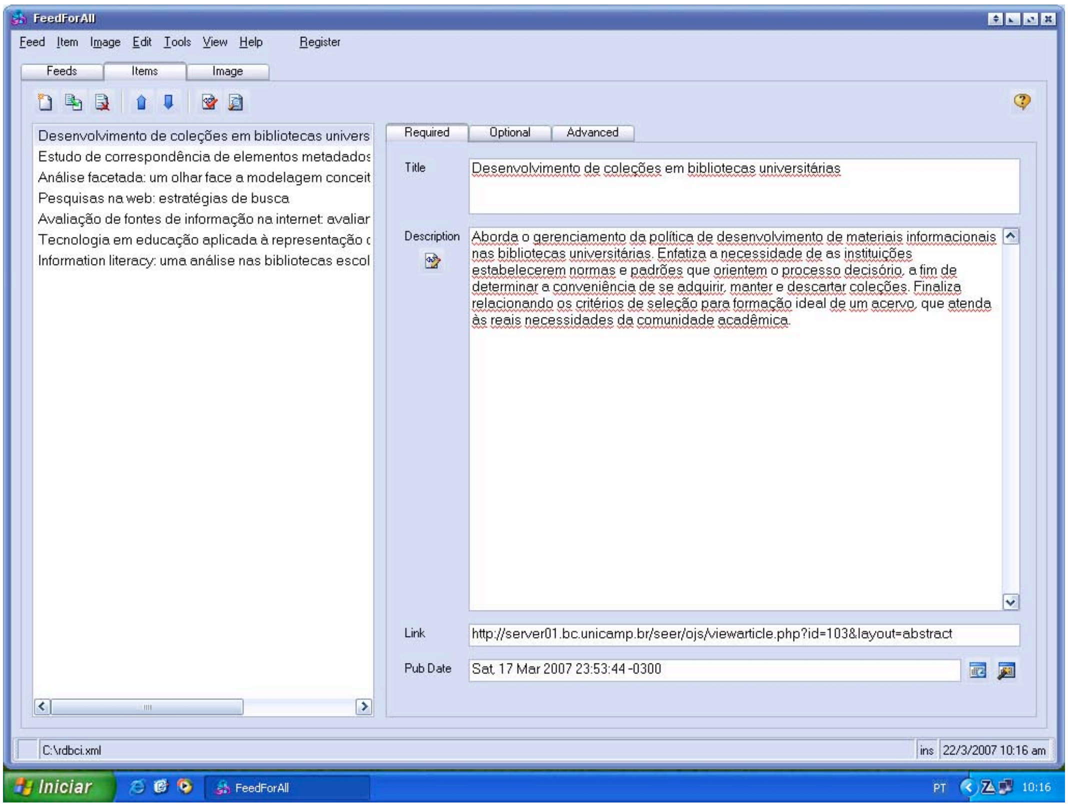
Although none of the sample journals presented feeds to their users, we could identify that there is at least one national journal that uses the OJS feed plug-in to generate RSS/Atom feeds automatically. It is called Qu@litas, an electronic journal edited by The Center of Applied and Social Sciences of the University of Paraíba.

After a testing period, these journal’s feeds were included into a content aggregator application specially created, using Netvibes service (http://www.netvibes.com), which provides a personalized page in which the author can manage several modules created from RSS/Atom feeds. Creating this “prototype” was a way of demonstrating the potential of a content aggregator application, starting from the simple process of creating Web feeds.

# 5 Results and Expectations

The RSS feeds created by the author to the 10 selected journals had as basis the last edition available in the website. The task of creating every feed lasted about 15 minutes, once it counted on the help of an RSS feed creation tool (FeedForAll), which made the creation of documents easier, without being necessary to write down codes which are particular of the RSS format.

Once the software is installed, it was quite simple to create RSS feeds. First, it is necessary to fill in the channel’s basic information: title, link (the URL of the webpage that corresponds to the channel) and description (a brief description of the content of the feed). Once the feed is created, it is necessary to add items. This task corresponds to the addition of metadata related to the articles. The indispensable information of each item are the same for creating a feed: the title of the article, the link (location of the page where the article can be found) and description (a summary of the article), as shown in Figure 3, which illustrates the filling of the required fields from “Items".


FeedForAll
Feed Item Image Edit Tools View Help Register
Feeds Items Image
Required Optional Advanced
Iniciar
FeedForAll
PT
10:16

Desenvolvimento de coleções em bibliotecas univers
Estudo de correspondência de elementos metadados
Análise facetada: um olhar face a modelagem conceit
Pesquisas na web: estratégias de busca
Avaliação de fontes de informação na internet: avaliar
Tecnologia em educação aplicada à representação c
Information literacy: uma análise nas bibliotecas escol

Title: Desenvolvimento de coleções em bibliotecas universitárias

Description: Aborda o gerenciamento da política de desenvolvimento de materiais informacionais nas bibliotecas universitárias. Enfatiza a necessidade de as instituições estabelecerem normas e padrões que orientem o processo decisório, a fim de determinar a conveniência de se adquirir, manter e descartar coleções. Finaliza relacionando os critérios de seleção para formação ideal de um acervo, que atenda às reais necessidades da comunidade acadêmica.

Link: http://server01.bc.unicamp.br/seer/ojs/viewarticle.php?id=103&layout=abstract

Pub Date: Sat, 17 Mar 2007 23:53:44 -0300

C:\rdbci.xml ins 22/3/2007 10:16 am

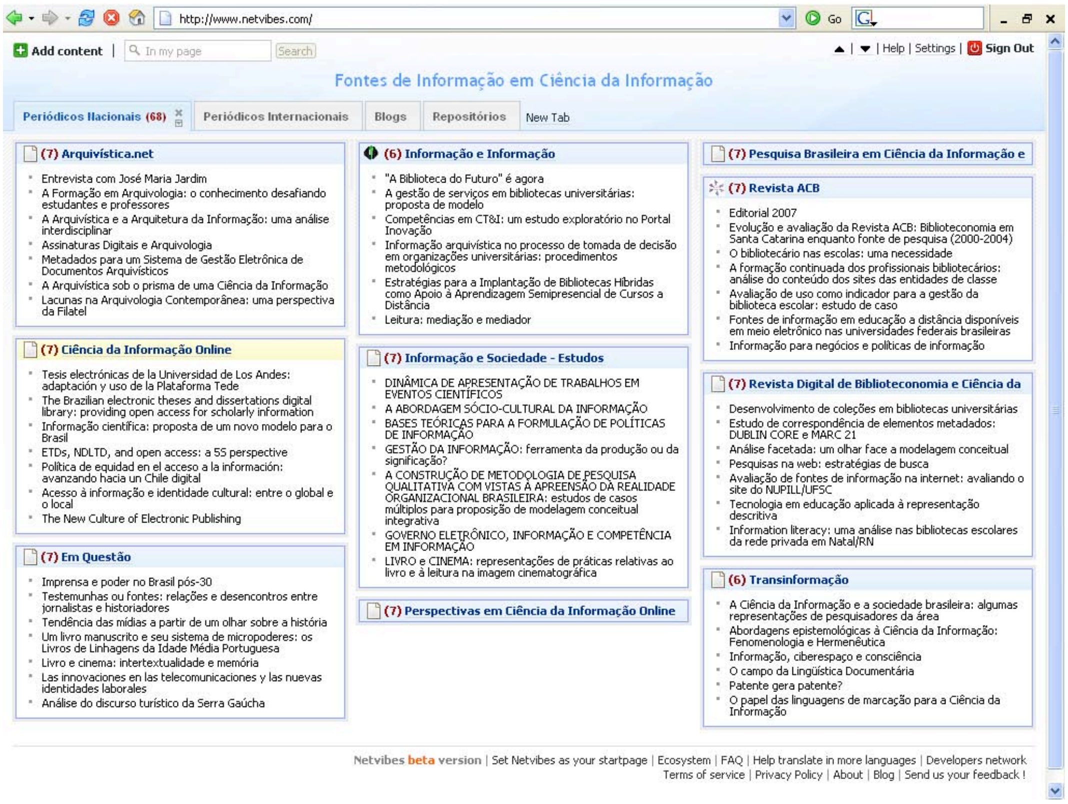
**Figure 3: Screenshot of RSSForAll tool**

In a second moment, we created a directory list containing the RSS feeds of all the periodicals being worked with, summarized in Table 1.

| Journal | Created RSS feed |
|---|---|
| Arquivística.net | http://www.rlalmeida.correiovip.com.br/arqnet/arqnet.xml |
| Ciência da Informação | http://www.rlalmeida.correiovip.com.br/cionline/ciinfo.xml |
| Em Questão | http://www.rlalmeida.correiovip.com.br/emquestao/emquestao.xml |
| Informação e Sociedade | http://www.rlalmeida.correiovip.com.br/ies/ies.xml |
| Informação e Informação | http://www.rlalmeida.correiovip.com.br/iei/iei.xml |
| Perspectivas em Ciência da Informação | http://www.rlalmeida.correiovip.com.br/pci/pci.xml |
| Pesquisa Brasileira em Ciência da Informação e Biblioteconomia | http://www.rlalmeida.correiovip.com.br/pbcib/pbcib.xml |
| Revista ACB | http://www.rlalmeida.correiovip.com.br/acb/acb.xml |
| RDBD | http://www.rlalmeida.correiovip.com.br/rdbci/rdbci.xml |
| Transinformação | http://www.rlalmeida.correiovip.com.br/transinfo/transinfo.xml |

**Table 1: Information Science Journals under SEER/OJS and its respective RSS feeds**

As a final result of this study, we propose a model of aggregation service on the specific content of Information Science e-journals based on Netvibes, a free service which uses the Ajax technology (Asynchronous Javascript And XML), in order to make the browser more interactive with the user, allowing him/her to create and manage models whose content come from Web feeds. It consists of an online service which was developed by means of XML and JavaScript. Once the feed(s) is(are) added, the application harvests the specific content and brings about, as a result, the titles and summaries of the updated articles. If the user wishes to access the whole content of any article, he/she will be sent to the correspondent page in the periodical itself, accessing the source or the document directly. The screen with all the aggregated periodicals can be seen in Figure 4.



Fontes de Informação em Ciência da Informação

(7) Arquivística.net

- Entrevista com José Maria Jardim
- A Formação em Arquivologia: o conhecimento desafiando estudantes e professores
- A Arquivística e a Arquitetura da Informação: uma análise interdisciplinar
- Assinaturas Digitais e Arquivologia
- Metadados para um Sistema de Gestão Eletrônica de Documentos Arquivísticos
- A Arquivística sob o prisma de uma Ciência da Informação
- Lacunas na Arquivologia Contemporânea: uma perspectiva da Filatel

(7) Ciência da Informação Online

- Tesis electrónicas de la Universidad de Los Andes: adaptación y uso de la Plataforma Tede
- The Brazilian electronic theses and dissertations digital library: providing open access for scholarly information
- Informação científica: proposta de um novo modelo para o Brasil
- ETDs, NDLTD, and open access: a 5S perspective
- Política de equidad en el acceso a la información: avanzando hacia un Chile digital
- Acesso à informação e identidade cultural: entre o global e o local
- The New Culture of Electronic Publishing

(7) Em Questão

- Imprensa e poder no Brasil pós-30
- Testemunhas ou fontes: relações e desencontros entre jornalistas e historiadores
- Tendência das mídias a partir de um olhar sobre a história
- Um livro manuscrito e seu sistema de micropoderes: os Livros de Linhagens da Idade Média Portuguesa
- Livro e cinema: intertextualidade e memória
- Las innovaciones en las telecomunicaciones y las nuevas identidades laborales
- Análise do discurso turístico da Serra Gaúcha

(6) Informação e Informação

- "A Biblioteca do Futuro" é agora
- A gestão de serviços em bibliotecas universitárias: proposta de modelo
- Competências em CT&I: um estudo exploratório no Portal Inovação
- Informação arquivística no processo de tomada de decisão em organizações universitárias: procedimentos metodológicos
- Estratégias para a Implantação de Bibliotecas Híbridas como Apoio à Aprendizagem Semipresencial de Cursos a Distância
- Leitura: mediação e mediador

(7) Informação e Sociedade - Estudos

- DINÂMICA DE APRESENTAÇÃO DE TRABALHOS EM EVENTOS CIENTÍFICOS
- A ABORDAGEM SÓCIO-CULTURAL DA INFORMAÇÃO
- BASES TEÓRICAS PARA A FORMULAÇÃO DE POLÍTICAS DE INFORMAÇÃO
- GESTÃO DA INFORMAÇÃO: ferramenta da produção ou da significação?
- A CONSTRUÇÃO DE METODOLOGIA DE PESQUISA QUALITATIVA COM VISTAS À APREENSÃO DA REALIDADE ORGANIZACIONAL BRASILEIRA: estudos de casos múltiplos para proposição de modelagem conceitual integrativa
- GOVERNO ELETRÔNICO, INFORMAÇÃO E COMPETÊNCIA EM INFORMAÇÃO
- LIVRO e CINEMA: representações de práticas relativas ao livro e à leitura na imagem cinematográfica

(7) Perspectivas em Ciência da Informação Online

(7) Pesquisa Brasileira em Ciência da Informação e

(7) Revista ACB

- Editorial 2007
- Evolução e avaliação da Revista ACB: Biblioteconomia em Santa Catarina enquanto fonte de pesquisa (2000-2004)
- O bibliotecário nas escolas: uma necessidade
- A formação continuada dos profissionais bibliotecários: análise do conteúdo dos sites das entidades de classe
- Avaliação de uso como indicador para a gestão da biblioteca escolar: estudo de caso
- Fontes de informação em educação a distância disponíveis em meio eletrônico nas universidades federais brasileiras
- Informação para negócios e políticas de informação

(7) Revista Digital de Biblioteconomia e Ciência da

- Desenvolvimento de coleções em bibliotecas universitárias
- Estudo de correspondência de elementos metadados: DUBLIN CORE e MARC 21
- Análise facetada: um olhar face a modelagem conceitual
- Pesquisas na web: estratégias de busca
- Avaliação de fontes de informação na internet: avaliando o site do NUPILL/UFSC
- Tecnologia em educação aplicada à representação descritiva
- Information literacy: uma análise nas bibliotecas escolares da rede privada em Natal/RN

(6) Transinformação

- A Ciência da Informação e a sociedade brasileira: algumas representações de pesquisadores da área
- Abordagens epistemológicas à Ciência da Informação: Fenomenologia e Hermenêutica
- Informação, ciberespaço e consciência
- O campo da Lingüística Documentária
- Patente gera patente?
- O papel das linguagens de marcação para a Ciência da Informação

Netvibes beta version | Set Netvibes as your startpage | Ecosystem | FAQ | Help translate in more languages | Developers network
Terms of service | Privacy Policy | About | Blog | Send us your feedback !

**Figure 4: Screenshot of Netvibes application joining the 10 Brazilian Information Science Journals**

The model for this service shows one of the main applications for Web feeds nowadays. In this case, we present the dynamics of a mechanism which is able to gather, in one single Web page, the indication of the Brazilian journals articles in Information Science (with their respective summary), by the use of Web Syndication. Since then, this service could be advertised to the community of potential users, mainly researchers, professors and people engaged in post-graduation courses on Information Science. Through these feeds provided by the directory, these users will be able to locate easily or, if they want, to subscribe to receive the updates of the publications they wish by means of the aggregation service they prefer.

Once all this content is gathered, it is possible to search every journal of this collection simultaneously. This way, if the user wishes to research the word "ontology", for instance, he/she will have as a result all the articles which contain the word "ontology" in its title or in the summary, no matter which journal that may be.

# 6 Discussion

## 6.1 Advantages

According to our perception, the advantages to the reader are great. We will be able to count on a powerful tool with which we will be able to keep ourselves up-to-date in relation to several sources of information and, at the same time, make simultaneous researches on relevant content, which enables us to refine this search, raising the relevance of the recovered terms.

Saving time reading practically personalized information is a great advantage of using these kind of services which are able to join, in one single environment, a variety of contents produced by several different sources, with no need to access each site individually. Another characteristic of this system is that when a certain topic is selected by the user, the RSS technology offers the possibility of showing the full content of the document, with direct access to the source. That is, there is no copy of information or inappropriate seizure. The publishers, however, will aggregate value to the content of their publication and, consequently, will gain visibility to their publication, once RSS feed allows their users to read their content without going out of their way to visit. While this may seem a flaw at a first glance, it actually improves the visibility by making it easier for users to see it the way they want to. Because there are so many sources on the Web, most viewers won't come to the same site every day. By providing a feed, publishers are in front of them constantly, improving the chances for them to click through to an article that catches their attention.

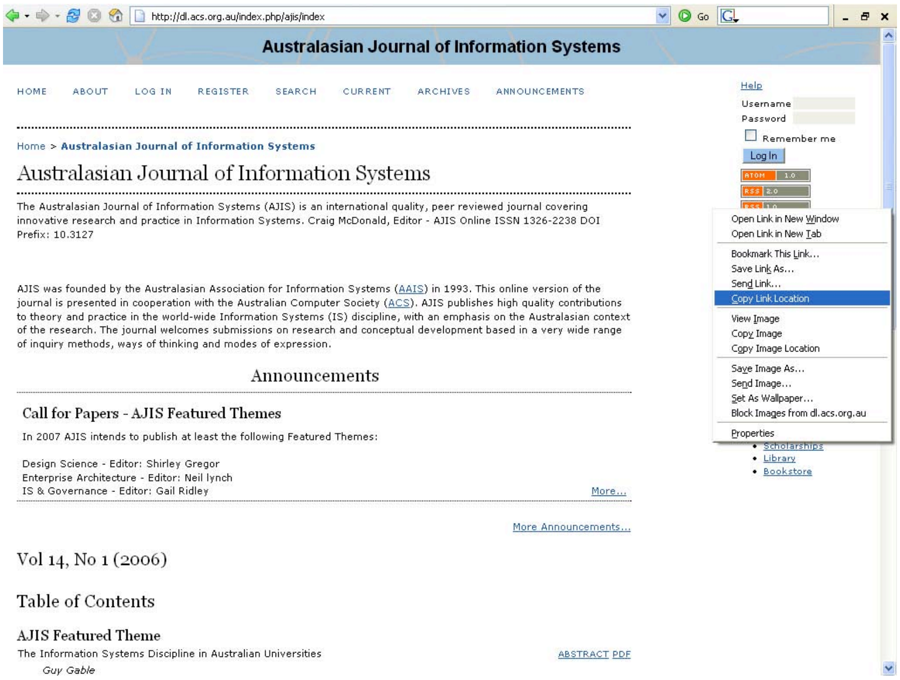
Nowadays, most journals which use OJS publishing system have a notifying service which offers the user the possibility of enrolling to receive, by e-mail, a notice with the summary of the new editions as they are published. If the user wishes to follow publications in a certain area, he/she will have to repeat this procedure for each journal. This means that he/she will receive several different notifications for each updating. Through a Web syndication service, the user does not need to enroll to keep updated, with the advantage of not having his/her mail box full.

## 6.2 Barriers

The main barrier for the implementation of this type of service seems to be the publishers' and users' lack of knowledge about the Web syndication technology. In Brazil, only the great newspapers which circulate around the country and some other specialized websites make feeds available to their users. Even in the academic context, with the exception of the courses on Computer Science, there is still ignorance of terms such as "feeds", "syndication" or "aggregator". In other countries, the adoption of RSS in information services is more common. However, we notice some resistance from the scientific editors. Even the group responsible for the development of the OJS admits that Atom/RSS feed plug-in is not a very well-known feature for their users, and few OJS e-journals make Web feeds available in their editions.

Differently from reading an e-mail, for instance, the current method of subscribing to a feed – copying the URL from the link (normally with .xml or .rss extension), and pasting it into a reader application – is not obvious to the new user. When doing this, the user normally sees XML codes on the screen. These are, at first, incomprehensible. It is normal for lay users to be confused with such information and, instead of subscribing the channel, they end up not doing that because they think they have committed some kind of mistake.

In the case of the feeds generated automatically through OJS plug-in, the subscription process is even more difficult. Once the desired format is chosen (Atom or RSS 1.0/2.0) with a click on the respective icon on the main page, the browser will ask the user to indicate an application to open an unknown document format. In order to avoid this discomfort, it is necessary to click the right button of the mouse on the icon which represents the chosen format, and select the option "Copy Link Location" (Figure 5) to, later, paste it into the reader application. This important piece of information, however, is not available in the journals analyzed. These are pretty user unfriendly and, probably, will be a barrier to widespread the adoption of RSS by 'non-techies'.



Australasian Journal of Information Systems

The Australasian Journal of Information Systems (AJIS) is an international quality, peer reviewed journal covering innovative research and practice in Information Systems. Craig McDonald, Editor - AJIS Online ISSN 1326-2238 DOI Prefix: 10.3127

AJIS was founded by the Australasian Association for Information Systems (AAIS) in 1993. This online version of the journal is presented in cooperation with the Australian Computer Society (ACS). AJIS publishes high quality contributions to theory and practice in the world-wide Information Systems (IS) discipline, with an emphasis on the Australasian context of the research. The journal welcomes submissions on research and conceptual development based in a very wide range of inquiry methods, ways of thinking and modes of expression.

Announcements

Call for Papers - AJIS Featured Themes

In 2007 AJIS intends to publish at least the following Featured Themes:

Design Science - Editor: Shirley Gregor
Enterprise Architecture - Editor: Neil lynch
IS & Governance - Editor: Gail Ridley

Vol 14, No 1 (2006)

Table of Contents

AJIS Featured Theme

The Information Systems Discipline in Australian Universities
*Guy Gable*

**Figure 5: Screenshot of an OJS Journal which makes Web feeds available. The procedure for subscribing the channel is not informed in the website.**

## 6.3 In the Future

In a near future, the Web Syndication technology might be widely known and broadly used by the ones who publish and the ones who acquire information in the Web. The availability of Web feeds in the periodical publications, no matter its nature may be, will be so natural that the user will not find it strange to meet indicative icons that a certain website or blog has an RSS channel. The process of subscription of a feed might be an extremely simple task and also transparent to the user. It will not be necessary to do copy&paste from the link which corresponds to the XML file, once the browsers will have efficient mechanisms to read feeds [8].

Independently of its format and version (RSS, Atom or even another one which may be created), the reader will learn, little by little, how to deal with a “power” that is hard to be imagined up to now: personalize the content of the information he/she wants to acquire, as well as produce new contents, through the use of resources offered by syndication technology, without the intervention of intermediates.

This way, when reading an article about “Semantic Web”, for instance, the user may, if he/she wants to, subscribe to the Web feeds about the same theme and the ones which are syndicated from other articles, data base, repositories, and so on. Everything turns out to be interoperating, thanks to the compatibility provided by the applications based on the XML language and its derivatives. From the user’s point of view, there will not have distinction among journals A, B or C in a certain area anymore. To him/her, it is as if there were one single source of information, easily available in his/her “personal digital library”.

In the next five years, there may be an explosion of new Information Retrieval Systems (IRS), with Web feed resources. This way, once the results have been recovered from a research in a database, users have the option of subscribing the feed related to that expression for search, and, so, keep themselves updated in relation to that specific matter. One of the existing services which follows this model is the one offered by the Agência Brasil (http://www.agenciabrasil.gov.br), the Brazilian government news agency which allows the user to create new channels based on their searches, more than offering RSS feeds for more than 120 different subjects.

The scope of this type of service is great and it is a good example that its potential is being tested by Ockham Alerting Service (http://alert.ockham.org), a current awareness service based on the National Science Foundation Digital Library content. According to their website, it demonstrates a standard-based method for collecting content, providing access to it and disseminating it on a regular basis in the form of an alerting service. The method includes: a) identifying OAI repositories with content of interest; b) using OAI to harvest content and store it in a central pile; c) indexing the content of the central pile; d) providing an SRU interface [9] to the index; e) allowing users to save the SRU URL's as "profiles" (RSS feeds); f) allowing users to have the profiles executed on a regular basis; g) making the results of searches available as HTML, e-mail, RSS, etc.

# 7 Conclusions

Despite the barriers identified in this study, we believe that the Web syndication technologies are viable to the integration of any Web-based information system, from search engines to publication systems, such as the case of OJS, presented in this study.

The reach of the Web syndication resources goes beyond the management of Internet content. It can also be a useful way of marketing, for instance, or even serve to notify users of the status of projects, monitor web statistics or otherwise replace the "notifications" that are now sent out as e-mail alerts.

Concerning the electronic journals, this type of technology seems to be welcome, once the simple inclusion of an RSS feed may aggregate a great value to the publication, not only as an intelligent way of divulging, but mainly because of the possibility of integrating with other contents, thanks to the interoperability which exists among the formats compatible with XML language.

This way, it seems that the new products and information services will have even more relationship with RSS/Atom feeds. The OJS feed plug-in is a proof of that. Before, it was available as an external archive to the OJS 1.x, and nowadays it includes recent releases of OJS 2.x.
Finally, after this investigation about Web syndication and the possibility of its use as resource for electronic journals, it is possible to summarize our conclusions based on the following topics:

1. The group of resources based on the Web syndication standards constitutes a technological innovation in the field of new reference services for information units, as well as for the development of potentialities of electronic journals;

2. It can be ascertained that the RSS is a meaningful tool for warning and automation of the content in the Web;

3. The publishers' commitment is essential for the dissemination of new products that use the technology of content syndication;

4. This technology is easily applied to the systems of information backup and of selective distribution;

5. New studies are necessary to widen the discussion about this issue, through new approaches and applications;

6. This technology is based on the information sharing paradigm. Therefore, it contributes to the generation of new knowledge.

## Notes and References

[1] According to Wikipedia, Blogosphere is the collective term encompassing all blogs as a community or social network.

[2] A "RSS feed" is a XML-based document that usually ends in .xml or .rss and is a slimmed-down version of a website created to be easily syndicated. It contains a list of items or entries of content metadata. News websites and blogs are common sources for RSS feeds, but feeds are also used to deliver structured information, such as articles from periodicals.

[3] The eXtensible Markup Language (XML) is a W3C-recommended general-purpose markup language for creating specion-purpose languages, capable of describing many different kinds of data.

[4] Crayon (http://crayon.net) is the abbreviation of Create Your Own Newspaper, a personalized information service that offers users the possibility of creating its own journal with links for more than 100 sources of news available in the Web.

[5] "What is Atom?" (http://www.atomenabled.org)

[6] For a full discussion of the history of web syndication, see Wikipedia. History of web syndication technology: http://en.wikipedia.org/wiki/History_of_web_syndication_technology

[7] A selected list of OJS journals is available on the PKP website: http://pkp.sfu.ca/ojs-journals

[8] Nowadays, the Firefox browser identifies if a webpage uses RSS showing the feed icon in the browser's status bar. However, it is not an RSS complete client, being necessary the installment of extensions to make its support even more powerful.

[9] SRU (Search/Retrieve via URL) is a standard search protocol for Internet search queries, utilizing CQL (Common Query Language), a standard query syntax for representing queries. Standards for SRU are promulgated by the United States Library of Congress.

# Changing Content Industry Structures: The Case of Digital Newspapers on ePaper Mobile Devices

*Leo Van Audenhove; Simon Delaere; Pieter Ballon; Michael Van Bossuyt*

SMIT-IBBT, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium
e-mail: leo.van.audenhove@vub.ac.be, simon.delaere@vub.ac.be,
pieter.ballon@vub.ac.be, michael.van.bossuyt@vub.ac.be

## Abstract

The proposed paper analyses the changes in business models employed by the stakeholders in the newspaper value network, in the context of a new type of electronic reading device –the ePaper. This PDA–like device uses a new high–contrast, low–power screen technology (eInk), which holds the promise of a digital and mobile reading experience close to that of 'real' paper. The potential impact of massive digitally distributed reading content –newspapers, but also magazines, books, and all other material previously printed on paper– on the traditional publishing value chain and its different constituent actors could be significant. For example, content aggregation roles already greatly dispersed by the internet could move further away from the traditional newspaper publishers; using logging data and RSS feeds on the device, newspaper advertising could become personalised and interactive; for newspaper publishers, production and distribution costs could go down and updated content could be sent to the device whenever needed etc. This paper is based on a large scale research project in Flanders/Belgium, which has brought together a device manufacturer, a financial newspaper publisher, a telecoms incumbent and several technological and social science research groups from Flemish universities. To complement the technological development and an extensive field trial with near–market devices, the authors analysed how this new technology might transform the traditional publishing value chain, what the strategic options of the different actors are, and what scenarios are possible and likely to occur in the development of ePaper publishing. To do this, they make use of the theoretical framework for business model analysis. Using literature study as well as empirical data (i.e. face to face interviews with important stakeholders from the newspaper and book publishing sectors), a number of scenarios for the re–definition of roles are outlined. The authors come to the conclusion that the choice for an open versus a closed architecture, along with the technological roadmap of the device, will be crucial in establishing a valid business model for ePaper. In this paper we complement the scenario study with information on the first commercial trials and products using electronic eInk based reading devices.

**Keywords:** electronic newspaper; mobile newspaper; electronic paper

# 1 Introduction

The rise of PC from the 1970s and the Internet and mobile communication from the 1990s have lured many self–proclaimed gurus in predicting that we are moving towards a paperless society. However, so far this idea has not materialised. If anything, the use of ICTs and the Internet seem to increase the use of paper, and the publishing industry is performing quite well despite all electronic information available. The main reasons why people still print electronic content on paper are 1) the portability of paper and 2) the high quality of the printed material. Visual displays still cause physical stress on its readers and the quality of the image is lower than on paper [1].

Different companies are searching for electronic alternatives for the traditional paper. One of the most recent additions is called eInk, a screen technology developed by a consortium consisting, among others, of Philips, Toppan Printing, Gruppo Espresso, Hearst Corporation, Motorola and Vivendi. The company's *electronic ink* – ink that carries a charge enabling it to be updated through electronics– allows for the production of so–called *Electronic Paper Displays (EPD)* possessing a paper–like high contrast appearance, ultra–low power consumption, and a relatively thin and light form factor. Theoretically, these devices could therefore be able to give the viewer the experience of reading from paper, while having the power of updatable information.

This paper analyses how the introduction of a device using this technology might provoke changes in business models, actors and roles in the (newspaper) publishing sector. It is based on the business modelling Work Package within a large scale government funded research project in Flanders (Belgium), called ePaper. The project brought together a device manufacturer (Philips/iRex Technologies), a financial newspaper publisher (De

Tijd), a telecoms incumbent (Belgacom), advertisers (Hypervision–Agency.com)/iMerge and several technological and social science research groups. To complement technological development of an ePaper device based on eInk technology, and an extensive field trial with near–market devices, the authors have analysed within this project how this new technology might transform the traditional publishing value chain, what are the strategic options of the different actors, and what scenarios are possible and likely to occur in the development of ePaper publishing. The potential impact of massive digitally distributed reading content on the traditional publishing value chain and its different constituent actors could be significant. For example, content aggregation roles already greatly dispersed by the internet could move further away from the traditional newspaper publishers; using logging data and RSS feeds on the device, newspaper advertising could become personalised and interactive; for newspaper publishers, production and distribution costs could go down and updated content could be sent to the device whenever needed etc.

In this paper, the results of our analysis will be briefly outlined. The methodological framework for business model analysis is concisely described. The paper focusses on the analysis of the ePaper value chain, and on the empirical elaboration of possible business model scenarios. The empirical basis for this work are expert interviews with representatives of the publishing industry in Flanders [2]. This analysis is complemented with information on the first commercial trials and products using electronic eInk based reading devices.

## 2 Approach and Methodology

Despite growing interest in business modelling in recent years, no clear definition of the term exists today. Different definitions emphasize diverging aspects such as the architecture of a product or service, a description of the roles of and the relations between companies, the ways in which business can be conducted, the way in which value is created etc. [3]. In this report, we use a definition, which tries to synthesize the most crucial elements in the mentioned literature and definitions [4]. We define a business model as: *'A description of how a company or a set of companies intends to create and capture value with a product or service. A business model defines the architecture of the product or service, the roles and relations of the company, its customers, partners and suppliers, and the physical, virtual and financial flows between them'.*

This definition relates to three levels of the business model: a *functional* level (dealing with the architecture of a product or a service), a *strategic/organisational* level (dealing with the roles and relations between actors and the physical and virtual flows between these actors) and a *financial* level (dealing with the sources of revenue of and the financial flows between the actors involved). In our analysis, we add to this a fourth level, i.e. the *value proposition*. This fourth level, which is the way value is created in the market, can be considered as a logical outcome of the strategic choices made on the other three levels when designing business models.

An important aspect of this definition is that is does not limit the focus of analysis to one specific firm, but instead takes into account a network of actors involved with the production, distribution and consumption of products and services. This reflects the growing complexity of innovation processes in what is called the network economy and society. From a financial perspective, the emphasis is on structuring the revenue streams and on creating models for revenue sharing.

In terms of the value chain, a concept coined by Porter to describe the primary value–adding activities of a firm or of a set of firms, this means looking at the whole chain [5]. In fact, most scholars agree that the increasing complexity and flexibility of business design means that the representation of business processes by a linear value chain has to be replaced by more fluid value networks, in which roles and functions can be combined in different ways by different actors. Business design is therefore increasingly about defining firms' boundaries and the level of horizontal and vertical integration. Taking into account the three basic levels of business modelling and the value proposition that is the outcome of these, a successful business model will emerge when a so–called *strategic fit* occurs between the different firms involved in the production of a product or a service, and on the different levels discussed, as well as between a firm's business model and the consumer [7].

## 3 The ePaper Value Chain

### 3.1 Value Chain and Network

We have started our business scenario analysis by analysing the ePaper value chain. This value chain contains the roles that are essential for the production and distribution of content on the ePaper device. It is important to point out that these roles may be taken up by diverging actors. In the ePaper value chain, we discern the roles of *Content Provision, Content Aggregation, Platform Content Aggregation, Platform Provision, Network*

*Operation* as well as *Service Provision, Advertising, Device Supply* and *Device Manufacturing*. The latter four roles are basically related to the strategies of other actors and to the business scenarios chosen, and are therefore not included in the value chain as such.

## 3.2 Roles and Actors in the Value Network

Below, we define the different roles in the ePaper value network. We indicate which actors are potentially interested in taking up any of the roles in the network. This implies that, besides looking at the newspaper sector, we also include the news production and publishing sectors in this value network; looking at the present functionalities of the ePaper device, content published on it will –at least initially– be of a written nature.

*Content Provision*. In the news and newspaper sector many actors take up this role (e.g. independent journalists, national and international news agencies, newspapers delivering syndicated content etc.) The newspaper itself acts as a producer for a lot of content; besides this, ePaper also provides a platform for other written content such as literature, magazines, trade journals, corporate publications etc. coming from a host of different providers.

*Content Aggregation*. In the news production sector, the newspaper is a typical example of an aggregator of content. Newspapers and magazines make a profession out of bundling content, services and advertising in a coherent editorial concept. These actors strongly believe that this aggregation function will remain an important task in the digital age. However, the digitisation of content and the subsequent creation of new communication platforms such as the Web, i–mode, iDTV etc. have spurred the development of alternative content aggregators.

*Platform Content Aggregation*. It is important to make a distinction between *Content Aggregation* and *Platform Content Aggregation*: the former relates to the filtering, editing and branding of content in an editorial concept, the latter points to the assembling of already aggregated content (e.g. newspapers, books, magazines, etc.) of different *Content Providers* and *Aggregators* onto an electronic platform. For example, *Newsstand.com* offers a broad selection of digitised international newspapers and magazines from different publishers on the Internet Platform. A crucial point of discussion surrounding ePaper is the degree to which content from newspapers and other providers will be offered in an aggregated or a desaggregated manner. In constructing business scenarios for the ePaper platform, a central variable will be who takes up the role of *Content Platform Aggregation*.

*Platform Provision,* i.e. the provision of a technical platform that links content and technology. This role is significant because it determines, to a large extent, the control of who publishes on the device and what is possible on it. This role can be divided into a server–side and a software/DRM function. The server–side function assures communication between the content provision and the ePaper device and therefore constitutes a potential bottleneck. The uncertainty on which actor will take up this function, renders the function into a possible source of conflict within the value network.

*Network operation*. This is the domain of telecommunications operators, whose services might be considered as substitutable commodities. In such case, *Network Operation* is reduced to the provision of a *pipeline* for the content. However, network operators worldwide are trying to broaden the scope of their operations from pure transmission to the offering of content–related services. Within ePaper, these actors might have the ambition to take up the roles of *Platform Content Aggregation* and *Content Aggregation*.

*Service Provision.* This is a crucial role in the ePaper value network, relating to who maintains the customer relationship and effectively markets the service. For the time being, this role cannot be identified in the value chain, since its positioning within this value chain depends from which actor takes up this role. The newspaper or its overarching publisher seems to be well–placed to do this, because –especially in subscription models– it has a unique relationship with its customers. However, when looking at the technological functionalities of ePaper, other actors –for example *Platform Content Aggregators*– could also take ups this role.

*Device Supply*. The question here is by whom and in which way the device is marketed. Again, this role cannot be identified in the value chain for the moment because it is dependent upon the business scenario chosen. Taking into account the cost of the device, we expect that this role will often coincide with the offering of content and services, and that the device will be offered in some sort of subscription model. However, other options, among which an ePaper reader as a simple consumer device remain possible.

*Device manufacturing*. At the moment there are only a few commercial eInk based devices on the market. iRex technologies—the company involved in the trial this paper is related to—developed the Iliad reader and Sony developed a Librie and Sony Reader. Both can display different content, but the Iliad was specifically developed with electronic newspapers in mind, whereas Librie and Sony Reader were developed to display ebooks.

*Advertising*. This role is already fully part of the traditional newspapers' value chain, with newspaper publishers in the role of *Content Aggregators* integrating advertisements coming from other parties. However, ePaper offers new opportunities for advertising, e.g. interactive and personalised ads, on the level of the electronic newspaper (*Content Aggregation)* as well as on the level of the device (*Platform Content Aggregation)*. The *Advertising* role will therefore be dependent upon the business scenario chosen. Initially it is not foreseen that the advertisers will play a central role in the ePaper value network: our interviews with the newspaper and magazine sector in Flanders have shown that these sectors are rather skeptical about highly personalised content and advertising.

# 4 About the Potential Scenarios for ePaper

The above discussion of the ePaper value network has made clear that this network contains several roles which can be taken up by different actors. Question is how these roles are complementary with the interests and strategies of existing actors. The digitisation of content implies that the role of *Content Aggregation* –which, in the offline world, is a clear prerogative of the newspaper editors– could shift towards the platform itself by means of *Platform Content Aggregation*. The roles of *Service Provision* and *Device Supply*, for their part, are closely linked to the business scenario chosen.

In order to gain insight into potential and probable business models, we use the scenario method, in which two uncertain variables are defined, along which four potential futures can be outlined. In the present context, many of these uncertainties are surrounding the ePaper device and possible business scenarios; based on the interviews and on our literature review, we were able to define two uncertainties which can be considered as crucial:

*Aggregation vs. Desaggregation*, i.e. the degree to which content is offered on the platform in an aggregated or desaggregated manner, defined from the perspective of the newspaper. *Aggregated* signifies that the newspaper can offer its content *as such* on the platform, whereas *desaggregated* means that the content on the device originates from different content providers and is more *fragmented*, i.e. less edited, packaged and branded.

*Open vs. Closed*, i.e. the degree to which the device is accessible for content originating from different content providers. A crucial question for determining this variable is whether –and if yes, to what degree– an exclusive link exists between the offering of content and the display of that content on the ePaper device.

It is striking that the different actors interviewed and studied have pronounced often conflicting opinions about the necessity of an open or a closed model and about the inevitability of the evolution of media towards a desaggregated model. Both variables may be used to create a co–ordinate system comprising four quadrants, with each quadrant representing a potential business scenario. We discern these scenarios: *(1) Newspaper model (Aggregated–Closed); (2) Kiosk model (Aggregated–Open); (3) iTunes model (Desaggregated–Closed); (4) Web model (Desaggregated–Open)*. Below, we shall describe four generic scenarios and analyse their potential.

# 5 Scenario 1 – The Newspaper Model on ePaper

## 5.1 Business Scenario Outline

In this scenario one party, the Content Aggregator, offers a particular service on the ePaper device. This scenario is largely similar to the experimental IBBT ePaper project, in which De Tijd publishes an electronic version of its newspaper onto the device. In principle this can be done in two ways: (1) the newspaper can be uploaded to the device as is, without any major adaptations to the structure; (2) the newspaper may, as Content Provider and Content Aggregator, make use of the new capabilities of this medium. In the latter case it can alter its service by (1) publishing up–to–date content multiple times per day, (2) offering specific information aimed at particular audience segments, (3) personalising content, (4) integrate personalised advertisements into the content etc. Whatever option is picked, the newspaper remains the primordial provider of content on the device.

In the figure below we have displayed the value network of this scenario in a generic fashion. Besides the newspaper's role of Content Provider and Content Aggregator, the ePaper device offers new opportunities to put content on the device originating from third party providers. In this scenario, we make the assumption that the newspaper itself might play a potential role; in other words, the newspaper could take up the role of Platform Content Aggregation –or part of that role (see figure). Two options exist for doing this:

1) The newspaper could complement its own content with content from its own publishing group, thereby enhancing the attractiveness of its own service and possibly also increasing revenues of its entire group. An important condition for this is the availability of a sufficiently large and complementary offer within this publishing house that can appeal to the targeted audience;

2) In case the newspaper wishes to offer content originating from third parties outside its own publishing group, then this content can be expected to be mainly complementary; other newspapers will have little inclination to publish their product on a competing platform. This hypothesis is confirmed by the Content Aggregators interviewed for this study, who clearly indicate that they are only prepared to provide content for a device which is administered by a neutral party.

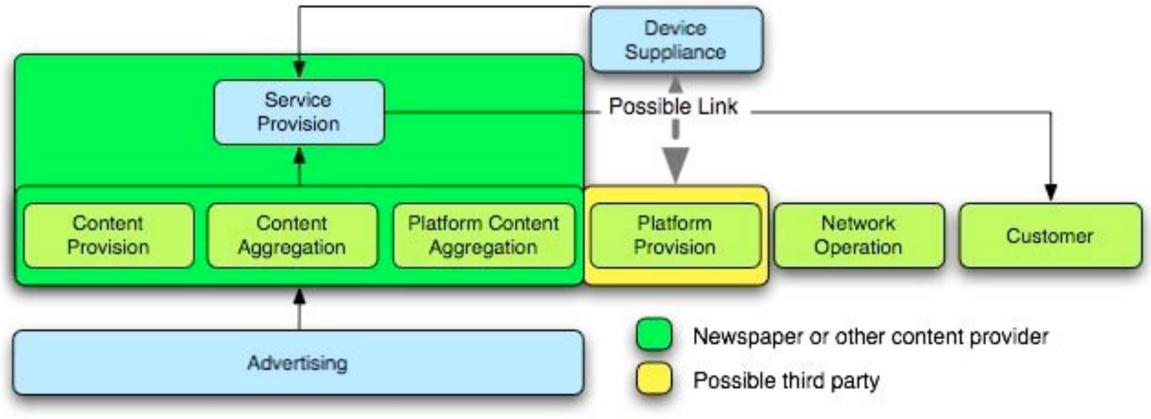


**Figure 1: Newspaper model value network**

If a newspaper integrates the roles of Content Provision, Content Aggregation and Platform Content Aggregation, then it is clear that this actor will market the service. It has considerable advantages over other parties in doing this: (1) an existing customer relationship, (2) content for which customers are prepared to pay and (3) a certain market intelligence.

The role of Platform Provisioning may be taken up by the newspaper itself or by a third party. Newspapers might well be interested in doing this, since a number of parties indicate that newspapers are, in a digital environment, prone to handle distribution themselves. Other potential actors are the Device Manufacturer, the Device Supplier or the Network Operator. The Device Supplier has a certain control over the device configuration, the standards used, the capabilities and limitations imposed by DRM etc. In the Flemish case, iRex is taking up this role by having developed a client as well as a server component, and is able to simultaneously offer tailored services to different parties; the functionalities of the architecture are negotiated with the newspaper in its different roles.

For marketing the device, two main options exist: (1) the customer may individually purchase an ePaper device and subsequently take a digital subscription to a newspaper; (2) the newspaper may offer the ePaper device as part of a subscription to the digital paper. In this project, it is clear that iRex, as a Device Supplier, has chosen the second model. The argument for this is that the ePaper device, unlike the iPod for example, does not have an unambiguous, easily recognisable functionality for the consumer, and that it is rather expensive at the moment. The device therefore seems easier to integrate into the market when being part of a subscription model. However, this also implies that the newspaper will need to carry the financial burden of pre–ordering the devices. As for the Device Supplier, this actor could create an additional revenue stream by also taking up the role of Platform Provider. In its turn, the Platform Provider could be inclined to shift towards the role of Platform Content Aggregator and publish services on the device itself. However, as it is the newspaper who markets the devices itself, this scenario seems rather implausible.

In case the actors choose to make use of personalised or more directed advertising, an exchange of information will need to take place between the Platform Provider, the Platform Content Provider (being the newspaper in this scenario) and the Advertiser. Firstly, the Advertiser will be interested in obtaining information about (1) the use of the platform and the characteristics of the user, and (2) which user has seen/clicked on which advertisement. This information is also important for the newspaper itself since clicking through on advertisements usually generates higher revenue.

## 5.2 Evaluation

In this scenario the newspaper plays a dominant role. It has a number of important advantages: a large reader base, a good customer relationship and content that customers are willing to pay for. The newspaper may address this reader base in order to try to make a large group of readers use ePaper as quickly as possible. In making this effort, marketing the ePaper device as part of a subscription offers a number of additional advantages. Firstly, readers will be more easily persuaded to switch to the technology; secondly, in the longer term this strategy might have a cost–reducing effect for the newspaper; and finally, the newspaper would be able to monitor the reading behaviour of its customers in order to better tune the content to reader preferences.

However, the functionality of ePaper as a digital reading platform for content originating from a large array of producers is threatened, particularly if the platform is too strictly protected by DRM and proprietary standards. In this case, this scenario might become alienated from the actual wishes and demands of the targeted audience (in this case, business professionals). In this sense, the use of ePaper as a mere digital substitute for the newspaper could be considered as a rather conservative reflex by newspapers in order to maintain readership in the digital era. Moreover, an initiative launched by only one newspaper or publishing house, might be boycotted by other players in the market.

The first commercial produces with ePaper readers are examples of this scenario. At the end of 2006 the Yantai Daily Media Group started publishing its main newspaper on the Iliad in China. In May of 2007 two newspaper of the Dutch PCM Group *De Volkskrant* and *NRC Handelsblad* will become available on the Iliad. There is so far little known about the projects and the agreement between PCM and iRex, but PCM is in discussion with other groups to extend services. This might indicate that PCM might also be interested in the kiosk model [7].

# 6. Scenario 2 – The Kiosk Model on ePaper

## 6.1 Business Scenario Outline

We call this the kiosk model by analogy with the newspaper kiosk. Currently, kiosks offer –besides a selection of national and foreign newspapers– a wide array of magazines, books etc. Transposed to the ePaper device, the user of this device has, in this scenario, access to a wide choice of textual media originating from different publishers. However, these publishers mainly continue to provide content in aggregated format. For the user, this scenario ads value because he can use the ePaper reader as a mobile platform for a large selection of content.

In the realm of the *audiobooks*, a platform similar to this one exists which is called *audible.com*. Audible is a platform for digital audiobooks which has a library of over 27,000 titles originating from 318 *Content Providers/Aggregators*. After installing a piece of software –either *iTunes* or *Audible Software*– files may be purchased and downloaded to a computer and subsequently to an mp3 player. Audible makes use of DRM to prevent files from being copied, but does not link its software to one particular device for using these files. According to the company, more than 200 devices are able to deal with the format used. In the realm of ebooks similar initiatives exist such as ebooks which brings together 80.000 titles from different publishers and mobipocket with 39.000 premium titles. eBooks distributes books in three standards i.e. Microsoft reader, Adobe reader and Mobipocket reader. Mobipocket uses its own standard.

In this scenario, an *intermediary* is a central actor in the value network. This intermediary takes up the role of *Platform Content Aggregation* and brings together content from diverging *Content Providers* en *Content Aggregators*. The main advantage for an intermediary is that it unites two markets, namely that of information providers and that of information users. If the intermediary succeeds in bringing a large segment of both markets to its platforms, significant network externalities occur on both these markets: the *Content Providers* gain access to a potentially larger customer base, while users have a much larger selection of content [8]. Following this strategy, Audible for example has succeeded to use the internet to create a *one–stop shop* for English language, digital audiobooks and has been able to further diversify into spoken newspapers, magazines, radio programmes and talk shows, which were distributed to 278,000 paying customers in 2006 [9].

In this scenario, it seems logical that the *Platform Content Aggregator* maintains the customer relationship or, put differently, that it takes up the role of *Service Provision*. The *Content Provider* or *Aggregator*, be it a newspaper or a publisher, uses the *Platform Content Aggregator* as an alternative distribution channel. In that case the newspaper could lose its relationship with the subscribed readers to the *Platform Content Aggregator.* In an online environment the latter actor could create a relationship with its customers, even if they don't take a newspaper subscription. A potential alternative to this model is that the newspaper, as a *Content Aggregator*, retains the role of *Service Provision*, but uses the platform to grant users access to a larger array of content.

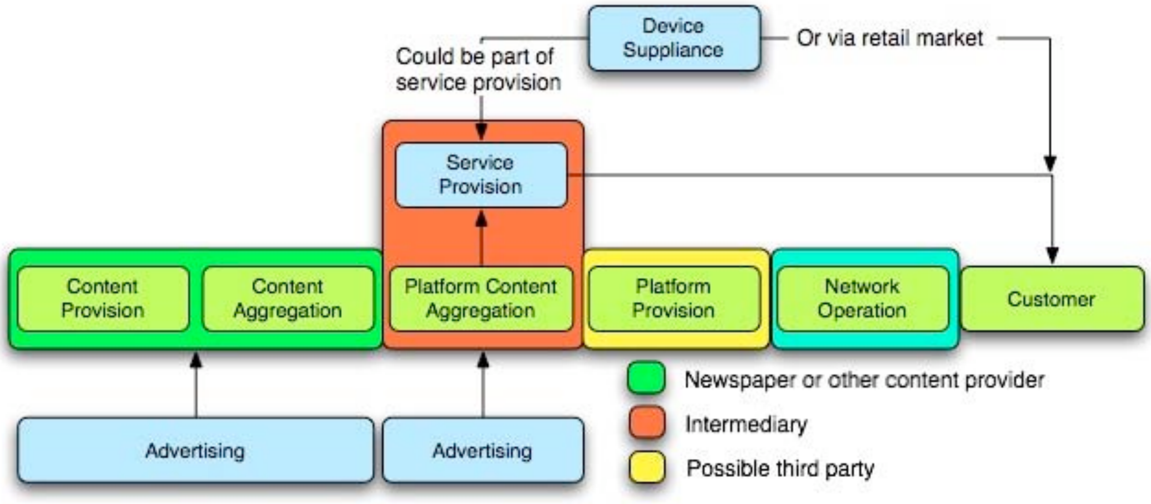


**Figure 2: Kiosk model value network**

It remains an open question who takes up the role of *Platform Provision* within this scenario. This role can be exerted by the *Platform Content Aggregator* itself, by the *Network Operator* or by a third party. In case the roles of *Platform Content Aggregation* and *Device Supply* are not combined, the *Platform Content Aggregator* –in this case the intermediary– faces two crucial challenges. On the one hand, this actor wishes –partly under pressure from the *Content Providers*– to prevent the copying of content, among other things by including DRM; on the other hand he wishes to offer his content on as much devices as possible. On the level of functional architecture, this party will therefore strive towards (1) the use of open standards that allow publication on multiple devices, or (2) the development of a proper solution that is subsequently supported by multiple producers. The latter strategy can only work if the intermediary has a sufficiently strong market position. A central question remains the role of the device manufacturers. Do they wish to sell their device as a piece of hardware with a number of technical service components, or do they also wish to take up other roles in the value chain, namely that of *Platform Content Aggregator?* (cf. next scenario). When transposing the scenario to the newspaper sector, the question is which party will take up the intermediary function. The establishment of a region– or nationwide intermediary could be a possibility that different actors seem to prefer –as was shown by the interviews.

In this scenario, advertisement might in principle play a role on two levels, namely that of the *Content Aggregation* (by a.o. newspapers and magazines) and that of the *Platform Content Aggregation*. As for the first level, an important issue here again is whether agreements can be reached and information exchanged between the *Platform Content Aggregator* and the *Content Aggregator* to allow personalised advertising on the level of the newspaper. After all, in the proposed scenario it will particularly be the *Platform Content Aggregator* which has disposal of a large amount of data concerning the user and content consumption behaviour. As for the second level (*Content Platform Aggregator)*, advertisements might be possible here as well. However, experience has shown that this only occurs in a limited way; the main reason for this is that the *Platform Content Aggregator* is deemed to remain a neutral party. Both *iTunes Music Store* and *Audible* –two intermediaries on the internet– do not allow publicity on their platforms, and have strict editorial guidelines as regards the presentation of products. Our interviews have clearly shown that advertisements on the level of the *Content Platform Aggregator* would not be readily accepted by *Content Aggregators*.

In this scenario the two options for marketing the device are open, and lot depends on the payment options used. In our example Audible offers several of these payment options: (1) a one–off payment per title, (2) a subscription granting a year long reduction on titles, (3) a subscription giving access to one title per month for a one year period or (4) a similar subscription allowing access to two monthly titles. In this case, the device is part of the *Service Provision*. However an ePaper device could also be marketed as a consumer device. The examples of payment methods for products and services mentioned above could also be implemented for the newspaper and (book) publishing sectors. In this scenario, it will likely be the *Platform Content Aggregator* which bundles services and device. However this is not a necessity: one of the interviewed *Content Aggregators* indicated that it was prepared to subsidise the device as part of a subscription *and* to grant access to third party content.

In this scenario, price–fixing and revenue sharing between *Platform Content Providers* on the one hand and *Content Providers* and *Content Aggregators* on the other hand, will be a difficult exercise and a possible source of conflict. The *iTunes* case in the music sector (cf. sub) constitutes a nice example of this: while a price of USD 0.99 per downloaded song is generally assumed to be too high, this price has to a large extent been imposed by the music industry [10]. A possible solution for avoiding conflict is the establishment of a *Platform Content Provider* within the sector in which the different actors participate.

## 6.2 Evaluation

This scenario offers interesting opportunities to stimulate the ePaper device as a mobile platform for different types of content originating from different parties while, from the publishers' perspective, the products offered retain their editorial function. It is less clear whether this scenario also contributes to the innovative use of the interactive capabilities of the device; this will require clear agreements between the *Platform Content Aggregator* and the *Content Providers* and *Aggregators.*

The introduction of an intermediary party as *Platform Content Provider* offers major advantages in terms of network externalities related to two–sided markets. However it also holds some threats: taking into account the economies of scale and network advantages created by internet and ICT–based platforms, this party could in time become a powerful actor, in particular if it maintains the customer relationship and if it has data on user preferences at its disposal. An additional threat is that the intermediary would shift toward *Content Aggregation* and *Content Provision*. In our example, Audible offers audiobooks that it has produced itself. Besides this, the launch of a new intermediary also implies larger necessary investments and limited brand awareness.

Setting up a totally new intermediary platform might proof to be a difficult exercise. Although all Flemish newspapers and publishers indicated to be in favour of the kiosk model actually setting up such a platform is another issue. Competition and mistrust might easily prevent this scenario. However, in other countries umbrella organisations representing or serving the newspaper industry already exist. E.g. the Joint Purchasing Association of the Danish Newspapers is an umbrella organisation aggregating demand for and purchasing paper for the different Danish newspapers. Such organisations might be the basis for an intermediary platform.

# 7 Scenario 3 – iTunes for ePaper

## 7.1 Business Scenario Outline

At first sight, the iTunes model seems to resemble the preceding model: here too, an intermediary partner takes up the role of Platform Content Aggregator, bringing together content from Content Providers and Aggregators. However, the scenario differs in two crucial points. Firstly, there is a certain degree of desaggregation. On the iTunes Music Store, users are able to download a single song. Transposed to the newspaper and publishing sector, this implies that separate articles could be purchased. We immediately need to add to this that desaggregation of newspapers will be trickier because the advertisements inserted are an important source of revenue. Secondly –and fundamentally differing– the same party (i.e. Apple) takes up the role of Platform Provision and of Device Supply, for Apple controls, via its software, the interaction between the iTunes Music Store and its device –the iPod– and songs downloaded via iTunes can only be played on the iPod.

A similar scenario can also be elaborated for the newspaper and publishing sector. Sony is currently aiming to do this for eBooks by using its new Sony eReader. This device can only access content from Sony's own content site Sony Connect. For this content, the Japanese firm has concluded agreements with a number of big publishing houses in the United States. In this scenario, the user still has access to a large offer originating from a number of Content Providers and Aggregators, but is forced to watch this content via a specific device, i.e. an ePaper reader. By analogy with the iTunes software, it would however be possible to print a selection [11].

As in the preceding scenario, the intermediary fulfils a crucial role in terms of uniting offer and demand. However, in this scenario the intermediary integrates even more roles, i.e. that of Platform Content Aggregation, Platform Provision, Service Provision and Device Supply (as well as Device Manufacturing). Especially in the iTunes case, where Apple has reached a US market share of more than 70 percent of mp3 players with its iPod, the combination of Platform Provision and Device Supply results in a fairly dominant position [12]. In this scenario too, there is a certain danger that the Platform Content Aggregator gradually shifts towards Content Aggregation and Content Provision; through the desaggregation of content coming from Content Providers and Aggregators, the Platform Content Aggregator is able to personalise its service to users even better.

In the iTunes case, a link exists between the iTunes Music Store, iTunes software and the iPod. The iTunes software gives access to the iTunes Music Store and takes care of file transfers to the iPod. The files on the iTunes Music Store are protected by DRM and Apple uses a proprietary encoding standard for its files, i.e. AAC. This way, files can only be transferred to four different iPods; however the software does allow content from third parties to be loaded onto the device in mp3 or AAC. For ePaper a similar—or even stricter—scenario could be chosen, in which the device itself (and not the PC) acts as the interface between the store and the platform. Moreover, the publishing sector could use a strong push–model, in which up–to–date content is pushed towards a device after the user has indicated which content is of interest to him or her.

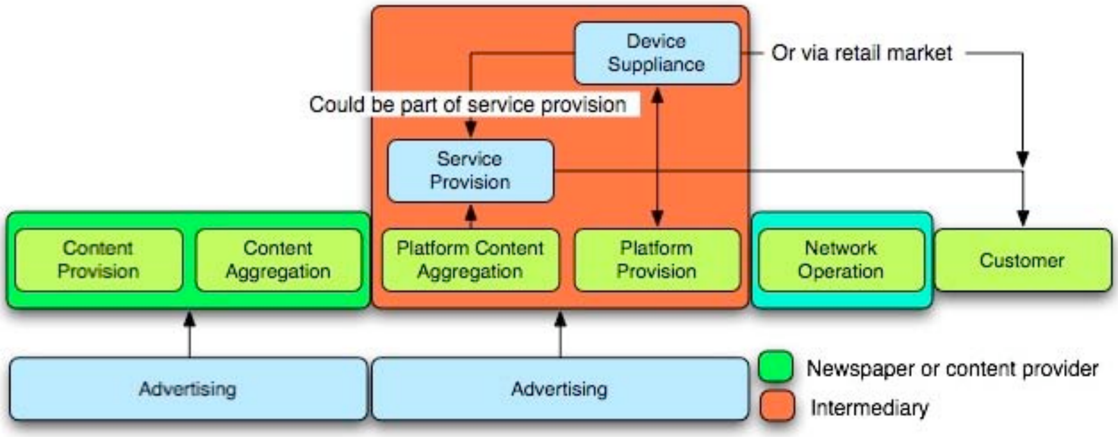


**Figure 3: iTunes model value network**

Taking into account this integration, it seems obvious that the Platform Content Aggregator is also responsible for Service Provision and thus maintains the relationship with the customers. Here too one can wonder about the plausibility of a scenario in which the newspaper, as Content Provider and Aggregator, takes up its own part of Service Provision. Finally, the Advertising role can be exerted on the same two levels as in the previous scenario, so the same issue apply.

In this scenario, different payment methods are equally possible; in that sense, it largely resembles the previous scenario. As it is assumed here that content can be accessed in a desaggregated format, separate articles from different Content Providers may be purchased. This necessitates new ways for paying this content, among which micro–payments. In case the Network Operator takes up the role of Platform Provisioning –or part of that role–, it may be well placed to take care of billing in this model.

A particularity in this scenario is that a larger number of roles are combined, among which Platform Content Aggregation, Platform Provision, Service Provision and Device Supply. This gives the opportunity, for the actor taking up these roles, to generate revenues on different levels: (1) as a percentage on sold content or subscriptions, (2) on the basis of devices sold or (3) on the basis of a service component aimed at Content Providers and Aggregators. Option (1) and (3) may eventually be combined as one percentage on content sold, including service provision. The price that can be asked by an intermediary for selling content depends on the negotiations with the Content Providers and Aggregators and what the bargaining power of these latter actors is. The intermediary could also strategically opt to position itself between these two revenue streams. Although little is officially known about this, it is generally assumed that Apple only generates limited profit out of its iTunes Music Store and instead focuses mainly on iPod sales.

Within this scenario, it is again possible to insert advertising on two levels, i.e. on the newspaper level (or even within a separate article), and on the level of the platform. Because access to desaggregated content is possible, it seems more logical within this scenario to administer at least part of the advertising on the platform level. Besides this, it is also the intermediary which possesses the knowledge about device and platform use as well as user preferences, which it could exploit as a third revenue stream. However, it seems unlikely that newspapers and publishing houses would hand over an important portion of their advertising revenues to the intermediary without any compensation.

## 7.2 Evaluation

In this scenario, the user has access to desaggregated content, i.e. individual articles from newspapers, magazines etc. This type of service clearly fits closer to the changes in reading behaviour of modern newspaper readers, as well as to changes in users' experiences with other ICT devices.

The intermediary party which integrates the roles of Platform Content Aggregation, Service Provision and Device Supply, threatens to become dominant within this scenario, which might render the publishing sector reluctant towards participating in it. Moreover, this sector traditionally attributes high value to the editorial concept with which it links its brand names, and possibly fears that excessive desaggregation will turn their content into an easily substitutable commodity. Finally, if the intermediary party protects content and devices by using DRM and proprietary standards, the user will in turn be rather reluctant to purchase such a device.

# 8 Scenario 4 – The Web on ePaper

## 8.1 Business Scenario Outline

In this scenario the ePaper device may be considered as a new gateway to the Web. The device has little or no protection by DRM or proprietary standards, so the user can upload any content –coming from the Web or produced by him/herself– onto the device. In a sense, the role of *Content Aggregation* shifts to the user by becoming that of *Content Selection*: the user actively searches for information from newspapers, weblogs, government websites, discussion forums, newsgroups, entertainment companies etc. This *prosumer* can also create information himself and make that information available to others.

All this does not necessarily mean that the user is not prepared to pay for content. He/she can still purchase certain types of content, albeit directly from the *Content Providers/Aggregators* and *Platform Content Aggregators*. Thus, while these latter roles continue to exist, the user has access to a large number of actors which individually make content available; the user is not necessarily tied to one actor.

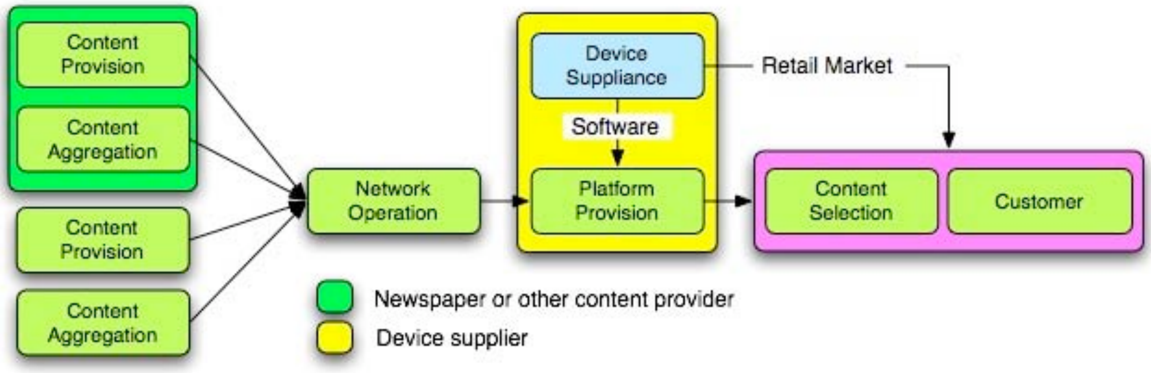


**Figure 4: web model value network**

The value network of the web model strongly differs from the other scenarios. Firstly, in this model *Content Provision, Content Aggregation* and *Platform Content Aggregation* are vertically aligned. The consumer has individual access to the content of one or more of these actors and newspapers, as *Content Aggregators*, directly compete with other *Content Aggregators* such as Google News, Newsstand etc. as well as with individual *Content Providers*. Secondly, the role of *Platform Content Aggregation* (at least at the device level) no longer exists; on the one hand, this role largely taken over by the user, while on the other hand one could argue that *search engines* also take up part of it. Thirdly, *Platform Provision* can still occur in the shape of software making up the interface between the internet and the device. Although this software could protect part of the content using DRM, the *Device Supplier* will not be inclined to consider this option. To the extent that *Content Providers* are only willing to publish their content on devices that protect this information, it is possible that pressure is exerted in order to include DRM solutions on these devices. The same goes for standards: as device sales are crucial for the *Device Supplier* in this scenario, he will be prone to support multiple and open standards.

In this scenario, it is more difficult to monitor the use of the device. Every *Content Provider* is able to track which of its content is downloaded, but the possibilities to gather information on what the user does with this content, are rather limited. These functionalities could be incorporated into the interfacing software of the device (as *adware* or *spyware*); however, these types of monitoring are usually strongly disapproved of by the user.

In this model, it seems fairly implausible that one party would market the device as part of a subscription; the consumer will rather buy such a device by itself. Although iRex has indicated that it would primarily focus on the B2B market, it is not inconceivable that another manufacturer would brand a similar device as a consumer product. This scenario becomes more plausible if multiple *Device Manufacturers* compete with each other on a device level. On the *Content Provision* and *Aggregation* levels, the revenues are generated by the individual actors.

## 8.2 Evaluation

This scenario probably fits in best with the desires and expectations of the user; he or she potentially gets access to a very broad range of content. However, it remains to be seen whether the different parties are willing to realise this scenario. Newspapers are primarily interested in finding new distribution channels for their product, and not in a device that offers desaggregated contest and on which they have to face full competition from free internet services. The device manufacturers for their part possibly face a *chicken–and–egg* dilemma if they cannot link the sale of devices (with the inherent distribution and marketing costs) to the guaranteed availability of content for the user.

# 9 Conclusion

In this study we have elaborated scenarios that describe *possible* roads towards a business model for ePaper. For doing this, we have used two fundamental uncertainties, being (1) the degree of aggregation versus desaggregation from the perspective of the newspaper, and (2) the degree to which the device is open for content originating from different providers. The combination of these variables has resulted in four scenarios: the newspaper model, the kiosk model, the iTunes model and the web model. To contextualise the scenarios we have conducted interviews with actors within the Flemish newspaper, publishing and telecommunications sector. Furthermore we have complemented the analysis with information on the first commercial trials currently running.

The described models are generic and represent only one type of business model. Besides the crucial uncertainties used in this study, too many variables exist –hence our choice for the scenario methodology. The eventual model depends on the strategic choices made by the different actors; in this regard, our interviews have already shown major differences in opinion between the actors involved. We have generically integrated these insights into the scenarios. The combination of the interviews, the literature review and the scenarios drawn up, has lead to a number of strategic considerations:

Both newspapers and publishers in general will continue to believe in the importance of editorial concepts and guidelines. They will therefore have little inclination to give this up in favour of a completely desaggregated system. The fact that a large number of customers is still prepared to pay for this service (be it in paper or for the online version of newspapers), certainly proves its relevance. In each of the scenarios, the newspaper's customer database offers a major advantage for marketing ePaper.

The newspaper has –much more than other media– a relationship with its customers. This is particularly the case for subscription readers –which form a large part of the audience in Flanders, but also in many other countries. Therefore, newspapers will mainly consider new distribution channels as a way to diversify their services, but will not be willing to give up this customer relationship, especially since the possibilities for monitoring news consumption offered by ePaper allow these newspapers to further deepen their knowledge about their customers.

Taking into account these arguments, scenario 1 seems to be an important plausible option. This is confirmed by the first commercial initiatives with ePaper devices. Both the Yantai Daily Media Group in China and the PCM Group in Holland have started with offering titles on the Iliad on an individual basis. Nevertheless, platforms such as iTunes, Audible, Rhapsody, Amazon etc. show that intermediaries in two–sided markets –aggregating *Content Providers/Aggregators* on the one hand and users of content on the other hand– can become a big success. Two–sided markets have significant network externalities that may be of particular benefit to users by creating a much broader offer of information. As newspaper markets are to a large extent delineated by language and national boundaries it will remain to be seen whether intermediaries will develop at this national level. In the present context, the position of the *Device Manufacturer* and the roles it will take up, constitute important and uncertain variables. For the moment, the actors involved seem to opt primarily for a B2B strategy. In the short term, this renders scenario 4 less plausible.

As mentioned, the question which scenario –or which derivative of such as scenario– will eventually become reality, largely depends on the strategies of and the negotiations between actors. Two final important remarks need to be made in this regard. Firstly, the scenarios *are not mutually exclusive*: it is perfectly possible for a newspaper and a *Device Manufacturer* to strive, in the short term, towards a newspaper model (scenario 1) while leaving room for elaborating other scenarios, such as a kiosk model (scenario 2). Secondly, it is not inconceivable that, as time passes, a shift occurs from scenario 1 to scenario 4. Particularly if eInk or similar technologies become more broadly adopted and multiple devices are launched, the pressure for creating open systems might increase. On the one hand it is important for newspapers to take this into account *a priori* and to avoid investing in systems and technology that create too much path dependency or that are not adaptable. On the other hand it remains to be seen whether this 'conservative' sector will grab the new opportunities this technology offers or whether it will be the Internet or electronics sector who will drive the initiatives.

## Notes and References

[1] SHAVER, D.; SHAVER, M. A. (2003). Books and Digital Technology. A new industry model. *Journal of Media Economics*, 16(2), 71–86

[2] Interviews with: iRex Technologies, Philips, Hypervision, Uitgeversbedrijf De Tijd, I–Merge, Belgacom, Lannoo, Magnet Magazines, Concentra, De Standaard Uitgeverij, De Standaard

[3] WEILL, P.; VITALE, M. (2001). *Place to Space: Migrating to eBusiness Models*, Boston: Harvard Business School Press; OVANS, A. (2000). E–Procurement at Schlumberger. *Harvard Business Review*, 78(3): 21–23; TIMMERS, P. (1998). Business Models for Electronic Markets. *EM– Electronic Markets*, vol. 8, no.2; SLYKOTSKY, A. J. (1996). *Value Migration – How to Think Several Moves Ahead of the Competition*. Boston: Harvard Business School Press.

[4] BALLON, P. (ed.) (2005). *Best Practice in Business Modelling for ICT Services*. Delft: TNO

[5] PORTER, M. (1985), *Competitive Advantage: Creating and Sustaining Superior Performance*, New York: Free Press.

[6] METHLIE, L.; PEDERSEN, P., (2001) *Understanding business models in mobile commerce*, Paper presented at WWRF 3, Stockholm, September. BOUWMAN, H. (2002). *Business Models for Innovative Telematics Applications*, Enschede: Telematica Instituut; FABER, E., BALLON, P., BOUWMAN, H., HAAKER, T., RIETKERK, O., STEEN, M., (2003) *Designing business models for mobile ICT services*, Paper presented at E–commerce workshop, Bled, June 9–11.

[7] HOOFT VAN HUYSDUYNEN, M. (2007) Volkskrant en NRC op electronisch papier, *FEM Business online*, http://www.fembusiness.nl/fembusiness/content/nieuws/55059/article.html

[8] CORTADE, T. (2006). A Strategic Guide on Two–Sided Markets. *Communications and Strategies*, No. 61, 1st Quarter, 17–37, EISENMANN, T.; PARKER, G.; VAN ALSTYNE, M. (2007) Strategies for two-sided markets, *Harvard Business Review,* oktober, 92-101.

[9] MACKENZIE, K. (2006). Audio books open a new chapter in digital age. *FT.COM Financial Times*, May 26

[10] KUSEK, D.; LEONHARD, G. (2005). *The Future of Music. Manifesto for the digital music revolution*. Berklee: Berklee College of Music.

[11] VAN AUDENHOVE, L. (2004) *The business scenario behind the iTunes Music Stores and the iPod.* B@Home Working Paper, Delft: TNO–STB

[12] SONY (2006) Sony and Borders to sell digital reading device, *Sony Electronic News and Information*, from: news.sel.sony.com (Accessed 5/16/2006)

# Introducing the e-newspaper – Audience Preferences and Demands

*Carina Ihlström Eriksson; Maria Åkesson*

School of Information Science, Computer and Electrical Engineering, Halmstad University
P.O. Box 823, S-301 19 Halmstad, Sweden
e-mail: carina.ihlstrom_eriksson@ide.hh.se; maria.akesson@ide.hh.se

## Abstract

This paper adds to the overall understanding of new media adoption in general and the promotion of the e-newspaper in particular by empirically studying the preferences and demands of the potential users. The e-newspaper is a newspaper published on e-paper technology. The findings in this paper is based on the results from two studies, i.e. an online questionnaire with 3626 respondents and an evaluation in real life settings with 10 families over a two week period. Our initial hypothesis was that: *users confronted with a vision of new technology and services are more positive to adopt than users with actual use experience of technology and services in an early stage of development with inherent technology problems.* The research question of the paper is: *How does use experience influence perceptions of preferences and demands for the e-newspaper?* The findings showed that the hypothesis proved to be false, the test persons that have an actual use experience of the e-newspaper, despite the shortcomings in the device and service, were more positive to adopt than the respondents that have experienced concept movies and prototypes with more advanced functionality and interface.

**Keywords:** new media adoption; user experience; e-newspaper

## 1 Introduction

New mobile devices are constantly being introduced to the market offering new opportunities for publishing mobile media content and services. It is very difficult however, for content providers to predict m-commerce markets due to the uncertainties related to adoption of new mobile technology and services [1]. Moreover, this situation is new not only to content providers, it is also new to the audience. The rapid introduction of mobile technology and new services has led to a situation where users are constantly trying out new appliances and new services. This in turn changes use patterns as well as creates new preferences and demands, which leads to uncertainty about what people want [1].

Mobile service adoption has been studied by many scholars, e.g. drivers for adoption and intentions to adopt mobile services [2], factors influencing adoption [3, 4], adoption patterns [5-7], and attitudes towards using mobile services [8]. Much of this research has been focused on the adoption of mobile devices as such, as without adoption of devices there is not any prospect for successful m-commerce [9]. On the other hand, we argue that without attractive mobile content and services there is no incitement for m-commerce. This is indicated by the fact that in spite of the high penetration of mobile phones, which in Sweden and Italy were as high as 110% in 2006 [10], m-commerce has not taken off as hoped for [11, 12].

In the DigiNews and UbiMedia projects we have studied the potential of a new innovation for the media sector, i.e. the e-newspaper published on e-paper technology. As the e-newspaper introduction concerns both a new device as well as new content, it makes it an interesting case to study from an adoption point of view. As argued by Sarker and Wells [9], there is a need to understand adoption from the perspective of the consumers themselves. We have studied the potential willingness to adopt a future e-newspaper by presenting an online questionnaire resulting in 3626 respondents. However, it is very uncertain to trust what people think they want before having an actual experience of a product or service, we therefore also performed a user evaluation of an actual e-newspaper over a two week period with 10 families.

The research question in this paper is: *How does use experience influence perceptions of preferences and demands for the e-newspaper?* The aim is to contribute to the understanding of new media adoption as well as to contribute to the newspaper organizations preparations for launching the e-newspaper. This challenge will be studied using the e-newspaper case described below and by testing the following hypothesis: *users confronted with a vision of new technology and services are more positive to adopt than users with actual use experience of technology and services in an early stage of development with inherent technology problems.*

The structure of this paper is as follows. In section 2 the e-newspaper case is presented followed by a description of the research method in section 3. The theoretical framework is presented in section 4. In section 5 the findings are presented and section 6 discuss the findings and conclude the paper.

## 2 The e-newspaper Case

This research has been conducted within two projects, i.e. DigiNews (ITEA 03015) and UbiMedia (Designing Ubiquitous Media Services through Action Research). The research started within the DigiNews project, which was a two year project including partners from Belgium, Spain, Netherlands, France and Sweden and consisted of several major technology firms, media houses and universities. The overall goal was to explore research and development issues for the future e-newspaper, i.e. a newspaper published on e-paper technology. After the DigiNews project ended in mid-year 2006, the research continued within the UbiMedia project, which is a Swedish project with partners from 9 Swedish newspaper, the Swedish Newspaper Publishers´ Association and Stampen. This two-year project targets the challenge of designing ubiquitous media services for a multitude of devices and contexts to be consumed anytime and anywhere.

Electronic paper (e-paper) is the common term for several different technologies that can be used to produce screens with a number of specific characteristics. The e-paper is reflecting, giving the same reader experience as paper (such as high contrast, good color representation and the possibility to read in sunlight). The e-paper is thin, flexible and non-sensitive. In addition, it does not require high battery performance – ultimately, the screen image is stable and fix even when there is no electrical voltage applied.

The e-newspaper is predicted to combine the readability and overview from the printed newspaper with the possibilities of online media such as constant updates, interactivity and video [13], and is even predicted to replace the printed edition in the long run [14]. The potential replacement of the printed newspaper with the e-newspaper would dramatically reduce production and distribution costs for the newspaper companies.

The introduction of the e-newspaper has already begun, during 2006 two experiments with e-newspapers in real life settings has been performed, the first with the financial paper De TIJD in Belgium [15] and the second with Sundsvalls Tidning in Sweden which is one of the studies presented in this paper. In China the Yantai Daily Media Group started to publish an e-newspaper in October 2006. In all these examples the device iRex iLiad (Figure 1) was used, which is one of the two available "reading devices" on the market today, using e-paper technology.

**Figure 1: iRex iLiad [16]**

iRex Technologies BV, a spin-off from Royal Philips Electronics, launched the iLiad, a first generation electronic reader product in April 2006. The iLiad includes an 8.1 inch screen with 16 levels of grey and 160 dpi resolution, Wi-Fi, USB ports and MP3 capabilities [16]. Using a special marker, readers can comment on articles and scribble their notes on the screen.

**Figure 2: Sony Reader [18]**

The other device on the market is the Sony Reader (Figure 2), which was launched during the fall of 2006 on the U.S. market. The Sony Reader has a 6-inch screen, weight is less than 9 ounces and one can do 7.500 page views for each charge by an AC adapter. It can hold up to 80 eBooks at the same time, and allows PDFs, personal documents, newsfeeds, blogs and JPEGs. Sony offers books for the device on a new web site called Sony Connects [17].

Just to show how big this industry is expected to be, we give an example of analysts from IDTechEx who forecast plastic electronics will be a $30 billion industry by 2015, and could reach as much as $250 billion by 2025 [19].

# 3 Research Method

In the DigiNews and UbiMedia projects, described above, we have conducted several studies concerning audience preferences and demands of the future e-newspaper. In this paper we report from two of these studies, i.e. a survey with 3626 respondents and an evaluation of an early version of an e-newspaper with 10 families over a two week period.

**The Survey**

The survey was done through a web-based questionnaire. We presented the questionnaires at the news sites of the three Swedish newspapers that we have collaborated with in developing e-newspaper prototypes within the DigiNews project, i.e. Aftonbladet, Göteborgs-Posten and Sundsvalls Tidning (Table 1). Aftonbladet is a tabloid with the most visited news site in Sweden, Göteborgs-Posten is a local morning paper covering Göteborg (the second largest city in Sweden) and its surroundings, and Sundsvalls Tidning is a local morning paper in the north of Sweden.

| *Newspaper* | *URL* | *Unique visitors/day* | *No. of respondents* |
|---|---|---|---|
| Aftonbladet | aftonbladet.se | 1.200.000 | 3757 |
| Göteborgsposten | gp.se | 41.500 | 135 |
| Sundsvalls Tidning | st.nu | 14.500 | 447 |

**Table 1: Newspapers hosts for questionnaires and number of respondents**

The questionnaire was divided in four parts concerning background data, business models for electronic news, preferences for future electronic news and use of mobile media services. In total, 127 questions were asked and 4339 respondents answered the questionnaire. The respondents that had given an age under 15, those who did not complete or answered the questions included in this study contradictorily were excluded form the data set, resulting in a dataset containing 3626 respondents. In this paper we report from the background questions and questions regarding preferences for future electronic news.

We choose to use online questionnaires because that allowed us to show concept videos and prototypes for the respondents to obtain an understanding of the e-newspaper concept. Moreover, Buchanan and Smith [20] have argued that web samples can be as representative as or more representative than traditionally collected samples because of the heterogeneity of the online population. Although, admittedly there are inherent problems in controlling whom responds to online questionnaires. Control for cases with multiple submissions from the same IP number was handled in the data collection. Since the e-newspaper concept was not known to all potential respondents we provided them with the possibility to read more about e-paper technology on a separate page which consisted of a simplistic picture of the concept as well as links for further reading.

Further, we provided three concept videos of future e-newspaper scenarios in conjunction to the questionnaire for the respondents to watch. The movies envisioned the benefit of the e-newspaper for three different personas: the business women, the student and the senior citizen. Close ups on the designed user interface together with examples of functions showed the future e-newspaper in detail. The scenarios were based on the assumed preferences of the three personas and showed how a future e-newspaper could support their media consumption in different contexts. Watching these videos provided the respondents with an idea of what functionality the future e-newspaper could provide.

During the DigiNews project different prototypes (Figure 3) where developed for PC:s and tablet PC:s to be able to test conceptual ideas. These prototypes were developed together with newspaper designers and used contents from the newspaper partners. The prototypes also served as a way to explain how a future e-newspaper may look like and where presented with the introduction to the questionnaire. The respondents could download and test the prototypes on their own computer before they answered the questions.

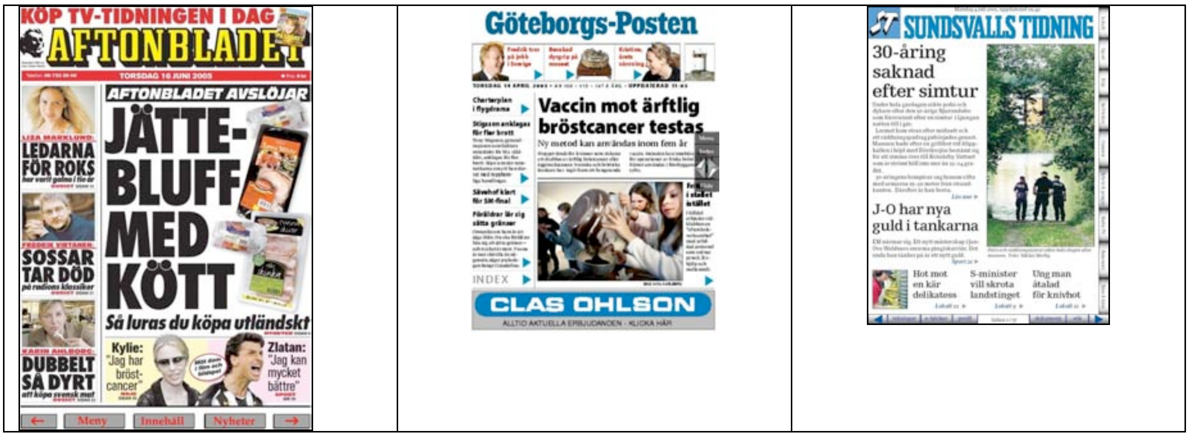

KÖP TV-TIDNINGEN I DAG
AFTONBLADET
TORSDAG 16 JUNI 2005
AFTONBLADET AVSLÖJAR
JÄTTE-BLUFF MED KÖTT
Så luras du köpa utländskt
LEDARNA FÖR ROKS
SOSSAR TAR DÖD
DUBBELT SÅ DYRT
Kylie: "Jag har bröst-cancer"
Zlatan: "Jag kan mycket bättre"
Meny
Innehåll
Nyheter
Göteborgs-Posten
Vaccin mot ärftlig bröstcancer testas
INDEX
CLAS OHLSON
SUNDSVALLS TIDNING
30-åring saknad efter simtur
J-O har nya guld i tankarna
Hot mot en kär delikatess
S-minister vill skrota landstinget
Ung man åtalad för knivhot

**Figure 3: Interactive e-newspaper prototype**

The questions about preferences for a future e-newspaper regarded the e-paper device as well as the content and services. Some of the questions were statements with a 7-grade Lickert scale and others were multiple choice questions. The responses to the questionnaire were analyzed using SPSS v14.0. The analysis focused on calculation of mean scores and standard deviations for each statement and on frequencies and percentages for multiple choice questions. The goal was to generate an overview of what that was comparable to the results from the e-newspaper test persons.

**The Evaluation**

The evaluation was conducted with 10 families who tested an early version of an e-newspaper (Figure 4) published on the iRex iLiad in real-life settings over a two week period in the autumn of 2006. The e-newspaper of Sundsvalls Tidning was published twice daily, at 6 pm and 1 am, and was downloadable via Internet. The respondents were foremost selected to represent different types of households such as singles, couples, families with children, and senior citizens, to secure different use patterns, but we also tried to get differences in gender, ages, occupation and education. In two of the families both adults participated in the evaluation resulting in a total of 12 respondents (but only one of the extra family members answered the questionnaire – giving a total of 11 respondents to that part of the evaluation).

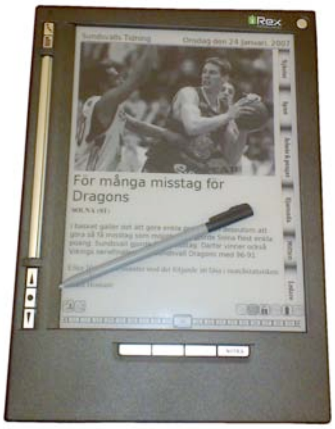


**Figure 4: E-newspaper prototype**

The two-week evaluation started with a meeting in Sundsvall, where the respondents were introduced to the device, and got a questionnaire about their media and reading habits. After two weeks of e-newspaper use, the respondents were visited in their homes for an interview about their experiences and preferences of the e-newspaper. A semi-structured interview approach [21], with an interview guide was used. These interviews were recorded and transcribed. Finally, they received a questionnaire consisting of 14 questions, partly matching the questions in the survey above.

However, as the e-paper technology in the iRex iLiad is in an early stage of development, there are some limitations compared to the e-newspaper prototypes described above. For example, the iLiad only presents 16 grayscale and have several limitations in the navigation system regulated by the device.

## 4 Theoretical Framework

One of the major topics in m-commerce research is user's adoption of mobile devices and services. Most of this research has been related to mobile telephony and services in 3G networks using theoretical frameworks like Innovation and Diffusion Theory [22] and Technology Acceptance Model [23]. Roger´s [22] Innovation and Diffusion Theory explains among other things the Innovation-Decision Process, which contains five stages. In the second stage, the persuasion stage, the general perception of the innovation is developed which is explained by the perceived attributes, relative advantage, compatibility, complexity, observability and trialability. The later two are related to how users can experience the new technology before adoption which is the topic of this paper [3]. Rogers [22] define *observability* as the degree to which the result brought by the technology and the technology itself is visible before adopting the technology and *trialability* as the degree to which a technology can be experimented with before adoption.

There are several studies addressing use experience and exposure in adoption processes. Sarker and Wells [9], for example, took an approach grounded in users actual practice and designed a framework for studying key issues related to mobile device use and adoption, including aspects of mobility. This framework is built as an input-process-output model. The inputs are user characteristics, communication/task characteristics, technology characteristics, modality of mobility, and surrounding context. The process consists of *exploration and experimentation* as one sub process and *assessment of experience* as another. Output refers to actual adoption behaviors, such as continuity of use over time. As this framework is based on study of motivations and circumstances surrounding individual's adoption and use of mobile devices, some of the issues described in the framework are related to communicational tasks such as voice communication, SMS, e-mail not applicable in this study.

In another study, the role of exposure to the adoption process was studied [3]. *Exposure* is defined as the degree to which an individual has acquired or exchanged information about the technology and its usage. The suggested model included exposure in form of *trial, communication*, and *observation*. The findings in this study suggest that the level of exposure of a new technology has an effect on the user's attitude towards that technology and thereby strengthens or weakens the user's intention to adopt. Trial and communication proved to be more effective than observation. The conclusion drawn is that exposure is likely to facilitate adoption of m-commerce.

In this study we are addressing adoption of a new technology together with its content. This was also the case in a study concerning adoption of services in mobile phones it was found that mobile services are adopted according to several patterns despite having the same technology base [5]. Allowing user to try the services rather than the technology was the focus in these tests. There are also studies indicating that *use situation and mobility* has a significant effect on intention to use a mobile service since user perceive services differently in different situations [24].

A summary of this literature review on factors influencing intention to use relevant for this study, i.e. related to use experience, are presented in Table 2.

| *Factor* | *Reference* |
|---|---|
| Observability and trialability | Rogers (1995) |
| Exploration, experimentation and assessment of experience | Sarker and Wells (2003) |
| Exposure (trial, communication, and observation) | Khalifa and Cheng (2002) |
| Trial of service (not technology) | Carlsson, *et al.* (2005) |
| Use situation and mobility | Mallat *et al.* (2006) |

**Table 2: Summary of factors related to use experience**

In this study we have let respondents experience the e-newspaper in two different ways. In the first study the respondents could observe visions of future e-newspaper usage in different situations watching concept videos. Further the respondents could try and experiment with e-newspaper prototypes. However, they were not exposed to the actual e-paper technology. In the second study, users tried the e-newspaper prototypes implemented in an e-paper device with its constraints in their everyday situations such as at home, commuting to work, at work etc. They were allowed to experiment with the e-newspaper prototypes for two weeks. In the following we describe and compare how these differences have affected the audience preferences and demands in the two studies.

## 5 Findings

In section we first present the background data to the respondents from the survey (Table 3) and the test persons in the evaluation. Thereafter we compare the results on preferences and opinions regarding the e-newspaper from the two studies. The test persons in the evaluation were 3 women and 9 men with the average age of 39,7. Their educational level was: elementary (3), grammar (5), and university level (5). 9 of the test persons work full time, 2 were students, and 1 a senior citizen. 9 of them subscribe to printed newspapers and all 12 read online news. Finally, 5 regularly use mobile services.

| **Demographic data** | *All* | *Men* | *Women* |
|---|---|---|---|
| *No of* | 3626 | 2216 | 1410 |
| *%* | 100 | 61,1 | 38,9 |
| *Mid age* | 37,1 | 37,9 | 35,7 |

| **Background data in percentage of total data set** | | | | | |
|---|---|---|---|---|---|
| *Occupation* | Full time | 55 | *Income* | Low | 30,0 |
| | Part time | 7,7 | | Medium | 41,9 |
| | Unemployed | 5,8 | | High | 28,1 |
| | Senior citizen | 6,3 | *Education* | Elementary | 9,1 |
| | Student | 19,0 | | Grammar | 45,5 |
| | Sick leave | 3,6 | | University | 44,0 |
| | Other | 2,6 | | Other | 1,5 |

| | *Newspaper subscriber* | *Read online news* | *Possession of mobile phone* | *Use mobile services* |
|---|---|---|---|---|
| Yes | 48,8 | 99,4 | 97,5 | 53,8 |
| No | 51,2 | 0,6 | 2,5 | 46,2 |

**Table 3: Demographic and background data of the questionnaire respondents**

The first topic for comparison regards the reason for considering exchanging the traditional printed newspaper with an e-newspaper. This question was asked in both studies and the respondents were given statements that they answered on a 7-grade Likert scale. The mean scored from both studies are presented in Table 4. *Availability anywhere* and *Added value such as new services* are the most important reasons for both groups. The least important reason in the survey was *Environmental reasons* and in the evaluation *Time savings*. Notable is that all reasons were rated higher by the test persons apart from *Time saving*. Some of the test persons mentioned that the e-paper device should not only contain their morning paper but also support all their reading, as illustrated by one of the test persons: "*I want to read everything on this device…it has to be good enough for that to make up for the inconvenience of downloading and updating. It has to be as good as what I have today*."

| *What reasons are critical if you sometime in the future would exchange your traditional printed newspaper to an e-newspaper?* | *Survey* | | *Test persons* |
|---|---|---|---|
| | *Mean* | *Std dev* | *Mean* |
| Environmental reasons | 3,6 | 2,59 | 4,8 |
| Cost savings | 4,0 | 2,64 | 5,6 |
| Time savings | 3,9 | 2,76 | 3,9 |
| Availability anywhere | 4,8 | 2,69 | 6,3 |
| Satisfaction with new technology | 4,0 | 2,64 | 6,1 |
| Added value such as new services | 4,2 | 2,60 | 6,0 |

**Table 4: Reasons for exchanging traditional newspaper**

The second question asked in both studies concerned what added services that the respondents regarded as interesting to include in an e-newspaper (Table 5). Both studies show that *Archive*, i.e. the possibility of saving newspapers from previous days, is the most interesting added service followed by *Personalization* and *Community information*. *Personal information* was regarded as the least interesting in both studies. Interestingly, the test persons scored all added services to be more interesting than the respondents in the survey. During the interviews the test persons gave additional input to preferred added services, e.g. the possibility of cutting out and save items from the printed edition was seen as an aspect that needed to be transferred to the e-newspaper. Some also mentioned the possibility of e-commerce as an attractive add-on. Other aspects mentioned were environmental, e.g. not cutting down trees and less decontamination due to less distribution by vehicles, and as one of the test persons said: "*I do not know how much time it would take to get used to it, but I think it is better than recycling old newspapers*".

| *Apart from reading the news, what added services do you think should be included in the e-newspaper?* | *Survey* | | *Test persons* |
|---|---|---|---|
| | *Mean* | *Std dev* | *Mean* |
| Personalization | 4,4 | 2,65 | 5,3 |
| Community information | 4,4 | 2,64 | 5 |
| Personal information | 2,9 | 2,40 | 3,4 |
| General information | 3,7 | 2,51 | 4,6 |
| Archive | 5,1 | 2,62 | 6,2 |
| E-commerce | 3,5 | 2,50 | 3,6 |
| Entertainment | 4,0 | 2,58 | 4,5 |

**Table 5: Added services preferences**

Next, the respondents were asked about the acceptable cost level compared to the printed newspaper (Table 6). In general, the test persons from the evaluation are more inclined to pay the same price or higher than the respondents from the survey. Some of the test persons mentioned that the e-newspaper had to be cheaper than the printed edition due to the saved printing costs for the publishers. Others mentioned that the price had to be lower than the printed edition, due to less content in the e-newspaper.

| *What cost level is acceptable for you to change to an e-newspaper?* | *Survey* | | *Test persons* | |
|---|---|---|---|---|
| | *No* | *%* | *No* | *%* |
| Cheaper than the traditional printed newspaper | 2119 | 58,4 | 5 | 45,4 |
| Same price | 371 | 10,2 | 4 | 36,4 |
| Can be more expensive if there is added value | 281 | 7,7 | 2 | 18,2 |
| Price is unessential | 171 | 4,7 | 0 | 0 |
| Missing | 684 | 18,9 | 0 | 0 |
| Total | 3626 | 100 | 11 | 100 |

**Table 6: Acceptable cost level**

Thereafter the respondents were asked how they think the e-paper device should be financed (Table 7). The most preferred model in both studies is inclusion in subscription.

| *How do you think the e-paper device should be financed/ paid for?* | *Survey* | | *Test persons* | |
|---|---|---|---|---|
| | *No* | *%* | *No* | *%* |
| Hire-purchase | 281 | 7,7 | 2 | 18,2 |
| Buy the device | 711 | 19,6 | 1 | 9,1 |
| Included in subscription | 1708 | 47,1 | 8 | 72,7 |
| Other | 592 | 16,3 | 0 | 0 |
| Missing | 334 | 9,2 | 0 | 0 |
| Total | 3626 | 100 | 11 | 100 |

**Table 7: Finance of device**

Further, they were asked about their willingness to exchange the traditional newspaper with the e-newspaper in the future (Table 8). Surprisingly, all test persons answered yes compared to two thirds of the respondents from the survey.

| *Would you consider to, some time in the future, exchange your traditional printed newspaper for the e-newspaper?* | *Survey* | | *Test persons* | |
|---|---|---|---|---|
| | *No* | *%* | *No* | *%* |
| Yes | 2431 | 67,1 | 11 | 100 |
| No | 1070 | 29,5 | 0 | 0 |
| Missing | 125 | 3,4 | 0 | 0 |
| Total | 3626 | 100 | 11 | 100 |

**Table 8: Willingness to exchange the traditional newspaper**

Moreover, they were asked within which time frame they would be ready to read their newspaper on e-paper (Table 9). In both studies there were surprisingly many that were prepared to exchange today or within five years. However, 25% of the respondents in the survey did not answer this question indicating that it was difficult to decide.

| *Within which time frame are you ready to read your newspaper on e-paper?* | *Survey* | | *Test persons* | |
|---|---|---|---|---|
| | *No* | *%* | *No* | *%* |
| Today | 1521 | 41,9 | 6 | 54,5 |
| Within 5 years | 683 | 18,8 | 5 | 45,5 |
| Within 10 years | 209 | 5,8 | 0 | 0 |
| Within 20 years | 88 | 2,4 | 0 | 0 |
| Never | 223 | 6,2 | 0 | 0 |
| Missing | 902 | 24,8 | 0 | 0 |
| Total | 3626 | 100 | 11 | 100 |

**Table 9: Time frame**

Finally, the survey respondents were asked about the factors that influenced their decision to read their newspaper on e-paper (Table 10). *Stable technology* and *Easy to find content* were the factors that scored the highest. The least influencing factor was *Observability of use***.** The test persons were asked similar questions in the interviews.

| *How important are the following factors for you choosing to read your newspaper on e-paper?* | *Survey* | |
|---|---|---|
| | *Mean* | *Std dev* |
| The appearance of the device | 4,0 | 2,72 |
| Continuous updates of news | 4,8 | 2,80 |
| Added functions such as chat | 3,1 | 2,46 |
| Easy to use and handle | 4,6 | 2,81 |
| Stable technology | 4,9 | 2,81 |
| Observability of use | 1,9 | 1,87 |
| Environment friendly | 3,8 | 2,72 |
| That it is the latest technology | 3,0 | 2,41 |
| Easy to find content | 4,9 | 2,82 |

**Table 10: Influencing factors**

One of the major concerns of the test persons was the navigation of the e-newspaper that was very constrained by the device. Almost everyone mentioned that the navigation needed to be improved if they should consider exchanging the printed edition with the e-newspaper. The other most mentioned issue that needed to be addressed was the refresh rate of the display, i.e. it needed to update faster in order to create a pleasant reading experience. Some but not all test persons regarded color as essential for exchanging the e-newspaper. One of the test persons expressed: "*Color would be fun, but it is nothing that I prioritize as essential, I would exchange it even if it is not in color*". News updates during the day was also found important by several of the respondents.

# 6 Discussion and Conclusion

In this paper we have addressed the research question: *How does use experience influence perceptions of preferences and demands for the e-newspaper?* We did so by testing the hypothesis that users confronted with a vision of new technology and services are more positive to adopt than users with actual use experience of technology and services in an early stage of development with inherent technology problems. This hypothesis proved to be false, in this case it was the other way around. In spite of the e-paper device technical constraints and the early prototypes, the test persons in the evaluation were more positive to the e-newspaper. On the one hand, the respondents in the survey were not able to experience the actual e-paper technology, they could only read about it and watch concept videos of the future e-newspaper vision and interact with e-newspaper prototypes to get an understanding of the concept. On the other hand, the test persons in the evaluation who experienced the e-paper technology first hand, also experienced the bugs and limitations in this early stage of technology. The e-newspaper prototypes available for online experience in the survey were all in color and had well functioned navigation systems and interaction possibilities whereas the Sundsvalls Tidning in the evaluation was presented in 16 grey scales and had limited navigation options and interaction possibilities due to limitations with the technology. When we set out to test the hypothesis we believed that the respondents in the survey that were confronted with a vision of a multimedia e-newspaper in color with added services would be more positive. Even so, there are similar patterns in preferences and opinions in both studies. The relations between reasons and importance of different added services were very alike, indicating the relevance of the results in both studies. We can conclude that our findings are in line with previous research, i.e. that experiencing technology and services in different ways and in different situations have impact on intentions to use. Even though the respondents in the second study were exposed to technical and navigational difficulties they were very positive towards adopting the e-newspaper. We believe that trying the e-newspaper during two week in their everyday setting as well as being able to experiment with the services in the actual e-paper technology have contributed to this positive attitude.

By empirically testing potential adoption of the e-newspaper with two different approaches, we have contributed to the overall understanding of new media adoption in general and the promotion of the e-newspaper in particular. To summarize the findings according to newspaper organizations preparations for launching the e-newspaper, the following can be derived: The e-newspaper need to contain archive functions. Providing added value by personalization and by offering community information would increase the potential adoption. If the newspaper organizations offer added value according to the audience preferences, they could expect the same willingness to pay as for the printed edition. However, before launching the e-newspaper, the navigation has to be improved as well as the refresh rate of the display. As most respondents preferred to have the device financed by inclusion in the subscription, this could be considered as the initial alternative. Finally, as almost half of the respondents stated that they were ready to start reading the e-newspaper already today, it is time for the newspapers to start preparing for the e-newspaper introduction.

Further research includes a major e-newspaper test in real life settings with 5 Swedish newspapers and 50 families at different locations in Sweden. This test will build on the results from the studies presented in this paper and will include more added value and more services.

# References

[1] TILSON, D.; LYYTINEN, K. AND BAXTER, R. A Framework for selecting a Location Based Service (LBS) Strategy and Service Portfolio, in Proceedings of the 37th Annual Hawaii International Conference on System Sciences, Hawaiii. (CD-ROM), Computer Society Press (10 pages), 2004.

[2] ANCKAR, B.;, AND DÍNCAU, D. Value-Added Services in Mobile Commerce: An Analytical Framework and Empirical Findings from a National Consumer Survey, in Proceedings of the 35th Annual Hawaii International Conference on System Sciences, Hawaii. (CD-ROM), Computer Society Press (10 pages), 2002.

[3] KHALIFA, M., AND CHENG, S.K.N. Adoption of Mobile Commerce: Role of Exposure, in Proceedings of the 35th Annual Hawaii International Conference on System Sciences, Hawaii, (CD-ROM), Computer Society Press (7 pages), 2002.

[4] HONG S-J.; TAM K. Y.; KIM J. Mobile data service fuels the desire for uniqueness, Communications of the ACM, Vol 49 No 10, 2006, pp. 89-94.

[5] CARLSSON, C.; HYVÖNEN, K.; REPO, P.; WALDEN, P. Asynchronous Adoption Patterns of Mobile Services, in Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Hawaii, (CD-ROM), Computer Society Press (10 pages), 2005.

[6] MALHOTRA, A.; SEGARS, A. H. Investigating Wireless Web Adoption Patterns in the U.S. Communications of the ACM, Vol. 48, No. 10, 2005, pp. 105-110.

[7] KNUTSEN, L. A.; CONSTANTINOU, I.D.; AND DAMSGAARD, J. Acceptance and Perceptions of Advanced Mobile Services: Alterations during a Field Study, in Proceedings of International Conference on Mobile Business, Sydney, 2005, pp. 326-332.

[8] KNUTSEN, A. M-Service Expectancies and Attitudes: Linkages and Effects of First Impressions, in Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Hawaii, (CD-ROM), Computer Society Press (10 pages), 2005.

[9] SARKER, S.; WELLS, J D. Understanding mobile handheld device use and adoption. Communications of the ACM, Vol 46, No 12, 2003, pp. 35-40.

[10] 30 Countries Passed 100% Mobile Phone Penetration in Q1. Available at: http://www.telecommagazine.com/newsglobe/article.asp?HH_ID=AR_2148 (January 30th, 2007)

[11] CARLSSON, C.; CARLSSON, J.; HYVÖNEN, K.; PUHAKAINEN, J.; WALDEN, P. Adoption of Mobile Devices/Services – Searching for Answers with the UTAUT, in Proceedings of the 39th Hawaii International Conference on System Sciences, Hawaii. (CD-ROM), Computer Society Press (10 pages), 2006.

[12] HAMMOND, K. B2C e-Commerce 2000-2010: What Experts Predict. Business Strategy Review, Vol. 12, 2001, pp. 43-50.

[13] IHLSTRÖM, C.; ÅKESSON, M.; NORDQVIST, S. From Print to Web to e-paper - the challenge of designing the e-newspaper, in Proceedings of ICCC 8th International Conference on Electronic Publishing, ELPUB 2004, Brasilia, 2004, pp. 249-260.

[14] IHLSTRÖM, C. The e-newspaper innovation - converging print and online, in Proceedings of the International Workshop on Innovation and Media: Managing changes in Technology, Products and Processes, Stockholm, 2005.

[15] Belgium: e-paper test launch. Available at: http://www.editorsweblog.org/news/2006/02/belgium_epaper_test_launch.php (February 3rd, 2007)

[16] What is the iLiad? Available at: http://www.irextechnologies.com/products/iliad (April 10th, 2007)

[17] E-paper E-merging. Available at:

[18] http://www.signonsandiego.com/news/computing/personaltech/20060123-9999-mz1b23epaper.html (January, 24th, 2007)

[19] Sony Reader. Available at:

[20] http://products.sel.sony.com/pa/prs/reader_features.html (February, 16th, 2007)

[21] Oak Investment Partners invests in Plastic Logic. Available at: http://www.idtechex.com/printelecreview/en/articles/00000399.asp (December 19th, 2006)

[22] BUCHANAN, T.; SMITH, J. L. Using the Internet for psychological research: Personality testing on the World-Wide Web. British Journal of Psychology, 90, 1999, pp. 125-144.

[23] PATTON, M. Q. Qualitative Research & Evaluation Methods (3 ed.). Sage Publications, Inc. California, 2002.

[24] ROGERS, E. M. Diffusion of innovations. New York, The Free Press, 1995.

[25] DAVIS, F. D.; BAGOZZI, R. P.; AND WARSHAW, P. R. User acceptance of computer technology: A comparison of two theoretical models. Management Science, 35(8), 1989, pp. 982-1003.

[26] MALLAT, N.; ROSSI, M.; TUUNAINEN, V. K.; ÖÖRNI, A. The Impact of Use Context and Mobility on the Acceptance of Mobile Services, in Proceedings of the 39th Hawaii International Conference on System Sciences, Hawaii. (CD-ROM), Computer Society Press (10 pages), 2006.

# Centralized Content Portals: iTunes and the Publishing Industry

*Matthijs Leendertse; Leo Pennings*

TNO Information and Communication Technology, Department: ICT & Policy
Brassersplein 2, 2600 GB Delft, the Netherlands
e-mail: matthijs.leendertse@tno.nl; leo.pennings@tno.nl

## Abstract

This paper addresses new questions around media performance as a result of the rise of centralized content portals such as iTunes or MySpace. We first describe the rise of centralized content portals in different media industries, and discuss how these portals are creating a dominant position for themselves by using lock-in strategies. Then we describe the concept of media market performance, and discuss two important media performance concepts: access and diversity. Using scenario analysis, this paper describes three learning scenarios that outline the effects of different configurations of centralized content portals on behavior of publishers, users and advertisers, and through that content on access to and diversity of content.

**Keywords:** content portal; iTunes; access model; scenario analysis

## 1 Introduction

Centralized content portals are platforms that allow third party content suppliers to offer their digital content products to consumers. These platforms are typically developed by firms that are not rooted in the content industry, and as a result are based on business models that do not necessarily revive around content itself. Mobile operator portals for instance have been developed by mobile operators to facilitate content owners to sell content to users of mobile devices, with the intention of promoting their mobile internet services. NTT DoCoMo has been very successful with this concept in Japan with over 47 million subscribers in February 2007. The i-mode ecosystem allows mobile phone users to access a mobile portal where third party content providers that – abiding the standards of NTT DoCoMo – offer their content to end users. This content is charged on the mobile phone bill and NTT DoCoMo takes a cut of the revenues. Within Europe and the US, mobile operators have tried to mimic NTT DoCoMo's offering, but never succeeded in attracting significant groups of users.

One of the most prominent examples of centralized content portals in our part of the world has been Apple's iTunes Store. The stellar success of the iPod and its easy to use music software iTunes prompted Apple to start selling music through its iTunes Store. This centralized content portal is integrated in the iTunes software. With this integration between hardware, software and a centralized content portal, Apple proved to record companies that digital music could be monetized. One of the strengths of the iTunes Store has been the seamless integration of software (iTunes) and hardware (the iPod), allowing for purchases in the iTunes store to be immediately transferred to the iPod when the user connects its iPod to the PC. The sale of music is not particularly profitable for Apple as most of the revenues have to be transferred to the right holders, mostly to record companies. Apple's profits come from selling the hardware (iPods).

So what do these centralized content portals have in common? First and foremost they increase the ease of finding, selecting, purchasing and distributing digital content for end users. Second, these portals are usually not designed around the requirements of content suppliers. Third, these portals tend to exclude rival services. In other words, users of NTT DoCoMo's i-mode service or iPods are disabled or discouraged to access rival centralized content portals. The integration of hardware, software, payment mechanisms and in the case of mobile operator portals also network connectivity, effectively locks users into a centralized content portal.

## 2 Natural Tendency Towards Dominant Platforms

"What you see on the internet is very much the 'Highlander Theory'. There can be only one. There's only one search engine, there's only one big book retailer, one big online auction house, and so on. That's not necessarily a bad thing, as long as it's able to supply a super hot service at a reasonable price" (Greg Eden of AIM Digital in the Guardian Newspaper of August 17, 2006).

Media markets have a natural tendency towards concentration. The Dutch Media Authority labelled this tendency as the law of 3; in each media market (e.g. television, news or book publishing) in the Netherlands there were 3 large media companies competing [1]. Within the digital media sphere, there is not so much a tendency towards concentration in terms of content creators or packagers, but more in the sphere of facilitating services. Microsoft Windows is by far the dominant operating system; Google by far the dominating search engine (in particular in Europe) and iTunes dominates the paid for music downloads market. As Eden notes, indeed there can only be one.

Media technology that is used to distribute digital media to consumers can be divided into 4 main categories: PC based online services, digital television based services, Game Consoles and Mobile Platforms [2]. In all 4 categories we find evidence for the Highlander theory: a natural tendency towards a dominant standard:

- PC based online services: here content is downloaded over the open internet, and is typically accessed on a PC but also on audiovisual equipment using the PC as the central server. Apple pioneered this market with its iTunes Music Store and established a dominant position on the market for paid for music downloads. We can see the publishing industry is following this strategy. In 2006, Sony launched the ‘Sony Reader’, a device that allows users to access digital text through a device that provides an experience similar to paper. Sony copied Apple’s strategy for digital music, and bundled their Sony Reader with their so called ‘Sony Connect Store’ offers access to (premium) e-publishing products. This store is integrated in the software that synchronizes the content on the end user’s PC with the Sony Reader, again similar to iTunes.
- Digital Television based services: the digital television platform is increasingly used to distribute content other than video. Subscribers to a digital television service can only use the set top box of their supplier, and hence only access the (often third party) content these DTV suppliers offer. In other words, the digital television platform of the supplier (e.g. a cable or satellite company) is effectively the dominant standard for content for subscribers of that platform.
- Game consoles: all three so-called next generation game platforms – Microsoft’s Xbox 360, Sony’s PlayStation 3 and Nintendo’s Wii – feature an internet connection and have some sort of content offering beyond traditional gaming. Nintendo’s Wii for instance, offers a news channel where stories are linked to a map of the world and end users can search for content by scrolling the globe. Content for now is provided by the Associated Press and the service is offered for free. On its Xbox 360 platform, Microsoft offers a vast array of television and movie content to its US based clients and outlined plans to distribute the service in other territories as well. As owners of game consoles typically do not own game consoles of other brands, these users can only use the platform of Sony, Nintendo or Microsoft.
- Mobile platforms: mobile devices are increasingly sophisticated and have increasingly access to mobile data networks such as UMTS or WiFi 802.11. One of the main advantages of mobile products is that it can easily recycle existing content into a market where people are more willing to pay for access to information. However, yet again there are dominant standards emerging. For instance, many operators have mobile operator portals through which third party content can be searched, purchased and downloaded. The strongest asset of mobile phone companies with regards to mobile content is their payment system and payment relation with their consumers. Third parties that want to are dependent on the mobile phone company, that holds a near monopoly. Vodafone for instance discourages its subscribers from using other mobile content than offered through their Vodafone Live! mobile operator portal by charging for all data traffic outside the Vodafone Live! domain.

That begs the question as to why the rise of such dominant Highlander-esque portals is an important topic.

## 3 The Power of Lock-In

“The European Commission can confirm that it has sent a Statement of Objections to major record companies and Apple in relation to agreements between each record company and Apple that restrict music sales: consumers can only buy music from the iTunes' on-line store in their country of residence. Consumers are thus restricted in their choice of where to buy music, and consequently what music is available, and at what price. The Commission alleges in the Statement of Objections that these agreements violate the EC Treaty's rules prohibiting restrictive business practices (Article 81)” [3].

As the quote above demonstrates in the case of the European Commission against Apple and several record labels, dominant content portals can damage the interests of citizens. Suppliers of these content portals – be they Apple or Vodafone – can for instance determine pricing, availability, ranking, disclosure and usage restrictions on the content sold through their portals. This case of the EC against Apple is one of many. The Norwegian

Ombudsman for instance declared on January 24th 2007 that the iTunes Music Store is illegal because it only allows purchased music to be played on Apple's iPod devices. Rivalling MP3 players cannot be used to play the purchased content, effectively locking-in customers to their proprietary system. Such a lock-in strategy intends to prevent buyers from turning to alternative suppliers. For suppliers such strategies seem advantageous because lock-in allows them to raise prices without having to invest in innovation or product quality. Consumers however are dependent on one supplier, and their interests are potentially endangered. In addition, competitors of Apple are effectively restricted from engaging in competition at all, raising questions with regards to competition policy.

Lock-in strategies have existed for a long time within the media industry, mostly in the form of subscriptions, and are on the increase due to new media technologies. The most common lock-in strategies are contractual agreements, taking advantage of durable purchases that demand complementary compatible purchases in a later stage, supplying products that demand brand-specific training, developing propriety information and database standards that are not compatible with other databases, becoming the specialized supplier for specific products and offering loyalty services. [4, p. 117]. Most of these strategies involve increasing the costs for consumers to switch to an alternative supplier. This is especially apparent in markets where suppliers have the exclusive rights to a particular technology or system. Because digital media products involve many different layers of the communication system, suppliers can develop proprietary technology to take advantage of this. It is therefore not strange that within digital media, we see this tendency towards centralized content portals that effectively lock-in their users. The next natural question is how we can assess the effects of these portals.

## 4 Assessing Consequences of Centralized Content Portals

So how can we assess the effects of such centralized content portals on digital media markets? Within media economic theory, these effects can be studied by using media market performance criteria. The concept of market performance comes from welfare economics, where performance is traditionally assessed by economic indicators such as allocative efficiency and industry profitability ratios [5, 6]. Media economic scholars have adapted the notion of market performance to assess media markets. Rather than focusing solely on economic indicators, media market performance is assessed by social, political and cultural indicators such as media diversity, freedom of expression, access to media outlets and services are the most prominent [5, 7-9]. Media market performance could be described as an assessment of mass media from a public interest perspective [5, p. 62].

It is important to outline that market performance refers to the outcome of the total market, and should not be mistaken with the performance of individual firms or other actors. "Performance is, first and foremost, appraised with reference to a market, which comprises all the interacting buyers and sellers as a whole, rather than to individual economic agents such as firms" [10, p. 4]. The normative approach to market performance proposes several performance indicators to assess whether markets deliver what society wants [6].

Media policy based on media performance is based on media performance assessment, and typically concentration within the media is not greeted with great enthusiasm. The rise of centralized content portals that effectively lock-in consumers raises new questions around familiar media performance criteria, most notably access and diversity.

## 5 Access

Accessibility has also been an important criterion for media regulators, and has become more prominent in debates on the future of media policy. In contemporary media policy, access to communications is an central concept [8]. Access to communications can be defined as "the possibility for individuals, groups of individuals, organizations and instructions to share society's communications resources" [11, p. 204]. Access can be looked upon from different perspectives, such as access to markets or consumers or access to content by users. An important access performance indicator for users is affordability of content [8]. With an abundance of content available, access to that content is becoming increasingly important. Issues such as media education for groups that do not have the skills to access this content, households that do not have access to new media devices and infrastructures, but also economic accessibility as some content can only be accessed through payment.

Also, the availability of content is an important indicator of accessibility. For content suppliers, access to markets is an important factor. When for instance Apple would prevent some record companies from selling their content through iTunes, a large part of the digital music buying audience is shielded off from this content. This is not only detrimental to the record company in question, but also to the audience as it limits their access to content. With regards to publishing products, in particular news, educational and professional content, accessibility is an even more important issue. In light of the democratic and socio-cultural functions that media have apart from economic functions [12], it is important that citizens of democratic societies have access to information. When for instance a centralized portal for news content would be as dominant as the iTunes Store, restricting access of content suppliers to this platform could seriously undermine the democratic process.

In addition to publishers of information, access to these platforms is also important for advertisers. We see that many platforms are replacing traditional advertising outlets as the main facilitator of advertisements. Online, we see that Google is now dominating the advertising industry using its Adsense and Adwords advertising network. Within the digital television domain, we see that cable companies and IPTV providers moving towards the advertising market as well, taking over the roles of traditional broadcasters [13].

# 6 Diversity

Diversity is perhaps an even more important media market performance criterion in Western countries. The adjectives to diversity in government reports usually reflect the desired media performance: cultural diversity, opinion diversity, regional diversity, genre diversity, ethnic diversity etc. The concept of media diversity could be defined as the heterogeneity of the media[14, 15] McDonald and Dimmick [15] argue that diversity is a two dimensional construct: [1] a set of categories within a given distribution (e.g. content categories) and [2] the allocation of elements to these classifications (e.g. how many programs are devoted to the content category news).

The concept of media diversity can be deconstructed into three distinct forms of media diversity: source diversity, content diversity and audience exposure diversity [16]. Source diversity refers to the number of media outlets (TV channels, newspapers) and the ownership structures of these outlets and is traditionally measured using economic measures for competition such as the HHI index or Competition Ratios. Content diversity refers to actual media supply, and is mostly assessed by content analysis studies. These studies typically classify media content into predefined categories, for instance into content categories. Most policy research around diversity assumes that audiences provided with a diversity of content options also consume a diversity of content. However, the mere availability of diverse information does not necessarily equal exposure to diverse opinions and information. Without audience exposure to diverse content, availability of content has no effect on the political and socio-cultural functions of media. Exposure diversity is defined as "the diversity of content or sources consumed by audience members"[16]. Many indicators for media diversity have been formulated, as listed below:

- Media should reflect the various social, economic and cultural realities of the societies in which they operate, more less proportional.
- Media should offer more or less equal chances of access to the voices.
- Media should service as a forum for different interests.
- Media should offer relevant choices of content at one point in time and also variety over time. [17].

With the increased importance of centralized content portals, it is important to reassess to what extend these diversity indicators are being met.

# 7 Scenario Methodology

To hypothesize the effects of different configurations of content portals, we conducted a scenario exercise. The main purpose to develop scenarios is that they should paint distinct different pictures of the future with unique implications for strategic decision-making [18, 19]. We use secondary sources and media economic theory to develop the scenario lines [20]. In our case, we wanted to develop scenarios that explain the effects of open or closed and commercial and non-profit content portals on access to and diversity of e-publishing products. Therefore, we developed three sets of conditions for the scenarios, each with their own unique configuration of content portals:

1. One dominant content portal similar to the iTunes Music Store, owned by a provider of e-publishing hardware (often referred to as e-readers);
2. Several interoperable open content portals that have been developed by commercial search engines and software companies;
3. A web full of Wikis with freely accessible co-created content, in tandem with securely sealed off walled garden of traditional publishers.

In each of the scenarios we established the so-called rules of interaction [21]. These rules outline how the important actors respond to the above mentioned conditions. The actors that we included in the scenarios are based on the value chain: content creators, content packagers (publishers) and content distributors [12]. In addition, we included advertisers as an important actor for the simple reason that many consumer based publishing products are dependent on advertising revenues. Based on this, we developed a linear story line that depicts not only the conditions and rules of interaction, but also the consequences for access to content and the diversity of content. In order to keep the scenarios clear, the broader conclusions are discussed in a separate concluding paragraph. We choose a time path of 5 years for the scenarios, placing them in the year 2012. This time path was selected because the technological progress is advancing at such a rapid pace that looking further into the future (e.g. 10-15 years) would be near impossible.

These scenarios aim to help policy makers to assess the performance of future e-publishing markets based on content portals, and support the policy-making process. Policy is based on beliefs around the future benefits of content. That is why policy makers often conduct ex ante policy assessments, whereby the effects of different policy options are weighed [22]. Managers within e-publishing companies can use these scenarios to evaluate the strategic position of their organizations under different configurations of content portals, and can hence formulate counterstrategies [21].

# 8 Scenarios

Please find below three scenarios, in which we discuss the effects of different configurations of content portals on the behavior of the main market actors, and hence on accessibility and diversity of content. The scenarios do not try to describe utopist visions, or intend to prescribe one future as the best. Assessment of these scenarios is a normative matter for policy makers and an economic / strategic question for commercial publishers.

# 9 IPUB (Proprietary Standard)

The year is 2012. After the stellar success of the iPod, Apple has ventured into other content markets as well. In 2008, the company launched a hard disc based text reader dubbed the iPub. Within 4 years, this device has captured 85% of the market for digital reading devices. Yet again, Apple proved that user friendly hard- and software can significantly increase end users' appetite for digital content. In all major cities, people are reading their iPubs in public transport, in restaurants and increasingly also in schools and offices. Traditional publishers are now selling their content directly to the iPub, making use of its Wimax wireless connection. Digital newspapers or magazines are directly downloaded to the iPub when the user has a subscription. The iTunes store can be accessed on both the PC / Mac and on the device itself and offers the largest collection of e-publishing products in the Western world. In terms of sales, the iTunes store has overtaken Amazon.com, Barnes & Noble and Bol.com as the largest reseller of e-publishing products in both the US and Europe. The iTunes store provides copyright protected e-publishing products that can only be viewed on an iPub. In terms of prices, Apple has set fixed prices for different product types so that users are confronted with a simple to understand pricing mechanism. Books for instance can be priced at €5, €10, €15 or €20 in the German iTunes store. These prices are set by Apple, and individual suppliers that want to sell their products using the iTunes store have to abide their pricing scheme. Apple gets a fixed share of 20% on all sales. In exchange the company deals with all the handling, the platform and the payment mechanisms. Because of the dominant position of Apple's iPub on the e-reader market, Apple controls the dominant platform for premium e-publishing products. Other OEMs have significantly less power, and are therefore not able to attract content owners, let alone dictate pricing and format standards to them.

Traditional publishers dominate the premium e-publishing products in iTunes. They have signed agreements with Apple to distribute their products, and sometimes demand certain special features such as additional promotion in iTunes and exclusion of certain rival products. Smaller or even individual content creators such as journalists or writers lack the bargaining power of the large publishers. As a result, they pay higher royalties to Apple for selling their e-publishing products and find it difficult to negotiate special deals or promotional activities. Although Apple promotes itself as a corporate responsible company, it remains a commercial company and money logic dictates that Apple will give priority to the large publishers, and put the squeeze on their smaller rivals.

Publishers can also use the iTunes store to provide ad supported content for free. Because Apple does not earn money on freely distributed content by its cut on each purchase, it charges for delivery of ad supported content to end users through its platform.

The dominant position of the iTunes Store for e-publishing products makes it a walled garden in which Apple dictates standards with regards to content packaging, price setting, copyrights technology and disclosing and distributing content. Large publishers have preferred access to the iTunes Store and are more heavily promoted by Apple. Smaller suppliers find it more difficult to access this platform. As a result, many niche e-publishing products are not present or hard to find in the iTunes Store. It is primarily the mainstream content that is promoted, similar to the music offering in the iTunes Store. Already in 2006, the Guardian newspaper reported that: "(…) iTunes does sell a reasonable volume of niche music, but as a mainstream music retailer, it markets to and mostly attracts mainstream music fans" [23]. Four years after the introduction of the iPub, the same has happened for e-publishing products. Sure, the iTunes store contains a diverse range of content, but effectively the content offering contains of mainstream mass media e-publishing products. Consumers that do not own an iPub or do not have a iTunes Store account are barred from many e-publishing products as these are exclusively available on the iTunes Store. Many publishers use the iTunes Store as their sole or prime distribution platform for e-publishing products, not in the least because Apple's pricing policy allows them to set relatively high prices for their e-publishing products. The relatively high prices combined with the enormous reduction in printing and distribution costs make an interesting business case for publishers.

## 10 Interoperable Content Portals

It took some time, but in the three years from 2009 up till now, e-publishing has finally matured. Several open content portals have been developed that are used by a myriad of content suppliers. These platforms are truly cross-medial, i.e. they can be accessed on the open internet, on mobile devices and on e-readers with network connectivity. Search engines and software providers are the main drivers behind these open platforms. These facilitators use their content portals to push their payment systems and advertising networks. Google, Paypal (eBay) and Nokia Software are the three leading providers of such platforms in Europe, and offer a payment system for purchases of e-publishing products. These payment systems can also be incorporated in other websites, so content providers can also offer, sell and distribute content through their own portals.

International open source standards are used for content packaging, copyrights and metadata in order to optimize interoperability and retrievability of e-publishing products over various portals. Content owners can easily distribute their content to the different portals, and the open character of these content portals ensures that also small publishers can find their way to end users. As content is organized along the lines of standard metadata sets, the size of the publisher does not influence the degree of retrievability on the large portals. However, these larger publishers also have an extensive online presence themselves, which makes up almost 40% of their turnover in the e-publishing domain. For these publisher websites they often prefer to use the payment system and advertising networks of facilitating companies over proprietary systems.

The large content portals for their part are used by a large variety of publishers, from individual authors to multinational media conglomerates. Standardized open platforms with standardized copyright protection reduce transaction costs [24, 25]. Since the size of a firm depends foremost on the question whether it will pay to bring an extra exchange transaction under the organizing authority of a firm [26], lower transaction costs as a result of standardization would dampen the economic case for organizing transactions within a firm. In other words, the raison d'être for large publishing houses as content packagers is becoming rather questionable. More and more content creators bypass the publishing houses and directly market their products using open content portals.

Because these platforms are open, several non profit organizations are also accessing the market. This results in a negative price spiral. For instance, the rivalry between the free news e-publishing products of public broadcasters such as the BBC and ZDF and those of commercial newspaper has driven most quality papers to an advertisement supported free e-newspaper model.

The large content portals use the traditional motto of telecommunication providers that they are not interested in the message, but only in the facilitation of the message. This content agnostic view has allowed them to incorporate a myriad of content suppliers from various backgrounds, such as political parties, student organizations, individual authors or educational publishers. As a result, end users have access to a very diverse pallet of e-publishing products.

The high level of competition between suppliers drives down prices of e-publishing products, which in turns increases dependency on advertising revenues. This benefits the advertisement networks of the companies behind these content portals, such as Google's AdSense and Nokia's Ad Accelerator. Especially the smaller publishers do not have the means or the know-how to sell advertisements for their products, and can easily tap into the systems of the large advertisement networks. We predict that in April 2017, 5 years from now, advertisements will make up 70% of the income of e-publishers. Because the suppliers of centralized content portals monitor user behaviour on the portals, but also their reactions to certain types of information, they can greatly increase the effectiveness of advertisements in e-publishing products, which has a positive effect on demand for and prices of advertisements in e-publishing products.

## 11 iWiki-publishing

It is hard to imagine that only 5 years ago, Wikipedia was primarily an online encyclopedia. We have come a long way since then. The managing board of Wikimedia declared on January 4th 2012 that the Wiki format has now become the dominant method of publishing for digital text, photo and audiovisual content. Although the large publishing companies still play an important role with their print and online propositions, in terms of digital content Wikis are the undisputed market leaders in Europe, the US and East Asia in all major digital content categories, but especially when it comes to news content. Recognizing the importance of the Wiki movement, iWiki-publishing has been added to the Wiki and Oxford Dictionaries as:

> *"i-wi-ki-pub-lish-ing Pronunciation [ai- wee-kee-pub-lish-ing] –derived from a verb. Meaning: an e-publishing product that has been created on or for an open platform and is continuously subject to changes from visitors".*

As there is no central organization that creates or packages iWiki publishing products, there are no business models around this type of content. When Wikis were still relatively unimportant in the media industry, many non-profit organizations started to use this mechanism. Public broadcasters were particularly instrumental in promoting the Wiki concept, and en masse started to use Wiki from 2008 on. As a result, there is a large variety of iWiki publishing products and these are not only free, but also void of advertisements. The servers and sites that host the Wikis are maintained by donations from individuals. These servers and sites are non-profit foundations with a democratically elected managing board. Wikis are always non-profit, because it is impossible to divide the proceedings of co-created content over all the co-creators.

This leads us to the problem of iWiki-publishing. As there are no business models, it is hard to find sufficient funding for more specialized or professional content. Investigative journalism for instance requires relatively large investments, with journalists sometimes having to infiltrate organizations for prolonged period of time. The traditional print publishers are aware of this problem and offer their often more specialized content for free to subscribers of their printed material. They have created online walled gardens where subscribers can access the digital version and extras around the printed product. By doing so, these traditional publishers effectively defend their positions in print and lure users with a need for specialized content to their printed products. Product innovation for these more specialized forms of digital content such as investigative journalism are dependent on innovation in the portfolio of printed material that these traditional publishing houses have. Because they have a semi-monopoly on specialized content, prices for printed material are increased in order to make up for costs of their digital offering. This started at the beginning of the century with educational publishers that increased prices of text- and workbooks to cross subsidize investments in 'free' digital content [27], and has now spread to all e-publishing segments.

For advertisers it has become difficult to reach audiences of e-publishing products. They are restricted to the walled gardens of the traditional publishers, and these only give access to a large minority of the total audiences. Although these audiences are interesting for many advertisers (the average subscriber to printed publications are more affluent than the average reader of free iWiki-publishing products), e-publishing has become less interesting to advertisers than other media. This also benefits the traditional publishers, as they are often part of a larger media conglomerate that can utilize alternative media to lure advertisers to their platform. Nevertheless, access to high quality, specialized e-publishing content has become more difficult and content innovation is dependent on end users.

All in all, Wikis have greatly benefited access to content distribution platforms for individual authors and provide a diverse supply of digital content. However, the counter reaction is that specialized digital content is now more disclosed then ever and less affluent consumers might have difficulties to access this content.

## 12 Conclusions

The answer to the question to what extend different centralized content platforms are beneficial or detrimental to the e-publishing sector, is dependent on the subjective evaluation of the person who asks. Publishers have different interests than politicians, and we can be sure that politicians rooted in different political ideologies also differ in their evaluation criteria. However, it is safe to say that centralized content portals will most likely change the way e-publishing products are packaged, disclosed and distributed.

In the first scenario, we found benefits for publishers that are allowed access to the iTunes store. Prices are relatively high, whereas the percentage that has to be paid to iTunes in return for facilitating the sale of e-publishing products small in comparison with for instance traditional bookstores. Consumers might evaluate this scenario less favorably, as they are locked into hardware and can choose from a pre-selected set of content for relatively high prices. Governmental agents might also be less than happy, especially given the lack of competition.

The second scenario might be less interesting for publishers, as it remains difficult for them to develop profitable services in a very competitive environment. Publishers that want to attract users must invest in quality and innovation, because only then users will be persuaded to pay for access to e-publishing products. Smart use of the networks of the facilitators might help smaller publishers to develop profitable business models based on the advertising market. Consumers have the best deal in this scenario, because they can select content from an infinite number of sources, and prices are very subdued. Regulators might also look favorable upon this scenario, as competition is high.

The final scenario is a more scientific approach, whereby knowledge is not created by any organization but by a collective of co-creators. Publishers won't cheer for this scenario, since it is near impossible to create profitable e-publishing products. Rather, e-publishing has become a by product for printed material. There is no such thing as a free lunch, so money has to come from alternative sources. For consumers it is also perhaps a less positive scenario as access to specialized content is somewhat restricted and the relevance and authenticity of content can be questionable. Governments should also consider to what extend this type of market, where digital content is either free or a by product, is desirable for their policy goals.

These scenarios intend to help policy makers and strategists within publishing companies think about the future impact of centralized content portals. Going forward more empirical research into the effects of these portals on media market performance is required. Especially the reaction of e-publishers to these portals should be further analyzed. The next step of the research would be to develop quantitative models that predict reactions of e-publishers to the rise of centralized content portals. This enables us to empirically test the validity of the predictions made in the scenarios.

## Acknowledgements

This paper has been written within the framework of the FLEET project, a multidisciplinary research project on Flemish e-Publishing Trends that is financed by IWT.

## References

[1] Commissariaat voor de Media, *Mediaconcentratie in Beeld. Concentratie en Pluriformiteit van de Nederlandse Media 2001*. 2002.

[2] VAN WOLFWINKEL, R.; LEENDERTSE, M. *Deploying Broadband Services in a Competitive Environment*. in *Broadband Europe*. 2006. Geneva, Switzerland.

[3] Commission of the European Communities, *Competition: European Commission confirms sending a Statement of Objections against alleged territorial restrictions in on-line music sales to major record companies and Apple*. URL: (http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/07/126&format=HTML&aged=0&language=EN&guiLanguage=en [Accessed: 06-04-2007].

[4] SHAPIRO, C.; VARIAN, H.R., *Information Rules - a strategic guide to the network economy*. 1999, Boston: Harvard Business School Press.

[5] HENDRIKS, P., *Communications Policy and Industrial Dynamics in Media Markets: Toward a Theoretical Framework for Analyzing Media Industry Organization.* The Journal of Media Economics, 1995. **8**(2): p. 61-76.

[6] SCHERER, F.M.; ROSS, D., *Industrial Market Structure and Economic Performance*. 1990, Boston: Houghton Mifflin Company.

[7] VAN DER WURFF, R.; VAN CUILENBURG, J., *Impact of moderate and ruinous competition on diversity: the Dutch television market.* The Journal of Media Economics, 2001. **14**(4): p. 15-26.

[8] VAN CUILENBURG, J., *On competition, access and diversity in media, old and new: some remarks for communications policy in the information age.* New Media & Society, 1999. **1**(2): p. 183-207.

[9] MCQUAIL, D., *Media Performance.* 1992, London: Sage.

[10] WAYNE FU, W. *The S-C-P Framework: applying the structure-conduct-performance framework in the media industry analysis.* in *AEJMC Annual Convention.* 2003. Kansas City.

[11] VAN CUILENBURG, J.; MCQUAIL, D., *Media Policy Paradigm Shifts.* European Journal of Communication, 2003. **18**(2): p. 182-207.

[12] BARDOEL, J.; CUILENBURG, J.V., *Communicatiebeleid en Communicatiemarkt.* 2003, Amsterdam: Otto Cramwinckel.

[13] LEURDIJK, A.; LEENDERTSE, M.; DE MUNCK, S., *Reclame 2.0 De toekomst van reclame in een digitaal televisielandschap (Advertising 2.0. The future of advertising in a digital television landscape).* TNO: Delft.

[14] VAN CUILENBURG, J., *On monitoring media diversity, media profusion and media performance: Some regulator's*

a. *notes.* Communications, 2005. **30**: p. pp. 301-308.

[15] MCDONALD, D.; DIMMICK, J., *The Conceptualization and Measurement of Diversity.* Communication Research, 2003(February): p. 1-10.

[16] NAPOLI, P.M. *Television station ownership characteristics and commitment to public service: an analysis of public affairs programming.* in *the Association for Education in Journalism and Mass Communication.* 2002. Miami, USA.

[17] MCQUAIL, D., *Mass Communication Theory.* 2000, London: Sage Publications.

[18] CHERMACK, T.J.; MERWE VAN DER, L., *The role of constructivist learning in scenario planning.* Futures, 2003. **35**: p. 445-460.

[19] COURTNEY, H.; KIRKLAND, J.; VIGUERRIE, P., *Strategy under uncertainty.* Harvard Business Review, 1997(November - December): p. 67-79.

[20] LEENDERTSE, M. *Balancing Business and Public Interests - Theoretical scenarios on how standardization & copyrights regimes impact the structure, conduct and performance of the market for learning objects.* in *World Media Economics Conference.* 2004. Montreal, Canada.

[21] SCHOEMAKER, P.J.H., *Scenario Planning: A tool for strategic thinking.* Sloan Management Review, 1995. **Winter 1995**: p. 25-40.

[22] HOOGERWERF, A., *Beleid, processen en effecten*, in *Overheidsbeleid*, A. Hoogerwerf, Editor. 1998, Samsom Uitgeverij: Alphen aan de Rijn.

[23] The Guardian, *A musical tail of hits and misses.* URL: http://arts.guardian.co.uk/netmusic/story/0,,1852005,00.html [Accessed: 25-03 2007]. 2006.

[24] FUNK, J.L.; METHE, D.T., *Market- and committee-based mechanisms in the creation and diffusion of global industry standards: the case of mobile communication.* Research Policy, 2001. **30**: p. 589-610.

[25] WILLIAMSON, O.E., *Strategy Research: Governance and Competence Perspectives.* Strategic Management Journal, 1999. **20**: p. 1087-1108.

[26] COASE, R.H., *The firm, the market and the law.* 1988, Chicago: The University of Chicago Press.

[27] LEENDERTSE, M. *Policy & Performance of the Market for Digital Educational Content.* in *ICA.* 2005. New York City, USA.

# The Open Document Format and its Impact on Accessibility for Persons with a Reading Impairment

*Jan Engelen; Christophe Strobbe*

Kath. Univ. Leuven, Research Group on Document Architectures
Kasteelpark Arenberg 10, B-3001 Leuven (Belgium)
e-mail: Jan.Engelen@esat.kuleuven.be; Christophe.Strobbe@esat.kuleuven.be

## Abstract

It has become very common in the current information society to talk about "open" and to use this term as a quality mark. Open standards, open source software, open archives, open formats etc. are all very much promoted. In this contribution, we would like to focus on the file structure of documents such as texts, spreadsheets and presentations, and more specifically on the Open Document Format. ODF is becoming increasingly popular for many reasons, but it also is the first document format for use in office suites that has unique features built in (as of version ODF 1.1) for persons with a reading impairment such as low vision, blindness or dyslexia.

**Keywords:** open document format; accessibility; ODF; OOXML

## 1 What is a Document Format?

When we produce computer documents, be it a text document, a spreadsheet, a presentation etc., the result of our work has to be stored in computer files. Sometimes this is just one file (a Microsoft Office Word document, a slideshow presentation…), sometimes the contents of a document can only be restored from several files (a well known example are HTML webpages where text and images are stored in separate files).

When we look at the availability of documentation for document formats, it is possible to categorize them as "open" or "closed". Over the years, commercial software makers have come up with document formats that were tied to their own products. Examples of proprietary formats for text processing include Microsoft Word's binary format and WordPerfect's WP format. These formats often changed when new versions of the software products were released, and forced users to migrate to a newer version of the product. This upgrade cycle has to do with several things. On the one hand, a software maker may want to incorporate new features into the next version of their product, and this may necessitate changes to the format (commercial argument). On the other hand, users feel forced to buy the newer version of the product because they fear that they may no longer be able to exchange documents with users who have migrated to the newer version. Users may also feel frustrated because the changes to the document format and the document format itself are not well documented, so they cannot judge the impact of the changes.

At the other end of the spectrum are so-called "open" formats. These formats may have been created by a de facto or official standardization organization and through an open process, for example the HyperText Markup Language (HTML) of the World Wide Web Consortium (W3C). The availability of the W3C's Extensible Markup Language (XML), which is not a document format but a generic syntax for document formats and other data, has led to the creation of hundreds of formats, by standardization organizations as well as companies, individuals and online communities. With the advent of XML, and with the availability of many free or open-source XML parsers, creating, reading and implementing document formats came within the reach of many more users. Document formats passed through standardization or were based on standards, or for which documentation was available to everyone (free or for a fee) gained popularity for several reasons, one being that these characteristics appear to guarantee long-term readability and usability.

It would not be correct, however, to equate proprietary with "closed". The specification for Microsoft's Rich Text Format, for example, is available on Microsoft's web site[1]. Corel offers software development kits (SDKs) for the WordPerfect file format[2]. Portable Document Format (PDF) was created by Adobe Systems but its specification is publicly available, and subsets of the format have undergone or are undergoing standardization by the International Organization for Standardization (ISO). Open formats may also change, and may force users to get new versions of software products, just like changes to proprietary formats.

## 2 The Open Document format

The development of the Open Document Format (ODF) was initiated by developers of word processing and other software who wanted to make their product available in the public domain (open source software). In this particular case the major stimulus came from OpenOffice.org, an office suite that can be used freely by everyone. But as we will see later, there are nowadays many more products supporting this format.
The ODF format has the following characteristics:

- several XML files are produced to describe one document; minimally one has four XML files for any document (cf. table 1);
- the content description of these files is well documented and was the result of a public ("open") standardization procedure (more details about this later);
- the different files are usually (but not necessary) bundled together in one single ZIP file. Zipping files is nowadays a de facto standardized procedure for compressing and joining files and folders together. This action is always transparent for the user.

| | |
|---|---|
| **meta.xml** | information about the document (author, time of last save, ...) |
| **styles.xml** | styles that are used in the document |
| **content.xml** | main document content (text, tables, graphical elements) |
| **settings.xml** | document and view settings (such as magnification level and selected printer); these are usually application specific |

**Table 1: The four basic building blocks of an ODF document**

More details can be found in the relevant Wikipedia page[3] or in chapter 17 (Packages) of the complete ODF standard[4].

## 3 Why is this Format Important for Persons with Disabilities?

The ODF format is based on XML technology, which is promoted through the World Wide Web Consortium (W3C), and reuses formats whose accessibility has been verified through W3C's Web Access Initiative. Also when the ODF standard was developed, under the umbrella of the OASIS consortium, accessibility requirements were taken on board.

However, serious accessibility-related problems showed up when the US State of Massachusetts adopted the use of ODF as the only admissible interchange format for official documents in 2005. That decision has provoked a lot of criticism by groups that do not believe in open source solutions but also by organizations of handicapped persons fearing that they would be forced to use software with less accessibility provisions than their current, Microsoft-based, tools. That is why OASIS set up a special ODF accessibility subgroup in 2005.
Anyhow, there remains still a lot of confusion on the topic of accessibility to information. It is much more important for users with a visual or other impairment that the software they are using is accessible and usable then that the resulting document formats are accessible. In practice these files will never be read by human beings but only by machines. Despite this, the file format **is** important because it may or may not contain the data needed for an accessible reproduction.

At this point in time, general computer accessibility to Microsoft Windows-based software is quite good, especially for persons with a visual impairment. They can efficiently use their special hardware and a special computer program, called a *screenreader*, gives them access to the information and the commands on the computer screen. This is not because of Microsoft cared for this but because a whole group of external companies is building screen enlargement and screenreader software to be used on Windows platforms.
The software packages that support the ODF format currently are less accessible than the Microsoft office products although they are catching up rapidly. Promoters of Unix/Linux systems are especially convinced that this is only a matter of time because, for example, the Gnome Unix desktop is at the same time a so-called Recommended Engineering Accessibility framework[5]. These are frameworks that permit intimate interaction between general applications and accessibility software.

# 4 ODF Accessibility Guidelines

**Current status**

The ODF document format has, right from the beginning, been developed with accessibility in mind.
For this process one could rely heavily on the long term experience gathered around web accessibility via the WAI guidelines.
The details can be found in "Accessibility Guidelines for Implementations of Open Document Format v1.1. Draft 19, 14 March 2007" [6] and the major items are:

a) About the ODF format itself:

- Descriptive texts should be used for anything that is not text (graphs, pictures, sound inserts etc.). All necessary tagging is available.
- Tables and especially column and row headers must be marked up as such. This permits screenreaders to speak out table information together with the cell location information.
- A strict scheme of document divisions and corresponding headers should be maintained, using named stylesheets.
- There is a provision for logical description of navigation inside drawing layers.

b) About the software used for ODF production or conversion:

- The software must check the use of the accessibility features and stimulate authors to use them as much as possible.
- When converting a document in another format, all the accessible information must be kept and must remain available for further conversion, e.g. back into the original format.
- Users must be able to have the layout following general rules (e.g. on font size or color schemes) set up at the level of the operating system. Personal layout wishes (e.g. for persons with low vision) must always adhered to (stylesheet priority management).

Most of the above items have been incorporated into ODF 1.1

**Future work**

The ODF accessibility sub-committee has a number of work items planned for the next release of ODF. These are:

1. Background images. Access to any information contained in images used as backgrounds.
2. Navigation. The way in which people with disabilities can navigate round an individual slide in a presentation. Improving access to tabular data as is found in spreadsheets.
3. Multi-modalities. The provision of access in alternate modalities. For example, improving access to charts and graphs.
4. Reviewing ODF support for a wider range of disabilities.
5. More detailed support for spreadsheets. Easier access to header information, cell labels and formulas.

# 5 Accessibility Testing

At the CSUN 2007 conference, Jonathan Whiting and Aaron Anderson (WebAIM) gave a presentation on "Creating Accessible Content in OpenOffice.org"[7]. Another recent evolution is the development of software packages that audit the accessibility features of ODF documents. In 2006 IBM and the U.S. Department of Education organized a contest to produce such testing software. The winning solution ("ODF accessibility validation tool"") was developed by two American students (from Capitol College and Oklahoma University) and a Chinese student from Tsinghua University (Bejing). This was also announced at the well-known Technology & Persons with Disabilities Conference (CSUN 2007). The winning application and several others are given to the open source community via Sourceforge.org[8]. An online Open Document Format (ODF) Accessibility Evaluator is also available on the website of the Illinois Center for Information Technology Accessibility[9].

# 6 Why is ODF Readily Accepted by so Many Authorities and Companies?

One of the major goals of the Open Document Format is to guarantee access to content on very long time scales and this without technical legal barriers. In other words, efficient archiving with guaranteed future retrieval possibilities and the wish to become independent of Microsoft's business strategy are among the main drivers of adoption.

The fact that the early adopters are mainly public authorities has definitely increased the visibility of accessibility aspects in ODF as these authorities nowadays often have the legal obligation to consider the needs of all the citizens.

It is expected that ODF will slowly gain momentum mainly through acceptance by authorities.

The Massachusetts case had shown the weak point: very few software packages that natively use ODF were available in late 2005. And the incident also lead to the creation of the OASIS subgroup on Accessibility.

One of the early adopters is the Belgian government that has decreed that only open formats are acceptable as an exchange format between the Belgian authorities, and this from 2008 onwards. If realized in time, Belgium would be the first country to prohibit the use of closed document formats. As could be expected, Microsoft has reacted strongly.

## 7 How Popular is ODF in Reality?

By the end of 2006 there were eleven word processors, six spreadsheet programs and 5 presentation managers (Powerpoint-like programs) with support for ODF available. Furthermore, three groups are active in the development of conversions from Microsoft Word into the ODF format and vice versa. They are SUN Microsystems, the Open Document Foundation and the public domain Sourceforge.net project, "ODF Add-in for Microsoft Word"[10]. Within the UK Royal National Institute of the Blind, a project has been set up to turn ODF documents into the Daisy format, the new, worldwide accepted standard for talking books and multimedia documents. The ODF format is also used in the online text processing facilities of **docs.google.com**. Documents produced online can be stored in Microsoft Word, Microsoft RTF, and OpenDocument formats. As it is possible to upload and to download the online documents, the **docs.google.com** facility can in fact be used for file format changes too: upload in one format, download in another. PDF can be used as output, not as input.

## 8 Standardization

The ODF format v1.0 became an international ISO standard in 2006. After having been developed within a working group of the OASIS foundation, it was passed through the International Organization for Standardization, ISO, where it became ISO standard ISO/IEC 26300. The proponents of ODF have created several organizations for discussion and exchange of information. Two of them are very well known:

**Figure 1: Logo of the ODF Alliance[11]**

**Figure 2: Logo of Opendocument-xml.org[12]**

Meanwhile Microsoft has launched a counterattack by creating and promoting XML version of its proprietary office formats, and called them Office Open XML format (OOXML)[13]. This format will be used in Microsoft Office 2007.

In the beginning of 2007, this lead to a very controversial issue on the standardization of XML-based open document standards. The Open Document Standard became ISO 26300 (700 pages) through a very formal process typical for ISO work. Microsoft's alternative, OOXML (Office Open XML) was produced in one year by a technical committee[14] chaired by two Microsoft persons, contained many references to specific behavior of Microsoft software that were not documented and counted 6000 pages. It was ratified as ECMA-376 by ECMA International[15], which was consequently described as "a private association that drafts standards on demand"[16].

In late 2006, Microsoft wanted to put the OOXML specification on a fast ratification track within ISO because the document had already been ratified by ECMA. This created a lot of dismay within organizations that had been developing and promoting ODF. People were even asked to lobby with national ISO delegates to cancel the fast-track procedure. Websites listed arguments against the fast-track process, for example "EOOXML objections" on the Grokdoc website[17].

In spite of this, it was recently (April 2, 2007) announced[18] that ISO-Joint Technical Committee 1 has started the voting period for ISO/IEC standard DIS 29500…

## 9 And now?

ODF still seems to attract quite a lot of organizations. It is public domain and vendor independent, it is well defined (and not too complex) and it is accessible. However, what is even more important for reading impaired users is the fact that the software producing open office documents is made accessible. The existence of different types of converter plug-ins has taken away the major objection against the use of ODF as people can, for example, stay with the more traditional Windows-based platforms.

## Acknowledgment

The authors wish to acknowledge the support of Kris Van Hees (PhD student at K.U.Leuven) and Peter Korn (SUN Microsystems) during the compilation of this contribution.

## Notes and References

[1] Rich Text Format (RTF) Specification 1.6: http://msdn2.microsoft.com/en-us/library/aa140277(office.10).aspx.
Word 2003: Rich Text Format (RTF) Specification:
http://www.microsoft.com/downloads/details.aspx?familyid=AC57DE32-17F0-4B46-9E4E-467EF9BC5540&displaylang=en.

[2] http://apps.corel.com/partners_developers/csp/wordperfect_fileformatsdk.htm

[3] http://en.wikipedia.org/wiki/OpenDocument

[4] ODF standards:
The ISO version:
http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=43485&scopelist=PROGRAMME
The freely available OASIS version can be downloaded from:
http://www.oasis-open.org/committees/download.php/19274/OpenDocument-v1.0ed2-cs1.pdf
(OpenDocument 1.0, Second Edition ; 722 pages) or
http://docs.oasis-open.org/office/v1.1/OS/OpenDocument-v1.1.pdf (OpenDocument 1.1).

[5] Other examples are the Java Accessibility API, the Apple Accessibility API (with limited possibilities) and WAI ARIA for web based applications.
Info from Korn & Schwerdtfeger: *Accessiblity guidelines for ODF* (cf. ref. #7 below)

[6] KORN, P.; SCHWERDTFEGER, R. eds. *Accessibility Guidelines for Implementations of Open Document Format v1.1*. Draft 19, 14 March 2007. http://www.oasis-open.org/committees/download.php/22957/ODF_Accessibility_Guidelines_19_14Mar2007.odt
(only available in Open Document Format).

[7] http://webaim.org/presentations/2007/CSUN/ooo.htm

[8] Meanwhile there are already a number of ODF validator projects on Sourceforge.net. On April 8, 2007 we found: ODF_Accessibility_Tester, ValidODF, Accessibi Add-on component of OpenOffice, ODF Accessibility Validator, Check the accessibility of ODF file, **ODF accessibility validation tool**, odfav, av4odf, Validator for ODF Accessibility, SalihiODF and ODFable.
More details by filling in their name in the search project page of www.sourceforge.net.

[9] http://odf.cita.uiuc.edu/

[10] http://sourceforge.net/project/showfiles.php?group_id=169337

[11] http://www.odfalliance.org/

[12] http://opendocument.xml.org

[13] http://en.wikipedia.org/wiki/Office_Open_XML

[14] ECMA International TC45: http://www.ecma-international.org/memento/TC45.htm

[15] Formerly known as European Computer Manufacturers Association; http://www.ecma-international.org/.

[16] http://press.ffii.org/Press_releases/FFII_opposes_Fasttrack_adoption_of_Microsoft_OOXML_format_as_ISO_standard

[17] http://www.grokdoc.net/index.php/EOOXML_objections

[18] http://www.ecma-international.org/news/PressReleases/PR_TC45_April2007.htm

# Multimedia Modular Training Packages by EUAIN

*David Crombie[1]; George Ioannidis[2]; Neil McKenzie[1]*

[1] Research and Development Department, Dedicon
Molenpad 2, Amsterdam, The Netherlands
e-mail: dcrombie@dedicon.nl; nmckenzie@dedicon.nl
[2] Image Processing Department, TZI (Technologie Zentrum Informatik), University of Bremen
Postfach 33 04 40, D-28334 Bremen, Germany

## Abstract

The European Accessible Information Network (EUAIN) was established to support the move to incorporate accessibility within mainstream content processing environments. EUAIN has brought together a considerable base of knowledge that has now been structured into a series of training modules and curricula which are intended to meet the real needs at this point in time.In this paper we outline how the EUAIN training and learning framework is primarily intended to provide support for everyone who is directly involved in digital content creation and document distribution channels. This target audience requires general courses and training materials as well as domain-specific materials. These general training materials include information about digital document standards and formats, accessibility guidelines and different kinds of publishers and distribution channels. Also important is knowledge about accessibility and alternative forms of presentation that fulfil special requirements for print impaired people. The curricula are illustrated by good practices of accessible content publishing and good examples of accessible digital documents. The specific training materials are addressed to different branches of publishing (books, newspapers, magazines, etc.) and content creators (multimedia content designers, web designers, authors of e-learning content). A significant part of the materials are curricula that demonstrate tools and techniques for accessible content processing. Additionally, the training materials are in modular form to allow them to be adopted within courses and programs to meet the requirements of particular groups. These modular materials are also extensible and scalable, and it is our intention that many new curricula will be developed using this ever-growing resource base. Indeed, the newly-established PRO-ACCESS project is disseminating this information across the publishing industries.

**Keywords:** visually impaired; e-learning; EUAIN; digital publishing

# 1 Introduction

Structured information is the first step in the accessible information process. A document whose internal structure can be defined and its elements isolated and classified, without losing sight of the overall structure of the information, is a document that can be navigated.

Most adaptive technology allows the user to access a document, and to read it following the "outer" structure of the original. But if the same information also has an "inner" structure that allows the adaptive device to distinguish between a phrase and a measure, between a paragraph and a sentence, highlighting particular annotations, then the level of accessibility (and therefore usability) of the whole document will be greatly enhanced, allowing the user to move through it in the same way as those without impairments do when looking at a printed document, and following the same integral logic.

In an ideal world, all documents made available in electronic formats should contain this internal structure that benefits everyone. Highly-structured documents are becoming more and more popular due to reasons that very seldom pertain to making them accessible to people with disabilities. The move to XML related formats and associated standards for metadata has provided an impetus for far greater document structuring than before. Whatever the reasons behind those decisions are, the use of highly-structured information is of great benefit to anybody accessing them for any purpose.

In recent years, the market for accessibility and assistive technologies has started to gain recognition. It is clear that the integration of accessibility notions into mainstream technologies would provide previously unavailable opportunities in the provision of accessible multimedia information systems. It would open up modern information services and provide them to all types and levels of users, in both the software and the hardware domain. Additionally, new consumption and production devices and environments can be addressed from such platforms and this would provide very useful information provision opportunities indeed, such as information on mobile devices with additional speech assistance.

It is equally clear that we remain at the very beginning of the move to incorporate accessibility within mainstream content processing environments. The EUAIN consortium has brought together a considerable base of knowledge that has been structured into a series of training modules and curricula which we believe meet the actual needs at this point in time. These materials are also extensible and scalable, and it is our hope that many new curricula will be developed using this ever-growing resource base.

## 2 The EUAIN Project

The EUAIN project [2] is now nearing completion, and as such much interesting information has been brought together concerning the provision of published information for visually impaired end users. In order that the information brought together by the consortium can have a maximum effect on stakeholder communities in Accessible Information Processing the EUAIN network has created a comprehensive set of instructional training materials. These flexible materials can be used in different environments and work is now underway to translate them into multimedia materials. This paper is a presentation of these developed materials.

## 3 Training Materials

The training and learning framework was primarily constructed with the intention to provide support for everyone who directly effects digital content creation and decides about document distribution channels. This group requires general courses and training materials as well as domain-specific training materials.

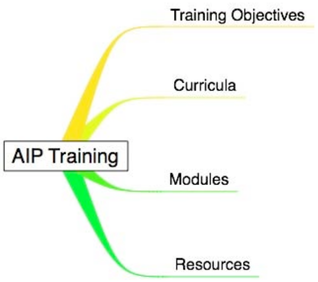


**Figure 1: Accessible Information Processing(AIP) Training**

The general training materials include information about digital document standards and formats, accessibility guidelines and different kinds of publishers and distribution channels. Also important is knowledge about accessibility and alternative forms of presentation that fulfill special requirements for print impaired people. The curricula are illustrated by good practices of accessible content publishing and good examples of accessible digital documents.

The specific training materials are addressed to different branches of publishing (books, newspapers, magazines, etc.) and content creators (multimedia content designers, web designers, authors of e-learning content). A significant part of the materials are curricula that demonstrate tools and techniques for accessible content processing. Additionally, the training materials are in modular form to allow them to be adopted within courses and programs to meet the requirements of particular groups.

In general, there are three themes. The first is related to different types of digital documents and their accessibility issues for print impaired people. The subject of the second theme is to discuss and demonstrate workflows for authoring tools and techniques that allow people to create documents accessible for all. The last theme addresses the processes that must be considered regarding content distribution and digital rights management.

The EUAIN training materials consist of:

- Practical examples of good practice;
- Illustrated explanations of good process management for accessible information production;

- Detailed explanations of approaches, technologies and tools;
- Detailed explanations and examples of benefits and weaknesses of different formats;
- Step by step, modular instructions for producing accessible information in different formats.

Furthermore, the educational process and especially the course materials are themselves a good example of accessible content creation.

After detailed consideration and advice from industry, there is also a requirement that the training materials should operate on several levels. These levels will become increasingly detailed and complex. In this way, different people can choose the level of detail that is required for their situation, or their position in the decision-making chain. In essence, the three levels are as follows:

- **Level 1: Descriptive** - Should teach actors to think about the issues and finding the solutions for their situation. There will only be simple explanations, and not detailed or over-technical information.

- **Level 2 : Decision making** - Should teach actors how to make the right decisions to implement accessible information processing. There will only be relatively simple explanations, and not detailed or over-technical information . These descriptions will link directly to level 3 detailed information.

- **Level 3 Training packages** - At this level, the detailed and more technical information is provided. This level essentially provides the answers to detailed questions and applications.

To this end we have constructed our training materials in such a way that people can choose exactly the most appropriate training packages for their local environments.

## 3.1 The Curricula

In order to target the EUAIN modular training packages at the correct market segments, it is important to understand the various targets of the curricula presented. The most relevant modules can then be presented to these audiences. As a starting point, the WAI Resources on Developing Web Accessibility Training and Presentations[1, 7, 8] have been used and adapted to be more specific to Accessible Information Processing:

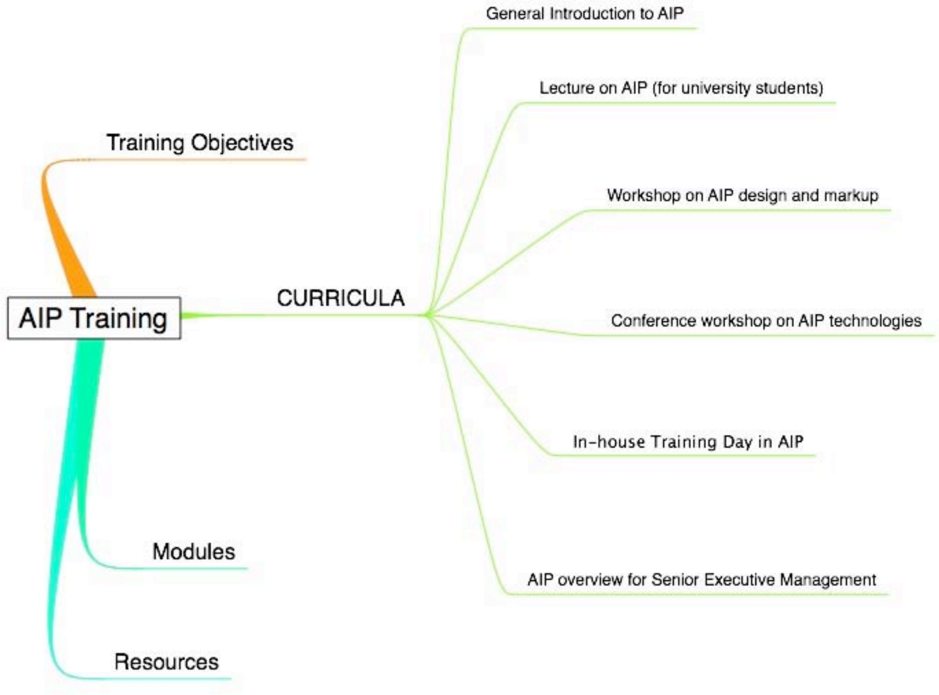

General Introduction to AIP
Lecture on AIP (for university students)
Training Objectives
Workshop on AIP design and markup
AIP Training
CURRICULA
Conference workshop on AIP technologies
In-house Training Day in AIP
Modules
AIP overview for Senior Executive Management
Resources

**Figure 2: An expansion of the Curricula section of the AIP training**

The example curricula currently available are:

- **General introduction to Accessible Information Processing:** A general introduction to a heterogeneous audience provides the background information required to understand why Accessible Information Processing is important in whatever environment the recipients of the training are involved in. The curricula is designed to take 40 minutes to present in order that it can be easily incorporated in other curricula and training sessions.

- **Accessible Information Processing Lecture:** a single two-hour lecture/presentation on AIP as part of a full semester's introductory course on general web design skills.

- **Workshop on AIP design and markup Context:** Hands-on workshop on Accessible Content design and mark up, for a class(~10-20 people), of content creators. The workshop assumes some knowledge of the Business case and the Market for Accessibility, and is taught with computers for learning assistance. The class has to be taken by someone who has a reasonable experience in Accessible Information Processing as the workshop requires a lot of interaction with the subject matter.

- **Conference Workshop on AIP technologies:** This curriculum specifies a ninety minute workshop which can be given at a conference or trade event. It is aimed at IT workers who have some knowledge of software design and development. It is assumed that the audience is familiar with the need for accessible information processing.

- **In-House Training Day on Accessible Information Processing:** In-house training at a publisher, content creation company, or software development company. The audience is assumed to have some level of knowledge of Accessible Information Processing. The training session requires considerable preparation(Possibly with the help of an organisation contact point) by the facilitator of the training in order that the training is relevant to the organisations specific field, workflows and authoring tools.

- **Accessible Information Processing Overview for Senior Management:** A brief presentation around a conference table during a senior management meeting. The focused delivery of this training aspect requires a familiarity with the material.

These curricula are aimed at different audiences and market segments. The materials for these curricula draw on a body of topic specific modules which have been brought together for use in training. Each sample curriculum highlights several objectives and learning outcomes for the topic and provides a list of resources relevant to those learning objectives. An estimated time frame for each curriculum is provided. They are designed in such a way that the curricula can be altered and personalised for more specific needs and situations.

## 3.2 Modular Training Packages

The Modules for the EUAIN training materials have been modeled on the structure of the WAI Resources on Developing Web Accessibility training and presentations. , The structure is as follows:

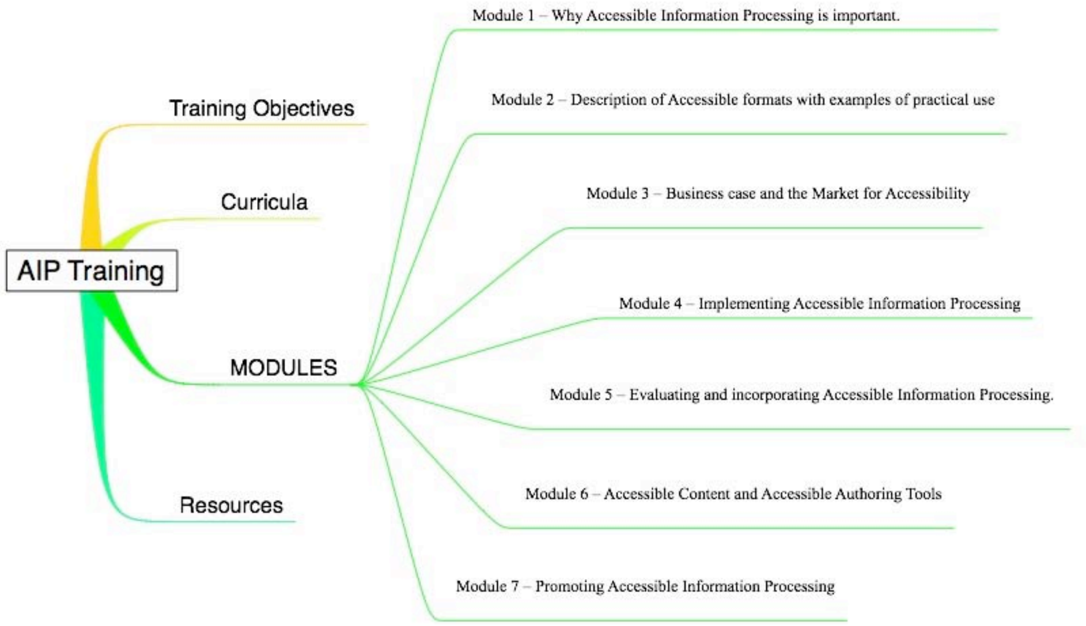

AIP Training
Training Objectives
Curricula
MODULES
Resources
Module 1 – Why Accessible Information Processing is important.
Module 2 – Description of Accessible formats with examples of practical use
Module 3 – Business case and the Market for Accessibility
Module 4 – Implementing Accessible Information Processing
Module 5 – Evaluating and incorporating Accessible Information Processing.
Module 6 – Accessible Content and Accessible Authoring Tools
Module 7 – Promoting Accessible Information Processing

**Figure 3: The Modules section of the AIP training packages.**

These modules consist of materials which can be used in the curricula described above. Each module focuses on a particular aspect of Accessible Information Processing and provides reusable materials such as hand-outs, presentations, software and other useful materials:

- **Module 1 – Why Accessible Information Processing is important.** This module provides an overview of Accessible Information Processing. This includes an overview of the topics in the other modules and should be seen as an introduction module to all the EUAIN training modules. The resources presented in this module are relevant to all the other modules and relate directly to all the curricula.

- **Module 2 – Description of Accessible formats with examples of practical use.** This module is aimed at providing a solid understanding of the formats relevant to Accessible Information Processing and how they are used in processes and workflows. As such it presents descriptions of how these formats used in both mainstream environments and environments that are more focused on print impaired users. An understanding of these formats is essential before conversion processes can be built out of the formats such that they can be incorporated into workflows and processes relevant to the stakeholders which these resources and training materials are presented to.

- **Module 3 – Business case and the Market for Accessibility.** It is important to understand not only the technical perspectives for formats and conversations required for Accessible Information Processing but also the business angle and the market relevance for incorporating Accessible Information Processing within existing industrial environments. This module targets decision makers and executives within stakeholding communities who have to assess the cost and benefits of implementing Accessible Information Processing.

- **Module 4 – Implementing Accessible Information Processing.** Accessible Information processing is a series of processes., but very few of these processes stand alone, in most cases, the accessibility component will be one in a chain of processes with the input and outputting feeding from and to other processes. This module describes how Accessible Information processing ties in with these other processes and workflows already in place in mainstream environments.

- **Module 5 – Evaluating and incorporating Accessible Information Processing.** In order to successfully incorporate Accessible Information Processing within existing workflows, it is important to first evaluate the workflows for accessibility. This relates to the formats, authoring tools and standards already in place in the processes within these workflows. This module focuses on evaluating this accessibility and how these evaluations can help point top answers on the best way to implement further accessibility.

- **Module 6 – Accessible Content and Accessible Authoring Tools.** In order to implement accessible information processing within organisations, clear understanding is required of how to create, modify and process content using both the tools available within mainstream organisations and also the accessibility conversion tools and assistive technologies. This module provides resources such to make this possible.

- **Module 7 – Promoting Accessible Information Processing.** This module ties the previous 6 modules together in order that participants of EUAIN learning packages can reuse their knowledge and understanding of Accessible Information Processing within heir organisation and further promote accessibility. This module ensures that accessibility is reverberated through the organisation and can be promoted from the top down.

Each module is intended for use as information which feeds into specific curricula but they have specific objectives, resources and learning outcomes such that they can be used as a stand alone information package.

## 3.3 Resources and Additional Materials

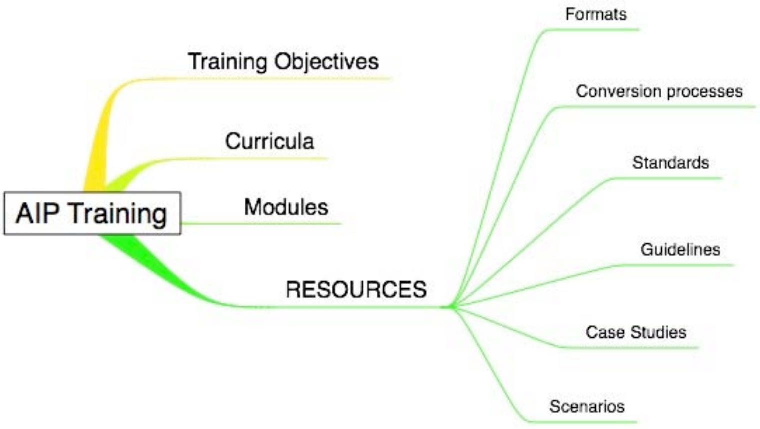


**Figure 4: Overview of AIP training resources**

**Formats**
In order to look at the various processes involved in Accessible Information Processing, it is essential to build up a finite list of the formats which are commonly used during these processes and interactions within supply chains. After careful consideration of several specialist organisations, publishers and users, we came up with the following list of formats.

- Printed paper
- Printed Braille
- Audio(Wav)
- ASCII Text
- HTML
- XML
- Multimedia Packages

It was felt that these descriptions covered all areas. There is a specific focus on formats for the print impaired, so formats such as bitmap or JPEG are considered to be components of more complex multimedia packages, as they are rarely dealt with without some sort of surrounding information or multimedia package.

**Conversion Processes**
Given that we have a finite list of formats used for accessible information processing, a conversion from every format to every other format provides us with a list of accessible information conversions. This provides us with a list of 42 conversion processes.

Each conversion process is dissected to provide:

- a description of the conversion from a accessible information processing perspective
- examples of the conversions use in real life case studies (see below)
- examples of the conversions use in hypothetical scenarios (see below)
- related guidelines and best practices
- A flow chart description of the process

For example:

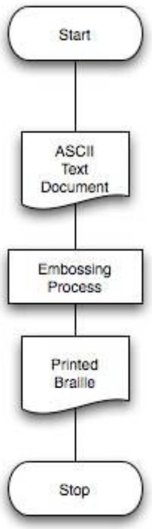


**Figure 5: Flow chart of ASCII to Braille Process**

**Standards**
As part of the work of EUAIN a deliverable entitled “Standards for Accessible Information Processing” was created. The deliverable is public and available on the EUAIN website. This information is used as a resource in many of the modules and curricula.

**Guidelines**
EUAIN is not the first project to tackle the issues of Accessible Information Processing and there are several sets of Guidelines and Best Practice already in existence. However, until now these have not been brought together in a systematic manner. EUAIN has collated this information in order to focus stakeholders on specific information based on their specific requirements. This information is available on the EUAIN website, the EUAIN wiki and it is also fed into the resources for the training materials.

**Case Studies and Scenarios**
Based on real-life examples of accessible content processing, we have prepared a number of Case Studies and scenarios to illustrate different aspects of accessible content processing. These Case Studies are drawn from different publishing sectors and address a variety of different issues. Each Case Study provides an in-depth examination of key factors and provides practical explanations of how the various processing stages were addressed to achieve accessible content. The case studies are constructed out of the same conversions which were described above:

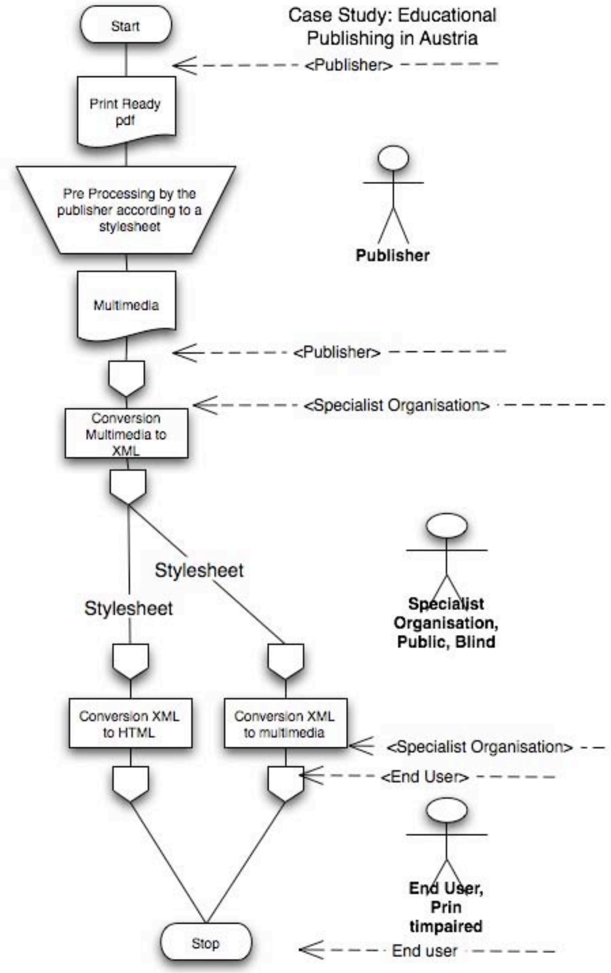

Case Study: Educational Publishing in Austria
Start
<Publisher>
Print Ready pdf
Pre Processing by the publisher according to a stylesheet
Publisher
Multimedia
<Publisher>
<Specialist Organisation>
Conversion Multimedia to XML
Stylesheet
Stylesheet
Specialist Organisation, Public, Blind
Conversion XML to HTML
Conversion XML to multimedia
<Specialist Organisation>
<End User>
End User, Prin timpaired
Stop
End user

**Figure 6: Example of a case study flow description**

This case study is then dissected in terms of:

- Actors involved in the information processing chain
- Conversions and processes used
- Standards used
- Guidelines used.

This body of information is also available on the EUAIN wiki[ref].

## 4 Pro-Access

Following on from EUAIN, the consortium intend to explore these themes further by taking part in several more practical implementations which tackle that which EUAIN raised. One such endeavour is the PRO-ACCESS[4] project which started recently.

The project will provide practical tools for publishers and content providers to address the targeted audience of primary and secondary students with specially formatted and accessible course materials on a timely basis with total respect of copyright.

This main objective of the project will be achieved following these steps:

- evaluate the actual situation in the involved countries, analysing on one side the needs of the disabled people and on the other the problems and the concerns arose from these request in the publishing sector, involving key schoolbook publishers and printed disabled people representatives in the process;

- define the production process needed to create accessible documents, starting from the achieved results of the EUAIN project;
- promote the results as wide as possible in the publishing sector;
- analysis of the content value chain in the education sector (authors, publishers, intermediaries, schools, students) to define a set of shared rules to managing rights.

Expected results:

- a set of ISO 9001 compliant Certification guidelines for publishers to create an accessible school materials in a standard way. The publishers who will follow these guideline will be appointed with a specific process certification;
- a standard license for publishers to be used to manage the relations between the publishers and the students, or the schools asking for special formatted materials;
- a set of materials devoted to create awareness in the print disable people environment and in the school environment, teachers and educational authorities in particular;
- a standard module for blended training courses for publishers and content providers in order provide them with all the information needed both on the technical and legal solutions defined in the project itself.

As these results come to the fore, they will be disseminated through similar channels to EUAIN.

## 5 Conclusion

The EUAIN network has provided some practical training solutions and it is now important to create broader awareness on these topics in the content producers market (i.e. publishers, Learning Object producers, digital content and software developers) and promote the adoption of collaborative and practical solutions to allow them quickly to make available these accessible materials. New projects such as PRO-ACCESS can help to achieve these goals.

The coherent and sustainable provision of accessible information cannot be tackled in isolation by individual actors in the information provision chain. While examples of good practice are emerging in the production sphere and in new collaborative distribution models, a European-wide approach offers far greater potential. In particular, a collaborative approach involving content producers and users' associations allows us to approach key aspects like rights clearance, definition of standard formats for exchanging content files, and finally actual increase of accessibility. As noted in the recent report produced for WIPO:

> *"This [EUAIN] is perhaps an example of a way forward more generally and work of this nature should perhaps be promoted more widely by governments and international agencies. It seems to be in everyone's interests that a desire to build in access from the start is both encouraged and facilitated by ensuring that what this requires in practice is widely understood and adopted."*[6]

By focusing closely key issues in this area (rights management, production processes, content value chain, and standard information exchange), we can make an important and lasting contribution to the Accessibility For All initiative and help to provide the educational building blocks needed to help make consumer needs more explicit to the designers of products and services for print impaired people.

## References

[1] http://www.w3.org/WAI/training/

[2] http://www.euain.org

[3] http://wiki.euain.org

[4] http://www.euain.org/proaccess

[5] http://www.ormee.net

[6] SULLIVAN, J., (2007) Study on Copyright Limitations and Exceptions for the Visually Impaired, SCCR15/7, WIPO, Geneva

[7] http://www.w3.org/WAI/

[8] http://www.w3.org/TR/WAI-WEBCONTENT/

# File-Formats for Preservation: Evaluating the Long-Term Stability of File-Formats

*Carl Rauch*[1]*; Harald Krottmaier*[2]*; Klaus Tochtermann*[3]

[1] Styria Media AG, Schönaugasse 64, 8010 Graz, Austria
e-mail: carl.rauch@styria.com
[2] Institute for Computer Graphics and Knowledge Visualization, Graz University of Technology
Infeldgasse 16c, 8010 Graz, Austria
e-mail: h.krottmaier@cgv.tugraz.at
[3] Knowledge Management Institute, Graz University of Technology, and Know-Center Graz
Infeldgasse 21/II, 8010 Graz, Austria
e-mail: klaus.tochtermann@tugraz.at

## Abstract

While some file-formats become unreadable after short periods, others remain interpretable over a long-term. Among the over 1.000 file-formats, some are better and some are less suited for long-term preservation. A standardized process for evaluating the stability of a file-format is described in this paper and its practical use is shown with file-formats for 3D-objects. Recommendations to users of 3D-applications are given in the last section of this article. Some of the results are used in PROBADO, a sophisticated search engine for non-traditional objects (such as 3D-documents, music etc.).

**Keywords:** digital preservation; evaluation metric; file-formats

## 1 Introduction

In file-format registries like PRONOM, filext or MyFileFormat, over 1.000 file formats are registered. Even when removing all depreciated formats and even when setting the focus on one type of digital records only, e.g. 3D-objects, the number of available file formats is big (in this case among others dxf/dwg, iges, 3ds/max, 3dm, obj ). While some file-formats depreciate over time, other file-formats are evolving. Formats, which were frequently used 10 years ago, are unreadable now as will many today's formats in ten years. But even slight modifications in the representation of digital objects can have major influences on their significance. An example would be a computer game with a slightly higher processing speed - it would become many times more difficult to play.

When a digital object needs to be available over a long-time period, users face the question, which file-format to choose for long-term preservation. Based on the concept of Utility Analysis [12] and on work done by Rauber, Strodl and Rauch [11], an evaluation process is described in this paper for analyzing and ranking file formats in terms of long-term reliability.

An evaluation of file-formats for 3D-objects is used for showing the process in practice. The remainder of this paper is organized as follows: Section 2 provides an overview over related work. In Section 3 the workflow and parameters for evaluating file-formats is described. In Section 4 the criteria for evaluating file-formats are shown in detail. A practical implementation for 3D-objects shows the feasibility of the described approach in Section 5.

## 2 Related Work

The work described in this paper is based on three research areas. The first basis is the area of digital preservation, where methods and workflows for comparing various preservation alternatives are developed and implemented. The second area are already existing initiatives to examine a file format's preservation risk. The third are file-format registries.

In the research area of digital preservation, several processes for evaluating preservation strategies were presented in the last couple of years. Among them are the test-bed workflow of the Dutch Preservation Test-bed [9] and the Utility Analysis workflow of the Vienna University of Technology [8]. As part of the DELOS

Network on Excellence project, these two workflows were combined to the DELOS digital preservation Test-bed's workflow [11], which is shown in Figure 1.

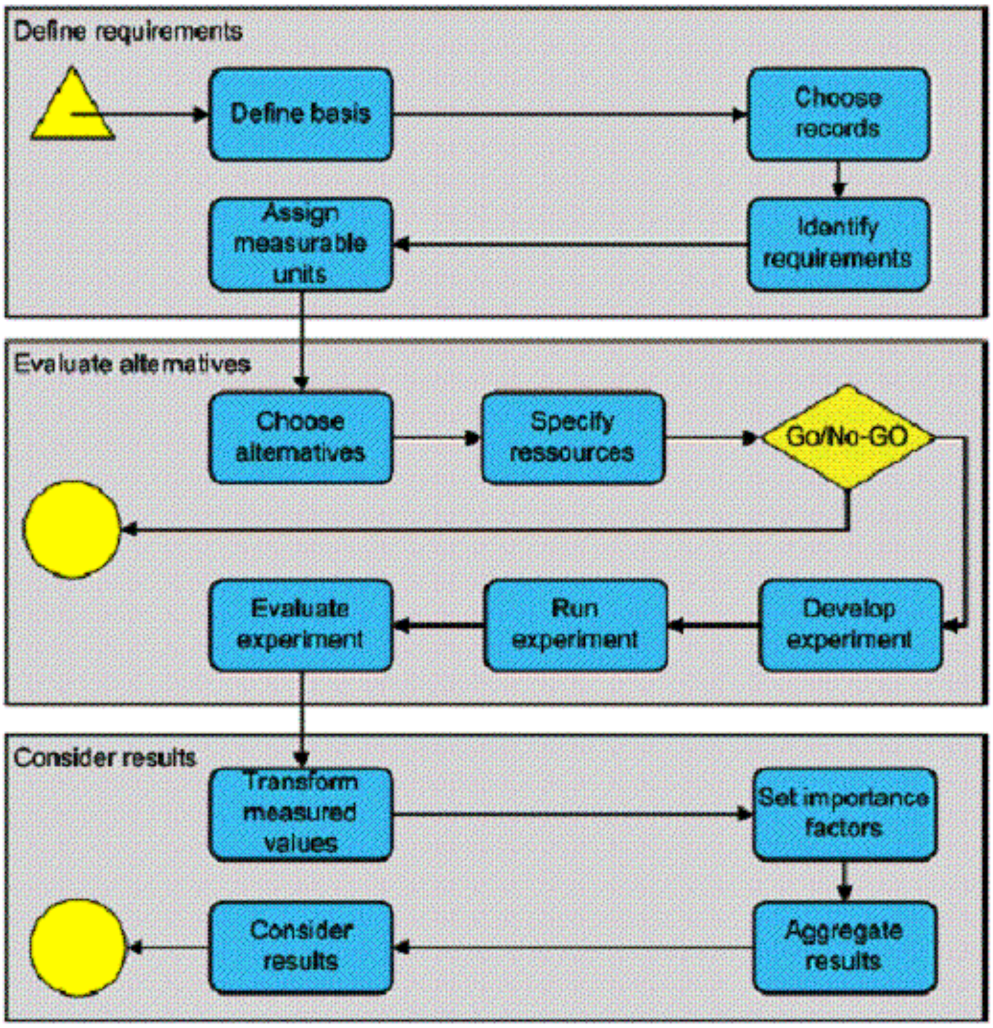


**Figure 1: Overview of DELOS Digital Preservation Test-bed's workflow [11]**

The DELOS workflow consists of three main parts: At the beginning, the requirements of an institution for a digital preservation strategy are defined. Here the record set, which is to be preserved, is selected, a list of criteria for evaluating the strategies defined and measurable units are assigned to each criterion. In the second part the evaluation takes place. After defining alternatives and resources to be tested, an experiment is developed and different preservation strategies are applied to the chosen objects. In the third part finally the results are examined by aggregating the performance of each alternative for the different criteria. This workflow forms the basis for the evaluation of the file-formats.

Another research is methods for evaluating the preservation risk. During the last couple of years two initiatives were started to evaluate the preservation risk of a file-format. First the INFORM system of the Online Computer Library Center [10]: There the durability of file-formats in a specific environment is evaluated, considering not only the reliability of a file-format itself, but also of the opening software, the hardware, of associated organizations, the digital archive and migration and derivative-based preservation plans. The main disadvantages of this system are, that for the assignment of a risk-factor to one of the six risk-areas, a high level of expertise is required for each individual environment. Thus the process needs highly qualified officers and cannot be standardized easily. The here proposed workflow suggests an alternative solution to these drawbacks.

A second initiative is the 'Virtual Remote Control' project of the Cornell University [4]. VRC focuses on the preservation of web pages. If the VRC-web-crawler detects a page with dysfunctional hyperlinks, longer downtimes or older server-software, the VRC-administrator is notified about the preservation risk of the web page. VRC provides some interesting insights on evaluating the preservation risk, however it is only focusing on web pages and the file-format itself plays a minor role.

The last research area on which this paper is based is file-format repositories. Several repositories exist, where different aspects of file-formats are stored. The best-known example is the PRONOM-database of the UK National Archives. In this archive the following information are stored (among others) about a file-format [7]:

- Name, Version and other Names
- Identifiers
- Family, Classification and Orientation
- Byte Order and Related File-Formats
- Release date and support end date

A second file-format registry is FILExt. In FILExt [3] the external and internal signatures of a file-format, the software programs able to interpret the format, the MIME types, the main producing company, the file-formats name and a description is given for each file-format.

Neither of the registries contains a specific measure on the reliability of a file-format. For both the information given needs to be interpreted by a file-format expert to evaluate the appropriateness of a format for digital preservation.

# 3 The File-Format Evaluation Process

Based on the workflow shown in Figure 1 a process for evaluating the reliability of file-formats is presented in this section. Due to the smaller scope - the DELOS workflow is designed for comparing whole preservation strategies including appearance, process characteristics and costs - the here shown process consists of less steps than the DELOS workflow. Most of these steps are standardized for all file-formats.

1. Review Requirements: The requirements for a reliable file format are structured in a criteria-tree. The criterion focuses on two areas: on technical characteristics and on the integration of the format within the marketplace. The criteria tree described in detail in Section 4 is the same for all file-formats in order to allow comparability;
2. Assign measurable categories: The second step is to assign measurable categories to each criterion. A metric is defined describing, how to convert the measured numbers into a zero-to-five scale (e.g. number of users between 10.000 and 100.000 is equal to '3' for the market penetration criterion). These conversion tables are described in more detail in Section 4 and are standardized for every evaluation run;
3. Choose alternatives: In this step file-formats are chosen, which are evaluated during a session of the workflow. In the here presented work, six file-formats for 3D-objects are evaluated as a proof-of-concept;
4. Evaluate file formats and transform values: Based on the seven sub-criteria of the criteria tree and on the measurable categories the file-formats are evaluated and a value between zero and five (five is the best) is assigned to every criterion of each file-format. These evaluation results do typically not change over time and are stored as a basis for the final aggregation;
5. Set importance factors: After the evaluation, each criterion is ranked with a percentage value according to the user's priorities; the sum of all percentages has to be 100 %. Each user can determine the importance of certain criteria for individual circumstances with values from 0 % (is not interesting at all) to 100 % (is the only relevant criterion);
6. Aggregate results: A final value per file-format is found by multiplying the value per criterion with its weight and summing these values up. The higher the value, the better a file-format is suited for long-term preservation. By aggregating the final values of several file-formats or by taking earlier evaluations as a reference, a clear ranking can be created. A measure suggested for file-formats is the preservation risk, which is calculated by dividing the final value per file-format by the maximum value possible (in the here described metric, the maximum possible number is five). This fulfillment percentage-value has then to be subtracted from one. The higher the preservation risk, the lower the probability of being able to interpret the file-format after a couple of years.

From the above listed steps, the requirement review and the assignment of measurable categories is standardized for every evaluation run. When evaluating file-formats, a user has to do the steps three to five for each run; the aggregation of results follows again a standardized scheme.

# 4 The File-Format Evaluation Tree

In this section the tree of requirements and the assignment of measurable categories are described. In order to compare and evaluate file-formats in terms of long-term reliability, criteria were defined and structured in a criteria-tree. The tree is based on a discussion process with the Department for Software Technology, Vienna University of Technology, the Austrian National Library, the Austrian Phonogrammarchiv and the Dutch Nationaal Archief; it is structuring all criteria, which are seen as important to measure the long-term reliability of a file-format. The tree is shown in Figure 2. The tree consists of two branches, the technical and the market characteristics.

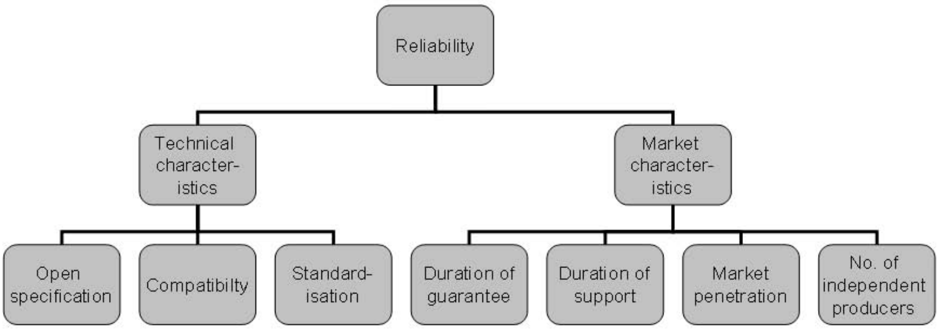


**Figure 2: Criteria-tree for evaluating the long-term reliability of file-formats**

The technical characteristics focus on the specification of a file-format. It consists of the following three sub-criteria:

- Open Specification: Is the specification of the file-format publicly available?
- Compatibility: Is the file-format supported and maintained by one or several software companies?
- Standardization: Is the file-format standardized by a recognized standardization agency, such as DIN or ISO?

The market characteristics focus on the acceptance and position of the file-format in the market. It is divided into the following sub-questions:

- Duration of guarantee: How long does the main producing software company guarantee to repair bugs in the interpreting software?
- Duration of support: How long does the main producing software company supports the interpreting of the file-format with its software?
- Market penetration: How many users are working with the file-format at the current time?
- Number of independent producers: How many software products exist, which are able to interpret the file-format?

In order to transform measurable units into values from zero to five the following transformation tables are suggested. The intervals are chosen in a way, which should bring a maximum distinction between typical software formats. By targeting the range of values, which typical software formats have, the differences between formats can be shown explicitly:

- Open Specification: Yes = 5, Partly available = 3, No = 0
- Compatibility: Number of software systems compatible with the format: 1 system = 1, 2 systems = 2, 3 systems = 3, 4 systems = 4, > 4 systems = 5
- Standardization: Yes = 5, Partly standardized = 3, No = 0
- Duration of guarantee: 0 years = 0 (no guarantee), > 0 years and <= 1 year = 1, > 1 year and <= 3 years = 2, > 3 years and <= 5 years = 3, > 5 years and <= 10 years = 4, > 10 years = 5
- Duration of support: 0 years = 0 (no support), > 0 years and <= 1 year = 1, > 1 year and <= 3 years = 2, > 3 years and <= 5 years = 3, > 5 years and <= 10 years = 4, > 10 years = 5
- Market penetration: < 100 users = 0, > 100 users and <= 10.000 users = 1, > 10.000 users and <= 100.000 users = 2, > 100.000 users and <= 1.000.000 users = 3, > 1.000.000 users and <= 10.000.000 users = 4, > 10.000.000 users = 5
- Number of independent producers (that support the software): 0 producer = 0, 1 producer = 1, 2 producers = 2, 3 producers = 3, 4 producers = 4, > 4 producers = 5

# 5 Evaluating Digital Objects for 3D-Data

As a proof-of-concept, file-formats for 3D-objects were evaluated and ranked according to their preservation risk. The steps three to six of the evaluation process are described in detail in this section.

The choice of alternatives is the first step, which needs to be done before an evaluation run. The following file-formats were selected, based on inputs from the PROBADO project [6]: Drawing Exchange Format DXF/DWG, Initial Graphics Exchange Specification IGES, 3D Studio 3DS/MAX, 3D Model 3DM and Object OBJ .

Based on publicly available sources, such as Internet queries and producer information, the file-formats were evaluated. Please note that the proof-of-concept is primarily done to show the functionality of the evaluation process and can not be seen as a final judgement on the performance of every file-format.

| Criterion | DXF/DWG | IGES | 3DS/MAX | 3DM | OBJ |
|---|---|---|---|---|---|
| Open Specification | 5 | 5 | 3 | 0 | 5 |
| Compatibility | 5 | 5 | 5 | 5 | 5 |
| Standardization | 0 | 5 | 0 | 0 | 0 |
| Duration of guarantee | 0 | 0 | 0 | 0 | 0 |
| Duration of support | 0 | 0 | 0 | 0 | 0 |
| Market penetration | 3 | 1 | 5 | 1 | 3 |
| No. of independent producers | 1 | 5 | 1 | 1 | 5 |

**Table 1: Evaluation results per file-format**

Some of the results are exemplarily described in more detail to clarify the evaluation process:

- Duration of guarantee / duration of support: No information was publicly available for these two criteria, so these criteria are always evaluated with zero (since all file-formats have the same value here, the ranking is not influenced). Data like these are typically given by software companies during sales negotiations.
- Open specification: Open specifications exist for the DXF/DWG [1], IGES [5] and 3DS/MAX [2] file-format. 3DS only gets three points, since the last found specification is from 1997, although 3DS is still under development by Autodesk.
- Compatibility of IGES: At the time of its creation IGES was compatible with most available software products. Meanwhile in PRONOM only one compatible software is listed: Adobe FrameMaker 2002; in a web-search additional software products, such as ModelPress Desktop, CrtlView or 3D Shop ModelScan are named (see http://www.programurl.com/, Date of Download: 09.03.2007). Additionally a conversion tool for Autodesk exists.
- Standardization of IGES: IGES has been standardized by the Department of Defense and the National Institute of Standards and Technology [5].
- Market penetration of 3DS MAX: Wikipedia [13] lists 42 software companies, which use the 3DS MAX format, among them major producer of computer games and animated movies.
- No. of independent producers of OBJ: According to Wikipedia, the OBJ file-format has been adopted by several software vendors and can be imported and exported to a number of software programs.

As can be seen, the above shown evaluations rely on Internet-sources only. We recommend a detailed clarification with software vendors before deciding for one format or another.

| Rank | File-Format | Preservation Risk |
|---|---|---|
| 1 | IGES | 40.00 % |
| 2 | OBJ | 48.57 % |
| 3 | DXF | 60.00 % |
| 4 | 3DS | 60.00 % |
| 5 | 3DM | 80.00 % |

**Table 2: The final evaluation result**

After the evaluation step importance factors are set for each criterion. These factors indicate how the end-user values certain criteria. In the here shown example, all seven criteria get the same weight – 14.29 %. The evaluation results are multiplied with the weight of its criterion and summed up per file-format. By taking the percentage value from the maximum possible value (which is five) and by subtracting it from 100, the preservation risk can be obtained. The final result is shown in Table 2. The differences between the file-formats in terms of preservation risk are significant and IGES is ranked top as a format for long-term preservation.

## 6 Conclusion

In this paper a methodology for evaluating file-formats in terms of reliability for long-term preservation is presented. In the first part the steps of the evaluation process are described in detail. In the second part of the paper a proof-of-concept is done for 3D-file-formats to show the functionality and details of the process in practice.

After evaluating several file-formats, a file-format list can be created, where all selected formats are ranked according to their preservation risk. Such a list could be maintained by a research institution or a library and could be continually updated. By including software companies and the open-source community into the evaluation process, the evaluation results can on the one hand become more precise and on the other hand become a motivation for improving the preservation reliability of file-formats. Additionally such a ranking could be added to existing file-format registries, such as PRONOM or the Global Digital Format Repository.

## Notes and References

[1] AutoCAD2006. *DXF Reference*, July 2005. URL http://www.autodesk.com/, Date of Download: 04.02.2006.

[2] Autodesk Ltd. *3D-Studio File Format*, January 1997. URL http://www.martinreddy.net/gfx/3d/3DS.spec, Date of Download: 04.02.2006.

[3] *FILExt - The File Extension Source*, 2007. URL http://FILExt.com, Date of Download: 31.01.2007.

[4] MCGOVERN, N. Y.; KENNEY, A. R.; ENTLICH, R.; KEHOE, W. R.; BUCKLEY, E. *Virtual Remote Control, Building a preservation risk management toolbox for web resources*. D-Lib Magazine Volume 10, Number 4 (2004).

[5] National Institute of Standards and Technology. Initial *Graphics Exchange Specification (IGES)*, April 1996. FIPS PUB 177-1.

[6] *PROBADO - Prototypischer Betrieb fuer Allgemeine Dokumente*, 2007. URL http://www.probado.de, Date of Download: 05.02.2007.

[7] *PRONOM, the technical registry*. URL http://www.nationalarchives.gov.uk/pronom/default.htm, Date of Download: 07.07.2006.

[8] RAUCH, C.; RAUBER, A. *Preserving digital media: Towards a preservation solution evaluation metric*. In Proceedings of the 7th International Conference on Asian Digital Libraries, Shanghai, ICADL 2004 (December 2004), Springer-Verlag Berlin, Germany, pp. 203–212.

[9] SLATS, J.; VERDEGEM, R. *Practical experiences of the Dutch Digital Preservation Testbed*. VINE, The journal of information and knowledge management systems, Volume 34, Number 2 (2004), 56–65.

[10] STANESCU, A. *Assessing the durability of formats in a digital preservation environment*. D-Lib Magazine 10, 11 (2004). URL http://www.dlib.org, Date of Download: 14.03.2005.

[11] STRODL, S.; RAUBER, A.; RAUCH, C.; HOFMAN, H.; DEBOLE, F.; AMATO, G. *The DELOS testbed for choosing a digital preservation strategy*. In Proceedings of the International Conference on Asian Digital Libraries, ICADL (2006), Springer-Verlag, Berlin, Germany.

[12] WEIRICH, P. Decision Space: *Multidimensional Utility Analysis*. Cambridge University Press, 2001. URL http://www.missouri.edu/weirichp, Date of Download: 03.08.2005.

[13] *WIKIPEDIA, The free Encyclopedia*, 2007. URL http://en.wikipedia.org, Date of Download: 20.02.2007.

# Beyond Publication – A Passage Through Project StORe

*Graham Pryor*

University of Edinburgh
Digital Library Division
George Square, Edinburgh, Scotland
e-mail: graham.pryor@ed.ac.uk

## Abstract

The principal aim of Project StORe is to provide middleware that will enable bi-directional links between source repositories of research data and the output repositories containing research publications derived from these data. This two-way link is intended to improve opportunities for information discovery and the curation of valuable research output. In immediate terms, it is expected to improve citation rates as a consequence of increasing the accessibility of research output. A survey of researchers in seven scientific disciplines was used to identify workflows and norms in the use of source and output repositories, with particular attention being paid to the existence of common attributes across disciplines, the functional enhancements to repositories considered to be desirable and perceived problems in the use of repositories. Cultural issues were also investigated. From the results of the survey, a generic technical specification was designed and a pilot environment created based upon the *UK Data Archive* (source repository) and the London School of Economics' *Research Articles Online* (output repository). A further link to a prototype institutional repository at the University of Essex was used as a control mechanism. The StORe middleware was designed using a Web 2.0 approach similar to existing FOAF (Friend Of A Friend) services such as Flickr and MySpace, but incorporating a federation of institutional, source and output repositories rather than one central area where digital objects are deposited. Researchers can deposit digital material in various formats at their institutional repositories until the data and publications are made available at linked source and output repositories. An enabling central portal provides an OAI-based aggregator service, which harvests the contents of the federation's repositories and provides a simple search facility. Whilst all digital objects are title visible, a key feature of the middleware is the Flickr-like option for regulating access, which gives researchers control over who can see objects they have designated 'non-public'. Using the StORe middleware, it will be possible to traverse the research data environment and its outputs by stepping seamlessly from within an electronic publication directly to the data upon which its findings were based, or linking instantly to all the publications that have resulted from a particular research dataset. It has already been endorsed by participating researchers as having the potential for integrating multiple data sets from different publications. Following completion of the pilot demonstrator, an independent evaluation undertaken by the National Centre for e-Social Science found it effective and easy to use. It may also be said to have broadened the meaning of the terms *publish* and *publication*.

**Keywords:** interoperability; research publications; institutional repositories; middleware

## 1 Introduction

Project StORe is an initiative funded by the UK's Joint Information Systems Committee within its 2005-7 Digital Repositories Programme.[1] StORe's principal aim is to attach new value to published research through the provision of two-way links between the output repositories that contain research publications and the source repositories of original and processed data from which those publications originated. Hence the project name, which is an acronym of **S**ource **t**o **O**utput **Re**positories. This bi-directional linkage is predicted to increase opportunities both for information discovery and the curation of valuable research data. Specifically, it will provide members of the research community with the means to navigate directly from within an electronic article to the source or synthesised data from which the article was derived; conversely, direct access will also be provided from source data to the publications associated with those data. Researchers will benefit from this linkage through an enhanced capacity to track the use and influence of their published research, as well as to engage in the more comprehensive dissemination of research and scholarship, which it is anticipated will increase the citation rate for research papers linked to their sources. Scientific researchers involved in the development phases of the project have already identified other advantages, such as the ability to conduct a reanalysis of source data as new methods emerge, a feature that should lead to improvements in the integrity of published results, whilst the potential for integrating multiple data sets from different publications has been perceived as promising time saved and more productive research.

On the subject of reanalysis, an incident reported in *Science* late last year[2] underwrites the potential value from being able to take a critical look at a published paper alongside its data. In a September 2006 paper in *Nature*, Swiss researchers cast serious doubts on a protein structure described in a 2001 *Science* paper by Geoffrey Chang's pioneering group at the Scripps Research Institute, San Diego. Upon investigation, Chang found that his homemade data-analysis program had inverted the electron-density map from which he had derived the final protein structure. Consequently, Chang and his colleagues had to retract three *Science* papers and report that two papers in other journals also contained erroneous structures. If his original paper and its data had been published together perhaps this mistake would have been discovered earlier.

Having referred here to the dual *publication* of a paper with its data opens up a more controversial realm than is first suggested by the design of a piece of functional middleware, since one may speculate that the provision of a mechanism for accessing not just electronic publications but also their underlying data raises fresh questions about the nature and meaning of the terms *scholarly* or *scientific publishing*. Reflecting on the open access publishing and repository movements, one detects a strong current of opinion that making data available does not constitute publication, which benign strategy contributes of course to the avoidance of unhelpful quarrels with publishers; but greater flexibility of interpretation and less defensiveness would be both appropriate and defensible, since the publication of scholarly papers and the dissemination of data are necessarily distinct acts, each being defined by their particular purpose. Any set of data selected specifically for inclusion in, or as the basis for a scientific paper is chosen with the principal purpose of helping to persuade the reader to accept a hypothesis or theory as proven, and its value is gauged by the degree to which it supports the effectiveness of the set piece of rhetoric that is the paper. The larger collection of data from a research programme, possibly archived in a source repository, does not serve that same purpose of persuasion. Indeed, it may be argued that by making this broader cache of data accessible via a link from a scientific paper to its source repository could even subvert the arguments in the paper, should there be weaknesses in the data or the research, although from a different perspective this does strengthen the case for the bi-directional link as a mechanism for ensuring the integrity of the source data. So whether making data from a source repository publicly available is an act of dissemination or publication, the answer is probably irrelevant. What is more to the point is the impact from enabling dual accessibility.

The impetus for Project StORe came from a belief held by members of the research library community that an achievable set of functional enhancements to both source and output repositories could be identified and built, on a generic basis, as a piece of middleware, and that this might be approached in a manner similar to the way in which digital library technologies have produced generic tools in other heterogeneous environments, such as ‘metasearch’ interfaces to publisher and local databases, metadata harvesters and link resolvers. These tools are based upon recent digital library protocols and standards such as OAI-PMH,[3] qualified Dublin Core[4] and OpenURL.[5] Project StORe was therefore conceived as a vehicle for undertaking the essential groundwork preparatory to building a production system solution that would meet the requirements for permitting useful interoperation between the two repository types, and it would be undertaken using the systems, standards and metadata protocols developed and used in other JISC projects, where appropriate, to ensure the widest possible interoperability. Its rationale would be that of a proof of concept, but from the start there was a firm aspiration to deliver an authentic pilot infrastructure capable of translation across multiple disciplines.

## 2 Methodology

In the first phase of the project a survey of researchers was conducted across seven scientific disciplines in the UK to understand their workflows and working philosophies, as well as to identify norms in the use of source and output repositories. The disciplines investigated were archaeology, astronomy, biochemistry, the biosciences, chemistry, physics and social sciences. The astronomy survey had a broader base, including members of the astronomy research community in the USA, in recognition of the internationally collaborative work undertaken by astronomy research teams at Edinburgh and Johns Hopkins universities and the discipline’s separate Mellon-funded analysis of repositories and applications. The survey, which was carried out over four months in 2006, first through an online questionnaire and subsequently by one-to-one interviews, addressed such issues as the existence of common attributes across disciplines (in terms of the data formats employed, the quality and method of metadata assignment, and the volume of data produced), the functional enhancements to repositories that were considered to be desirable, and the nature of problems experienced in the use of repositories. Cultural and organisational issues were also investigated, ranging from attitudes towards the concept of open access publishing to the measures employed for sharing and protecting data. Invitations to participate in the online questionnaire were sent to 3,700 scientific researchers and produced a return in excess of 10%, whilst the in-depth interviews were held with between 10 and 15 respondents per discipline, selected to ensure an equitable representation from all stages of the academic/research career path. Each individual discipline survey produced a published study that described the source and output repositories used by members of that discipline, including a brief history and statistical information on their use, with a detailed analysis of responses to the questionnaire and the structured interviews. These reports, which have been archived in the Edinburgh Research Archive (ERA), also incorporate scenarios and use cases.[6]

Project partners at university libraries identified staff to undertake the discipline surveys, with a view to exploiting their knowledge and the effectiveness of their relationships with researchers 'on the ground'. The libraries responsible for the survey work and the disciplines they surveyed are shown in Table 1.

| **Surveying University Library** | **Subject** |
|---|---|
| Edinburgh (lead) / Johns Hopkins | Astronomy |
| Birmingham | Physics |
| Imperial College | Chemistry |
| London School of Economics | Social Sciences |
| Manchester | Biosciences |
| University College London | Biochemistry |
| York (for the White Rose Partnership) | Archaeology |

**Table 1: Project partners & survey disciplines**

Whilst it was important to the design of a relevant and appropriate solution that actual research working practices and environments would be identified and understood, the survey team's principal role was to address the requirements for new functionality within source and output repositories that would permit interoperability from the point of ingest, so that authors of papers could insert links to data and to published/unpublished papers, associating newly deposited publications with data held in data repositories. It was anticipated that a number of new operations could be supported within the two types of repository, both for academic submitters and for repository users, including automatic link creation, automatic embedding of source repository metadata, and a facility to run operations upon data. The desirability of these features was explored in depth during the interviews.

Upon completion of the survey, a business analysis of the survey reports was undertaken by staff at the UK Data Archive (UKDA).[7] This analysis was used as the foundation for a generic technical specification of the proposed bi-directional link, with the aim of translating real requests for 'missing' functionality into a structured technical architecture. The assumptions and deductions made in the business analysis were then tested with research active staff at the University of Essex and with library professionals from the London School of Economics (LSE), leading to further refinements to the specification.

In the final phase of this development process, the generic technical specification has provided the platform for the pilot implementation of a working bi-directional link. This has featured social sciences data and publications exclusively, using the *UKDA* as the source repository and the LSE's *Research Articles Online* as the test output repository, augmented by a further link to a prototype institutional repository at the University of Essex, which served as a control mechanism. It should be emphasised that limiting the pilot to only one of the original seven disciplines has been necessary to meet the logistical constraints of a test environment, but in building that environment the full set of requirements established by the survey of seven disciplines has been incorporated with a view to proving the middleware as a generic, non-discipline specific tool.

Throughout, the rationale of Project StORe has been to anchor technical and user aspirations to the pursuit of practical benefits. During the pilot implementation, a critical element of the process has been user testing, involving members of the original cohort who responded to the survey, and at its conclusion the pilot demonstrator has been subject to a rigorous, independent evaluation by the National Centre for e-Social Science,[8] which has depended for its legitimacy upon user participation in a series of workshops.[9]

# 3 Survey and Analysis

A majority (85%) of respondents to the StORe survey judged the provision of a bi-directional link as likely to prove advantageous to the research process, with a small preference overall for an output to source link. Key benefits were described as an opportunity to access the large data sets it is not possible to reproduce in an article;

and more specifically, an output to source connection would enable the comparison of results, thereby providing the means to authenticate claims made, which was deemed to be of particular value where claims are considered controversial.

By selecting from prepared lists, respondents were asked to identify the data types and their formats that might be generated during research, with the range of data types given in the lists appearing to satisfy the majority as being representative, and with no data type receiving a nil response. A further 32 *Other* types were also declared but were found to describe either a sub-type of items from the lists, the name of experimental equipment or process-specific data sets. Nonetheless, across and within the seven disciplines, the volume, range and diversity of data produced was confirmed as considerable. Whilst generic types such as drawings, plots, images and text-based files scored highly, each showing in excess of 150 responses, noteworthy scores were attributed to more specialised types such as radiographic data (11), remote sensing surveys (15) and gene/protein sequences (42).

In terms of data format, image files, spreadsheets and word processed files comprised the majority, with around 200 responses each. In the next tier, plain text, database files, portable document format and tables/catalogues all scored more than 100 responses. Of the 76 *Other* formats volunteered by respondents, those that were not species from the main selection list tended to be proprietary and linked to specific discipline processes or equipment. Of greater significance to the design and maintenance of links from publications to their source data is that almost three quarters of the survey's respondents were found to generate and use complex data sets (i.e. data produced and held in combinations of data formats and files).

All of the seven disciplines identified barriers to the deposit of data or publications in repositories, citing time constraints, the bureaucracy imposed by repository administration and structures, or constraints arising from their own or others' intellectual property rights. A perceived inconsistency across all repositories was also reported in terms of content coverage and in the standards and methods used for keywords, metadata and data formats. It was in this latter area that the most powerful consensus was found amongst the survey cohort, with the appropriate assignment of metadata being roundly acknowledged as critical and demanding, both intellectually and in the time required to do it well. Perversely, this consensus on the need for good metadata did not necessarily translate into good practice, there being a high level of self-assignment and with limited evidence that standard schema or thesauri were being employed. Perceived responsibilities for metadata assignment are illustrated by the following table from the StORe questionnaire.

| | | |
|---|---|---|
| I decide which terms to use and I assign them | | **212** |
| Research colleagues assign metadata on the team's behalf | | **55** |
| Research support staff assign metadata on the team's behalf | | **22** |
| Metadata are assigned by library/information services staff | | **4** |
| Metadata are assigned by the repository administrators | | **37** |
| Metadata are generated automatically | | **63** |
| It is not known who assigns metadata | | **68** |
| Other (please specify) | | **37** |

**Table 2: The Assignment of Metadata to Research Data**

In order to establish whether there is a core set of metadata that might satisfy the needs of researchers in the seven disciplines, respondents were invited to identify key terms from a predefined list and to suggest their additional requirements. A large majority subscribed to the list as representing a functional generic suite of metadata, selecting such terms as project title, description and reference numbers, together with keywords, project and publication dates and format. Only 58 *Other* terms were suggested, and these were found to be highly discipline specific (e.g. archaeological period, celestial object, position and observation date, chemical entity, protein sequence).

As shown in Table 2, the subject of metadata provision revealed a broad spectrum of awareness and response amongst the survey cohort that was sustained when they were asked to indicate the point at which metadata are assigned. Assignment 'during file saving' attracted the highest score of 142, but there was insufficient evidence to deduce whether such a practice represented a properly structured activity or merely the casualty of afterthought. More reassuring were responses to the options 'Prior to data creation' (82), 'As part of the indexing process' (98) and 'When submitting data to the repository' (89). Of some concern were the 35 respondents to this question who believed no metadata were being assigned to their research output, with a further 75 admitting they were not sure at which stage metadata are assigned.

The disjunction between aspiration and practice in the assignment of metadata is perhaps explained by tensions between the prevailing research culture and embedded attitudes towards the support services. It was made clear during the StORe survey that researchers from all disciplines favoured self-reliance in matters associated with data management and the use of repositories, as opposed to the provision of institutional support from the library or other areas of professional expertise. The inherent culture of self-sufficiency within research groups or programmes, where normal practice is to manage all aspects of the research lifecycle internally, was evident from statements submitted during the StORe survey. Whilst this culture has given rise to the development of some highly effective data repositories focused on serving specific disciplines, the general effectiveness of a self-sufficient approach to accessing, organising, promulgating and curating data was not demonstrated across the scientific research spectrum.

National and international strategies for data deposit and preservation are of course already emerging. One can point, for example, to the Wellcome Trust's flagship initiative to mandate the deposit of research publications in the biosciences, which mandate is anticipated will extend to the deposit of data; or to the astronomy community's Virtual Observatory, an initiative to make all the astronomy data in the world easy to access.[10] They are not isolated examples, but when one considers the research milieu as a whole their considerable progress was found not to be typical. At the level of the individual researcher, whether asked about metadata assignment in particular or data management in general, responses such as "it's my problem, I'll deal with it" were commonplace. Whilst libraries have conducted advocacy campaigns on behalf of open access publishing and repositories, in some cases providing technical expertise to support the use of repositories, researchers canvassed by the StORe survey in most cases perceived there was no support available, they had little confidence in what support was known to be provided, and they claimed sufficient familiarity with information technology to consider themselves self-reliant. Yet at the same time as declaring they would not normally associate the management of research data with librarians, and evincing little apparent demand for assistance in seeking and navigating information, there was evidence of a clear requirement for information intermediaries to assist not only in the construction and maintenance of metadata but also in the preservation and curation of data. This dichotomy was reflected in a further aspect of the survey, which concerned researchers' attitudes towards making data available, and would prove a singular force in the design of the StORe middleware.

With few exceptions, respondents to the survey supported the statement that it should be a requirement for data from publicly funded research to be made freely available, but generally with the caveat that access should be restricted until results are published in a paper, in order to prevent data scavenging. Others noted that whilst this might be a creditable aspiration, without a data administrator it represented a potentially large burden from editing, compiling and sanctioning the release of data. In fact, both the provision of access and the sharing of data were found to be constrained by a lack of confidence in processes, and it was difficult to conclude whether some practices were deliberately designed to frustrate accessibility. For example, the storage of unique and original research on PCs and laptops was found to be common practice, and the failure to take a more relaxed approach to access was influenced by a perceived absence of adequate protection in networked systems. As one respondent described his data management regime: "data is held on secured CDs in encrypted format with only an identifying code. The codebook is kept physically separate".

The StORe survey revealed a range of diversity in practice and attitude, both within and between the seven disciplines, but with a consistently firm body of consensus when it came to explaining fundamental needs. When searching for information, a universal preference for simple keyword searching was declared and browsing amongst library shelves appears to have been replaced by browsing within repositories and other online resources. This practice is of course only effective when enabled by the functional efficacy of application and metadata structures, designed by system and data experts to meet the clamour for a 'Google-type' approach to searching.

# 4 The Generic Model

The business analysis that followed the StORe survey revealed sufficient shared ground between the disciplines to suggest the basis for a common model. To recap, an examination of the discipline-specific reports produced a majority in every discipline favouring two-way links between data repositories and publications, but with barriers to the actual deposit of data or publications found to be a consequence of time constraints, organisational bureaucracy or concerns over intellectual property rights, although the concept of data sharing was considered fundamental and important. A perceived inconsistency across all repositories in terms of coverage, standards and data formats was reported, with a simple 'Google type' approach to searching being preferred. Researchers from all disciplines also seemed to exhibit self-reliance in matters of data management and in the use of repositories, whilst recognising the need for assistance in the provision of some common minimum metadata.

Taking this level of consensus, the design of the model for a bi-directional link has adopted a Web 2.0 type approach, similar to existing FOAF (Friend Of A Friend) services such as Flickr or MySpace, but incorporating a federation of institutional, source and output repositories rather than one central area where digital objects are deposited. Articulation of a Web 2.0 rationale for the middleware has been a deliberate decision aimed at meeting cultural aspirations for self-determination and those individual anxieties concerning data ownership that were revealed during the survey, since it places control firmly in the hands of the researchers. In this model, objects deposited in federated repositories would be referenced by persistent identifiers that include domain identifiers, with researchers depositing digital material in various formats at their institutional repositories until the data and publications are ready to be made publicly available at linked source and output repositories. This focus on the institutional repository environment is predicted to have further value in providing a context for future implementations of asset-based research data repositories, in cases where global services from established discipline platforms such as astronomy's Virtual Observatory or the social sciences' UKDA are not provided, and discipline needs could be met instead by a regime of institutional data curation.

What may be described as the central StORe portal has been designed as an OAI-based aggregator service that will harvest the contents of a federation's repositories and provide a simple search facility based on centralised indexes. This basic level of searching can be enhanced for individual disciplines by the inclusion of domain ontologies, reflecting the need highlighted in the survey to enable discipline-specific terminology. All digital objects will be title visible to all, but researchers can restrict access to non-public objects to communities of project-specific colleagues, institutional colleagues, personal colleagues, or all of these. This is similar to the option for restricting access to family and/or friends in Flickr, in order to bar public access to private photographs, and is again a direct attempt to satisfy the demands of researchers to remain in command of their data.

Access management has proved to be a defining feature of the StORe middleware. Some data repositories are open to all enquirers, while others are password-protected, and in a scenario where users of open access research publications wish to view data in repositories to which access is normally controlled, a validation process will be required in order to allow temporary access rights. In this context we have investigated the authentication and authorisation issues involved with reference to the developing international work on Shibboleth, a federation-based architecture that enables organisations to build single sign-on environments for accessing Web-based resources.[11] Whilst it is not yet in place, it is planned that a production version of the central StORe portal will authenticate through Shibboleth, using a simple deposit interface to request the minimum amount of mandatory metadata for each object, identify the group or individual to which it is accessible and check whether it is a candidate for public submission. Until Shibboleth is adopted, we are applying a dummy Shibboleth mechanism for allocating user names and passwords. This will trigger an automatic process for setting up user accounts when legitimate users log in for the first time.

The minimum metadata required for any individual item is a title, provided the item is being associated with project data in a repository already assigned the metadata elements *author*, *title*, *geography*, *time*, *keywords* and *abstract*. The digital object will be deposited in the researcher's institutional repository, whilst the metadata and access conditions will be stored centrally; in turn, the search indexes will be built up from the centrally held metadata and harvesting from the objects themselves. This harvesting can also be used in the creation of the discipline-specific ontologies needed to satisfy metadata requirements that are not met by the generic core. Both source and output repositories in the federation will regularly trawl for potential acquisitions and, if a publication or data are accepted, the repository will supply a public link to a peer-reviewed version of the publication or to the data.

Hence, the generic model planned to be tested by the pilot demonstrator combines informal networking and sharing of data with a public access system that supports stronger links between data sources and publications. A user entering a StORe generic portal would log in to authenticate and the system will respond by determining his/her organisation, recorded preferences and known colleagues. Options would then be made available to browse any new activity of colleagues; to browse any objects available to the user (i.e. the user's own and other colleagues' objects); to search all discipline-specific or all repository-specific objects, with a further option to filter on a temporal basis; to deposit an object; to create a new project; to make an object available to another user; to request that an object be made available; to submit an object to an output repository for publication; to submit an object to a source repository for preservation; to download a repository object; and to edit, delete, organise or manage the user's own objects.

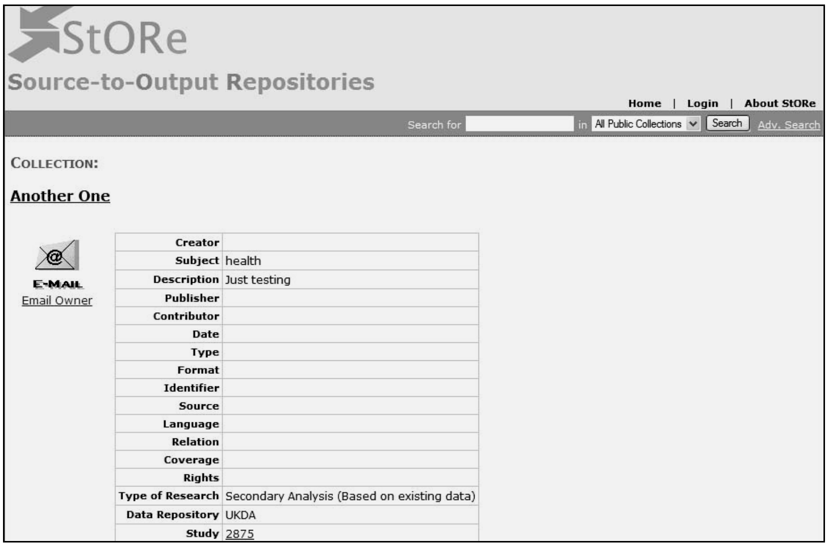
It was clear from the outset that the success of this model would be determined by three factors. Researcher acceptance of Web 2.0 technologies was essential, and we have been actively encouraged here by the younger members of the user testing cohort who already work routinely within that environment. Persuading researchers to use a third party portal for deposit into their local institutional repository was also acknowledged to be challenging, whilst the third and possibly most difficult obstacle lay in the resolution of potential security and policy objections to sharing sensitive data across institutions. Eventually, it was decided that these barriers could be broken down in a stage by stage approach that would embed a federation in the established publishing process and restrict the sharing of non-public data to institutional colleagues. A demonstration of simplicity would be the key to stage one, with the objects stored required to be identifiable only by title, discipline, project, file type and format, employing minimum Dublin Core metadata elements. In the second stage, each individual institutional repository would act as a portal to itself and all the domain specific source and output repositories in its federation, thereby preserving familiarity of the working environment but allowing the addition of Web 2.0 and FOAF features. Only at stage three would the concept of a StORe subject or domain portal be openly introduced to the discipline-specific elements of the federated repositories. Here, one solution to security concerns would be the temporary copying of protected objects to the portal for download within a prescribed period.

Looking beyond the pilot environment, this approach offers wider coverage, more choice of source and output repositories and more scope for Web 2.0 service features. There could even be a common interface for deposit to individual institutional repositories, and it was envisaged that listing of forthcoming conferences, wikis, and other networking facilities might encourage use. The final stage would see the full generic solution implemented, comprising the entire federated institutional, source and output repositories that have adopted the approach outlined in stage one. This solution is well placed to encourage cross-disciplinary research, a key driver in the modern research environment, although metadata mappings will have to be employed and even more additional features devised to encourage the use of such a universal portal.

# 5 A Passage Through Project Store

StORe's pilot demonstrator was built for a test federation using the UKDA as source repository and the LSE's *Research Articles Online* as the output repository, complemented by a prototype institutional repository at the University of Essex.[12] Options for linking to a commercial publisher had also been explored but were considered logistically too ambitious for a pilot implementation. The pilot was designed and implemented between November 2006 and April 2007, and what follows is an abridged system walkthrough showing how items (data and publications) are managed.[13] This description is of a standalone system, but in a live working environment access could be initiated within an electronic article in an output repository or from a source repository having an association with the federation.

In the pilot, as in a working system, it is possible for an unregistered user to search or browse across all or specific research collections in the federation, but any titles marked as private will not function as a hyperlink to their content. Collection metadata can, however, be seen via a *View Collection* link. If the research project from which the target collection was generated involved the secondary analysis of existing data, a link to the underlying data will already exist, and will take the user to the relevant Web page of the supporting source repository. If the collection owner has agreed, then a further link will appear, allowing users to send an email requesting additional details or to be granted access to items in the collection.


StORe
Source-to-Output Repositories
Home | Login | About StORe
Search for
in All Public Collections
Search
Adv. Search
E-MAIL
Email Owner

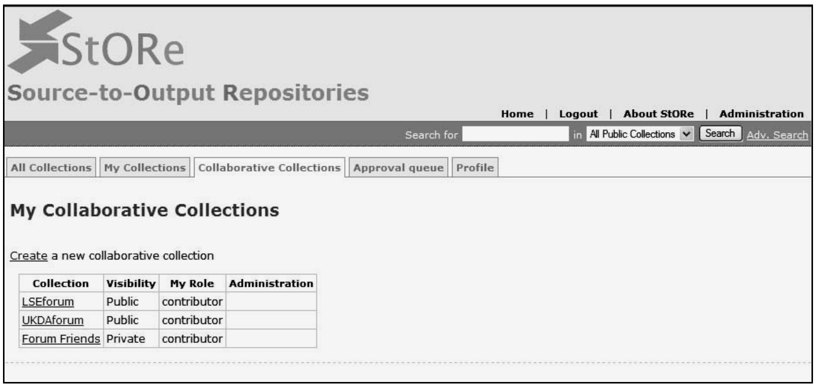
COLLECTION:

Another One

| | |
|---|---|
| Creator | |
| Subject | health |
| Description | Just testing |
| Publisher | |
| Contributor | |
| Date | |
| Type | |
| Format | |
| Identifier | |
| Source | |
| Language | |
| Relation | |
| Coverage | |
| Rights | |
| Type of Research | Secondary Analysis (Based on existing data) |
| Data Repository | UKDA |
| Study | 2875 |

**Figure 1: View Collection Metadata Screen**

Registered users logging in to the pilot federation can view the content of all items in their public and private solely-owned collections. They can also see those items in public collaborative collections with the UKDA or *Research Articles Online* where they are a contributor, and may view public or private collaborative collections made with project colleagues or 'friends', where they are identified as either contributor or administrator. Collaborative collections are linked via a unique *LinkID*. In the example below, the user (identified as Forum) is a member of a private collaborative collection created by another researcher in order to share documents with Forum. Each collaborative collection is distinguished as a collection type, either source/archive, output/publisher or user/researcher. The logged-in and authenticated user has access to full functionality and can create private or public, solely-owned or collaborative collections, including an option to allocate other registered users to a collaboration.


StORe
Source-to-Output Repositories
Home | Logout | About StORe | Administration
Search for
in All Public Collections
Search
Adv. Search
All Collections
My Collections
Collaborative Collections
Approval queue
Profile

My Collaborative Collections

Create a new collaborative collection

| Collection | Visibility | My Role | Administration |
|---|---|---|---|
| LSEforum | Public | contributor | |
| UKDAforum | Public | contributor | |
| Forum Friends | Private | contributor | |

**Figure 2: Collaborative Collections**

Figure 5, overleaf, shows how the process of adding metadata to a publication has been kept simple. At collection level, apart from the collection name and description, only subject terms and the type of research and study (if secondary research) are mandatory. All other Dublin core fields are optional. The subject terms can be directly typed into the box or chosen from a list of tags displayed at the right-hand side of the page. The type of research is selected from a drop-down menu (Figure 3),

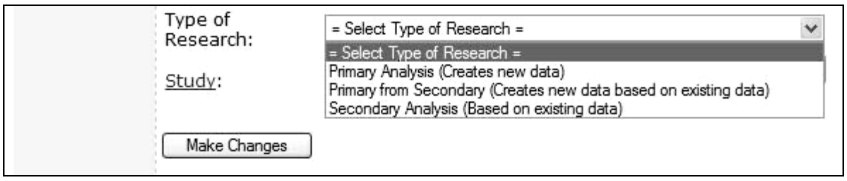

Type of Research:
= Select Type of Research =
= Select Type of Research =
Primary Analysis (Creates new data)
Primary from Secondary (Creates new data based on existing data)
Secondary Analysis (Based on existing data)
Study:
Make Changes

**Figure 3: Allocating Research Type**

and a study number corresponding to the number assigned to the corresponding data within the source repository is chosen from a further drop-down list (Figure 4).

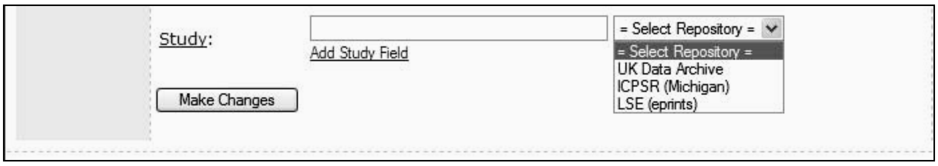
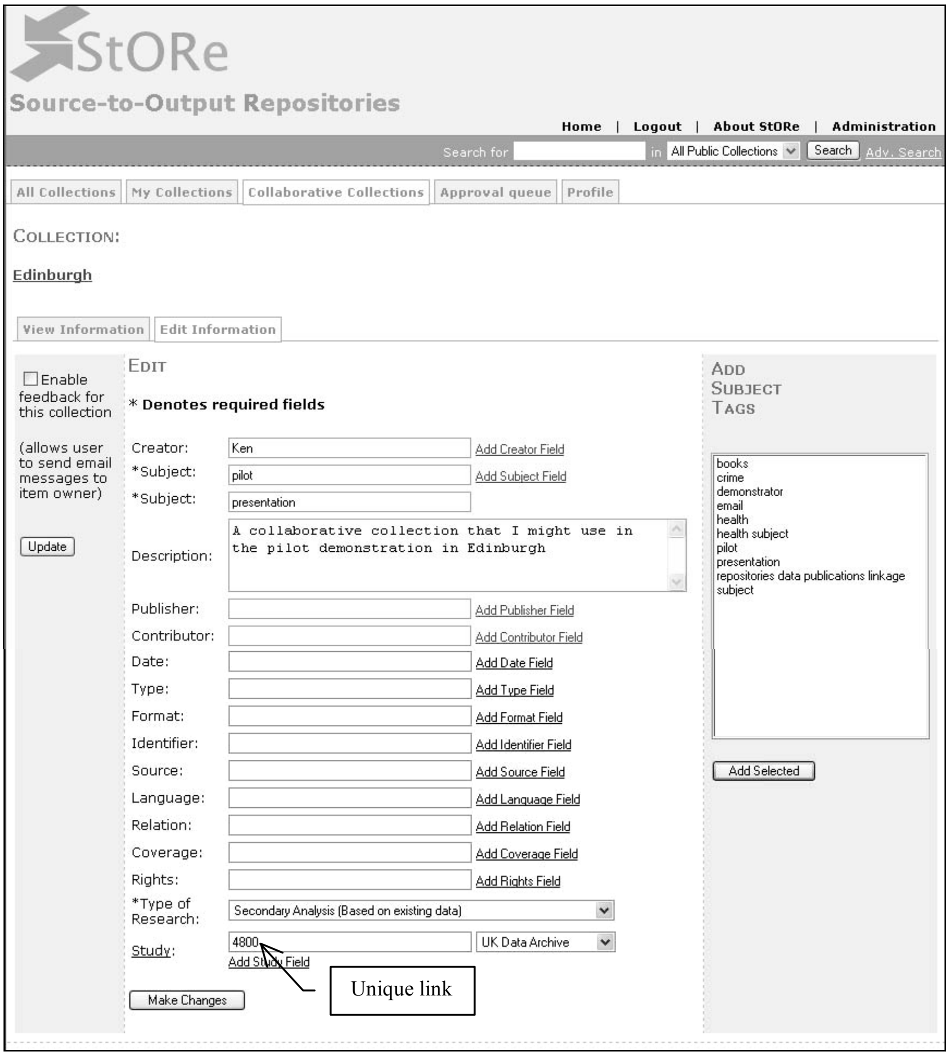


**Figure 4: Selecting The Study Number**

The study number then becomes a link to the appropriate page within the repository's web site.



COLLECTION:

Edinburgh

EDIT

* Denotes required fields

Creator: Ken

*Subject: pilot

*Subject: presentation

Description: A collaborative collection that I might use in the pilot demonstration in Edinburgh

Publisher:

Contributor:

Date:

Type:

Format:

Identifier:

Source:

Language:

Relation:

Coverage:

Rights:

*Type of Research: Secondary Analysis (Based on existing data)

Study: 4800 UK Data Archive

**Figure 5: Assignment of Metadata**

Moving a data item to the UKDA collaborative collection is a two-tier process. First, the data's identity is verified (which will enable publications based on this data to be moved and approved in *Research Articles Online*) and, where required, an embargo can be set by the data owner. Once verified, an acquisition number is assigned to the data item in the UKDA collaborative collection, which as already intimated will subsequently be assigned to any associated publications moved to a *Research Articles Online* collaborative collection. Upon approval by the UKDA this acquisition number is replaced by the actual research study number, which will function as a link to the data from its publications in *Research Articles Online*.

In StORe, individual items or folders are added to a collection either singly or bundled. Only the provision of an additional title and file name (or URL) is required, since each item adopts all the metadata associated with the collection itself. Files in different formats (Word or PDF documents, URLs, image files, etc.) are associated to an item or folder, and the Dublin Core fields may be edited if required to produce a more specific metadata record. When a scientific paper ready for publication is moved from a researcher's institutional repository into a collaborative collection owned by *Research Articles Online*, all the metadata associated with it moves as well. Simultaneously, the middleware automatically assigns a metadata term to identify the collection of origin, and confirms that corresponding data exists in the UKDA collaborative collection. It also provides functionality enabling the addition of further files or URLs to the item, or to add additional metadata.

# 6 Conclusions

The StORe pilot has demonstrated the feasibility of a bi-directional link within the specific context of a single discipline. However, despite the level of consensus identified by the survey, discipline variations would need to be managed during export of the StORe model across other domains. Individual institutional repositories will also contain different file types and formats, and will apply different metadata standards. For certain disciplines data interpretation, manipulation and methodology are as, if not more significant than access to the raw data, and although a simple search might cross disciplines, more advanced discipline-specific searches would be more in demand, with the resulting hit lists, relevance ranking and sorting being different for each discipline. Consequently, both subject and global portals will require different Web 2.0 features for each discipline.

Recognising the key preferences and practices of researchers interviewed during the StORe survey, the solution developed showed that traditional practices for the informal networking and sharing of data could be combined with a public access system supporting stronger links between data sources and publications. The StORe solution gives researchers the means to manage a level of privacy and access defined by themselves, countering expressions of apprehension towards full open access, which some saw as a threat to data ownership. It also offers a simple Google type search, preferred amongst the majority of those surveyed, and viewed by many as an effective tool for replacing the option of browsing amongst shelves in a library, although Boolean operators and wildcard functionality are made available for more advanced searches. Using the StORe middleware, researchers can move seamlessly around the research data environment and its outputs, stepping from within an electronic publication directly to the data upon which its findings were based, or linking instantly to all the publications that have resulted from a particular research dataset. By intrinsically connecting the process of publishing scientific papers with the provision of their underlying data, StORe has also broadened the connotation of the terms *publish* and *publication*.

# References

[1] http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories.aspx

[2] MILLER, G. *A Scientist's Nightmare: Software Problem Leads to Five Retractions*. Science, 22 December 2006, pp. 1856-1857

[3] An explanation of the Open Archives Initiative (OAI) and the OAI protocol for Metadata Harvesting may be found at http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/faq/oai/

[4] The Dublin Core metadata element set is explained at http://www.ukoln.ac.uk/metadata/resources/dc/

[5] An OpenURL demonstrator can be found at http://www.ukoln.ac.uk/distributed-systems/openurl/

[6] The individual discipline reports, together with the survey overview, may be examined as documents within the Edinburgh Research Archive, http://www.era.lib.ed.ac.uk/handle/1842/1412 or at the project wiki, http://jiscstore.jot.com/SurveyPhase

[7] UK Data Archive at the University of Essex, http://www.data-archive.ac.uk

[8] http://www.ncess.ac.uk

[9] The NCeSS evaluation plan can be examined at http://jiscstore.jot.com/EvaluationOfPilot

[10] http://www.virtualobservatory.org/

[11] http://www.athensams.net/federations/shibboleth_intro.aspx

[12] Both institutional output repositories have been constructed using open source software: *ePrints* at the LSE and *Fedora* at Essex.

[13] A full walkthrough of the StORe middleware can be accessed at the project wiki, http://jiscstore.jot.com/PilotDemonstrator

# Challenges in the Selection, Design and Implementation of an Online Submission and Peer Review System for STM Journals

*Judy Best; Richard Akerman*

Technology and Research; Canada Institute for Scientific and Technical Information
National Research Council of Canada, 1200 Montreal Road, Ottawa, Ontario K1A 0R6 Canada
e-mail: judy.best@nrc-cnrc.gc.ca; richard.akerman@nrc-cnrc.gc.ca

## Abstract

Two international scientific publishers collaborated to develop an Online Submission and Peer Review System (OSPREY) for their journals. Our goals were to meet market demand, increase editorial efficiency and streamline the transition from peer review to publishing. One of the publishers (National Research Council (NRC) Research Press, Canada) had previously purchased a third-party system that was subsequently discontinued by the vendor. Because of this experience and its complex requirements, we decided to build rather than buy a new system. The collaboration with the second publisher, Commonwealth Scientific and Industrial Research Organisation (CSIRO) Publishing, Australia, allowed sharing of resources within a common vision and goals. Agile development through the use of iterations allowed us to continuously add functionality, make improvements and incorporate new requirements. The development team included technical staff as well as stakeholders, future users, business analysts and project managers. The architecture chosen was based on open source technologies, with Java servlets and Java Server Pages for the Web interface. OSPREY currently supports 32 journals at the two publishers. Users accomplish all regular tasks in peer review (submission, selection and invitation of reviewers, submission of review, recommendations and decision) through the software. Editorial staff verifies submissions, sends correspondence and assigns customizable roles and tasks. All tasks are accomplished through a Web browser accessing the application on central servers at the publisher, with no special software or configuration required for any users. Currently, the system integrates with the publishing system by generating manuscript metadata in an XML format, although closer integration with a workflow management system is planned. Since OSPREY implementation, the number of submissions has risen, although marketing and higher ranking of the journals are also factors. For the future, we plan to add new functionality for business tasks and for parsing, tagging and linking of article references.

**Keywords:** on-line peer review; open source technologies; software architecture; workflow transition

## 1 Introduction

Our Online Submission and Peer Review System (OSPREY) is a web-based manuscript submission and peer review system used by scholarly publishers and societies to automate and streamline the publication process. It supports the submission of articles and the subsequent peer review process within a configurable automated electronic environment.

Communication with authors, reviewers and editors is handled by e-mail using customizable templates within the system. This is one of many features customizable by publisher or journal; others include copyright and reviewer forms and branding. Authors can upload a single file or multiple files consisting of many file types, and an Adobe Portable Document Format (PDF) file is created immediately. Metadata of accepted manuscripts in an XML format is integrated into the publishing system, however; some manual intervention is still required.

OSPREY is developed and maintained in collaboration between two leading Canadian and Australian scientific publishers. These are Commonwealth Scientific and Industrial Research Organisation (CSIRO) Publishing, Australia, and National Research Council (NRC) Research Press, part of the Canada Institute for Scientific and Technical Information (CISTI).

The objectives were to meet market demands, reduce turnaround times, increase efficiency within the editorial offices and to streamline the transition between peer review and publishing.

## 1.1 Background

**Publishing**

CISTI is a science library and a world leader in document delivery for all areas of science, technology, engineering and medicine. CISTI's publishing arm, NRC Research Press, has been a traditional publisher since 1929 and currently publishes 16 international print and online STM journals. With its resources and expertise in place, NRC Research Press began offering its print and electronic publishing services to other Canadian publishers in the late 1990s; as a result, NRC Research Press also publishes 15 client journals.

CSIRO Publishing operates as an independent science and technology publisher with a global reputation for quality products and services. Its internationally recognized publishing programme covers a wide range of scientific disciplines, including agriculture, plant and animal sciences, and environmental management. CSIRO Publishing publishes content in print and online. CSIRO Publishing is an autonomous business unit within CSIRO.

NRC Research Press moved into the electronic publishing world by first publishing content in PDF format on the Web for its subscribers in 1996 and later implementing a process to generate SGML metadata for searching, distribution to aggregators and dynamic generation of table of contents and abstract HTML pages on the Internet. NRC Research Press has since implemented an XML publishing system (Fig. 1) in which content is tagged according to a very rich custom Document type Definition (DTD) and published in print, PDF and HTML formats.

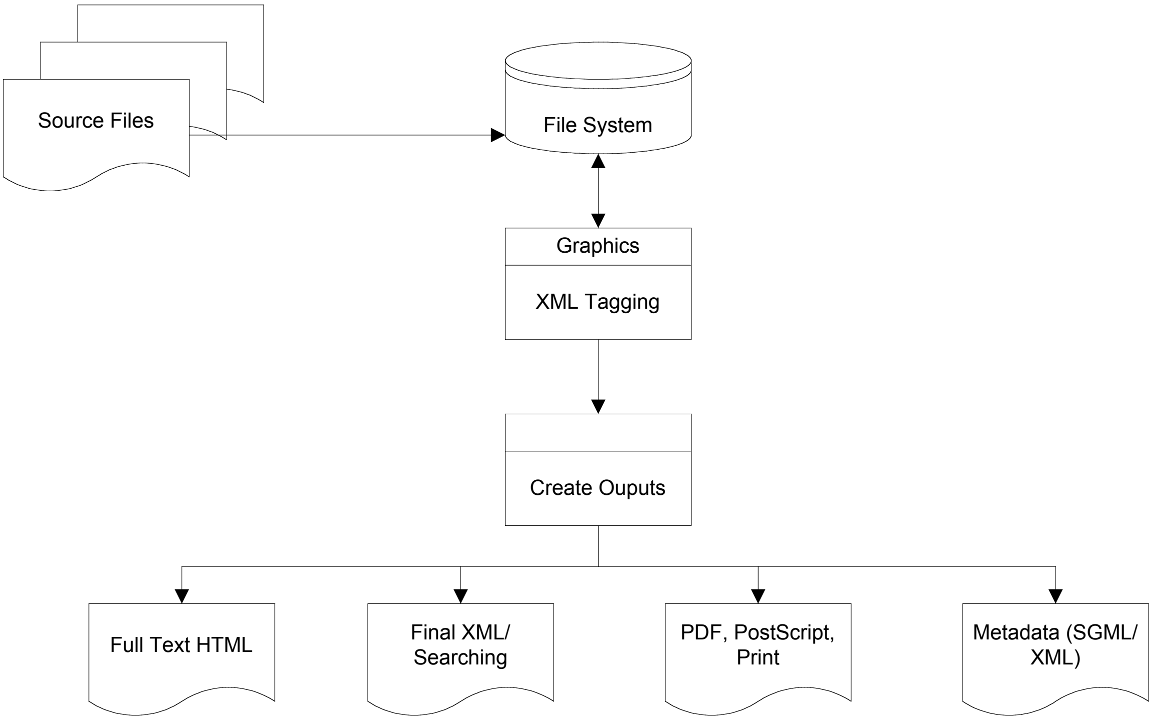


**Figure 1: NRC Research Press Publishing Process**

**Transition to Electronic Peer Review**

Traditionally, the submission of manuscripts, the peer review process and the management of the editorial process were paper-based and manual. Authors, reviewers and the journal offices would rely on mail, fax and courier services to deliver manuscripts, reviews and decisions throughout the workflow. Gradually the journal offices moved toward using e-mail for quicker transmission of manuscripts.

In early 2000, NRC Research Press purchased its first online submission and peer review system (PaperPath 2000). This new technology would help bridge the gap in the digital world between the peer review process and publication. It allowed authors and reviewers to submit manuscripts and reviews using a web browser and editorial offices to manage the workflow using third-party client software installed on their desktops. PaperPath 2000 was implemented in 15 journal offices over a 10-month period. In late fall of 2001 the vendor discontinued PaperPath 2000, leaving NRC Research Press with an unsupported product.

NRC Research Press supported PaperPath 2000 for another year and then began its search for a new online peer review system. Many Commercial-Off-The-Shelf (COTS) or licensed products were evolving, however; they did not adequately support a single sign-on, our diverse editorial workflows or our need for a multi-language (English and French) user interface with the potential to expand to other languages. There are many other factors to take into account when deciding whether to purchase or build, after evaluating each option, a decision was made to build in-house. Factors in making the decision were:

- available features in COTS or licensed product
- associated costs (one-time costs and maintenance and support)
- our diverse editorial workflows
- requirement for an English and French interface
- CISTI's technical skills and infrastructure
- storage of confidential data off site
- accessibility to data for customer relationship management
- integration to publishing and subscription management systems
- past experiences with the purchase and implementation of its previous online peer review application

Given the factors listed above we determined that the best approach was to leverage our internal capabilties and ensure our continued access to the source code and systems.

A significant effort was put into developing a solution that would ensure ease of use, minimal overhead and support costs, flexibility, scalability and future growth of features. The system was developed to current standards, using Java and XML for the application and Oracle as its database engine. The system architecture uses a component-based design methodology that enables flexibility and the potential for growth. It supports loose coupling and high cohesion, not only from a functional point of view, but also in the underlying data architecture.

# 2 Methodology

## 2.1 Collaboration

A key benefit to a collaborative approach is the ability to share resources (i.e. people and money) and to gather a broader set of requirements. Our experiences have shown us that identifying the roles and responsibilities, following sound project management principles and effectively communicating among team members, users and stakeholders are critical for success.

The two organizations shared a common vision, priorities and goals, helping us to develop ways to work collectively and to communicate effectively. Working together enabled us to draw on the skills and experiences of two scientific publishers. The stakeholders were instrumental in giving the project the priority and support required to develop OSPREY.

## 2.2 Development Approach

The international collaboration required rapid response to requests and iterations in the development of the user interface. For these reasons, an agile development methodology was chosen [1]. Iterations, which included use cases, analysis, design, implementation and tests, were 6 weeks long. That meant that both sides of the partnership could see new working functionality frequently. This approach made it possible to continuously add new functionality and make improvements. Each iteration could deliver minimal functionality. As new requirements were uncovered, a new iteration would replace the previous one.

Although each organization would install and support OSPREY, the technical development was completed at CISTI by two developers. An infrastructure was put in place to manage concurrent versions of source files and a centralized build function of the application.

The technical project team consisted of two developers, a system administrator, a database administrator and an application architect from CISTI. Representatives from CISTI and CSIRO rounded out the team and included stakeholders, users, business analysts and project managers.

## 2.3 OSPREY Architecture

There were several options available to deliver a web-based application to clients: Java Applets, Visual Basic, or server-side solutions such as Perl, ColdFusion, ASP or Java Servlets. CISTI's previous implementation of PaperPath 2000 required specific client-side software and hardware which added an additional burden to our technical support team. It was determined early in the project that the best means to meet the objectives was to implement a web-based application, requiring only a standard web browser as the client interface and no additional software on the client side. A Windows platform was chosen for development, however; the application could be ported to a Linux environment if required.

### 2.3.1 Open Source Technologies

Java, which is platform independent, was used as the programming language. The web interface was implemented with Java servlets and Java Server Pages (JSP), allowing loose coupling with the client-side.[2] CISTI had previously demonstrated the power of these technologies through the development of other succesful web applications. The Model View Controller (MVC) design pattern was selected to facilitate rapid development, ease of presentation and consistent application behaviour. MVC is useful in achieving a separation of the business logic, the system control and presentation layer of the application.

The Data Access Object (DAO) pattern is used to allow abstract Enterprise Information System (EIS) independent data access.[3] The OSPREY DAO implementation was designed in the simplest form to allow for maximum speed of development and minimal knowledge to maintain. Consideration was taken to ensure the basic structure of the DAO framework would allow it to be easily extended in future and allow even greater flexibility in selecting DataSources**.**

Tomcat is used to serve dynamic servlet and JSP pages, while Apache is the web server, serving HTML pages to the user. Using Tomcat for development helped to ensure that the code would be portable and would not use proprietary packages, libraries and classes that are not otherwise available. It has the further advantage of integrating relatively seamlessly with the Apache web server and being open source.

OSPREY was designed to allow a single user to be logged into multiple journals or multiple instances of the same journal from the same HttpSession.[4] This design challenge prompted the creation of a very simple JournalSession framework. A user has one JournalSession for each authentication to an OSPREY journal. Each JournalSession is uniquely identified, holds a reference to the user's name information, and has the basic capacity to store and retrieve attributes.

Reviews and editor decisions are captured in XML and are translated into HTML or plain text format. Metadata in an XML format is exported and imported into the publishing system.

### 2.3.2 Conversion Service and Software

This service forwards requests to the appropriate servers, performs the transformation, and returns the created PDF to the user. Java Remote Method Invocation (RMI) is used to communicate between the application and the conversion service [5]. It was essential for us to support multiple files and multiple file types in one submission and create a single PDF immediately. To enable this speed and flexibility of conversion, it had to be scalable. The architecture is designed so that the software used to perform the conversions can be easily replaced.

PDF files are created by two different software packages, depending on their file format. LaTex manuscripts are converted using open source software, MikTex.[6] Adlib Express Server (third-party licensed software) processes all other file types. Adlib supports up to 300 different file formats and is upgraded frequently to meet the demand of converting newer versions of source files.[7]

### 2.3.3 Other Technologies

Manuscript data (names, addresses, manuscript data, correspondence, manuscript tracking dates) are stored in an Oracle 9i database, while all versions of the generated PDF and original submitted files are stored in a central Network Attached Storage (NAS) system.

### 2.3.4 Implementation

**Data Conversion**

To maintain review and manuscript history, we wanted to migrate as much data as possible to OSPREY. The data were analyzed, and a mapping between the Paperpath 2000 database and OSPREY was created in combination with scripts to extract, validate and import the data. The validation and testing were very time-consuming and should not be underestimated in projects of this type. A trial data conversion was completed prior to moving to production.

**The Production Environment**

As previously discussed, to fit into the current technical infrastructure of each organization, Windows 2000 was chosen as the operating system and Apache and Tomcat as the web/application server. The OSPREY interface is web based and supports leading web browsers. The file-conversion programs (Adlib and MikTeX) convert several different file types into a single PDF file for reviewing purposes which are not visible to the user.

Four servers and a central storage device support the production environment (Fig. 2) for OSPREY at CISTI. Three additional servers are in place to support development, testing and failover:

- Web/application server
- Database server
- Two conversion servers
- Network attached storage

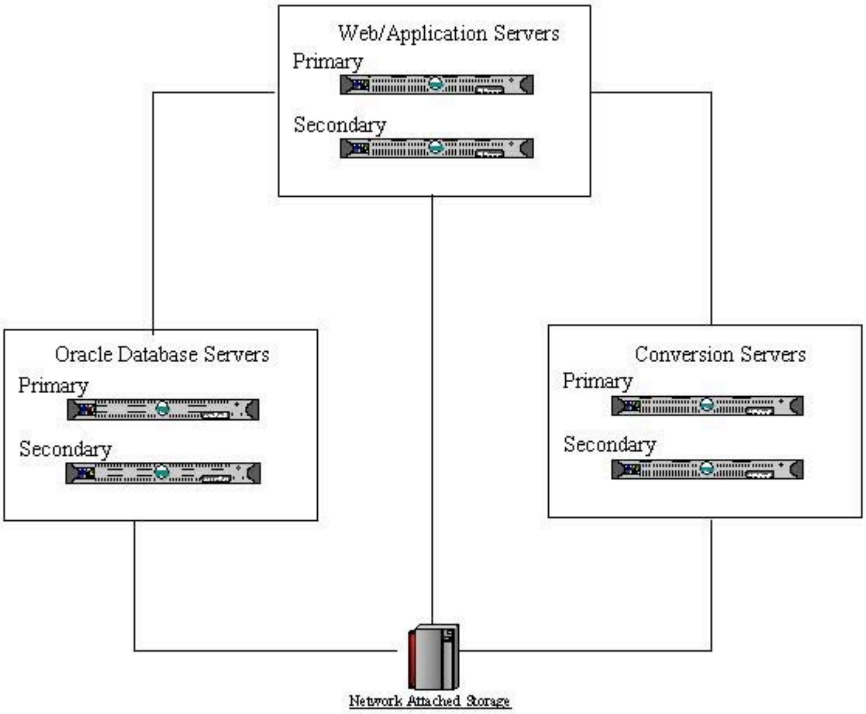


**Figure 2: NRC Research Press Production Environment**

**Integration with NRC Research Press Publishing System**

Upon acceptance of a manuscript, an XML file is created and sent to the publishing system (Fig. 3). Some processes are handled manually, while the manuscript, text, figures and supplementary data files are moved automatically to the appropriate file folders on the NAS.

```
<?xml version="1.0" encoding="UTF-8"?>
<EXPORT>
<ITEM>
<ITEM_INFO>
<EXPORT_DATE>13/09/2005</EXPORT_DATE>
</ITEM_INFO>
<MANUSCRIPT>
<TITLE>Life history variation among populations of Canadian Toads in Alberta, Canada</TITLE>
<NUMBER>4727</NUMBER>
<TYPE>Article</TYPE>
<DATE_SUBMITTED>31/08/2004</DATE_SUBMITTED>
<DATE_FINALIZED>13/09/2005</DATE_FINALIZED>
<OUTCOME>Accepted</OUTCOME>
<ABSTRACT>Development of appropriate conservation plans relies on life history information and how
life history traits vary across populations of a species. Such data are lacking for many amphibians,
including the Canadian Toad (Bufo hemiophrys Cope, 1886). Here we use skeletochronology to estimate
size-at-age, growth rates, age at maturity, and longevity of toads from nine populations along a latitudinal
gradient in Alberta, Canada. Size of individual toads within each year class was highly variable, but age and
size (measured as
```

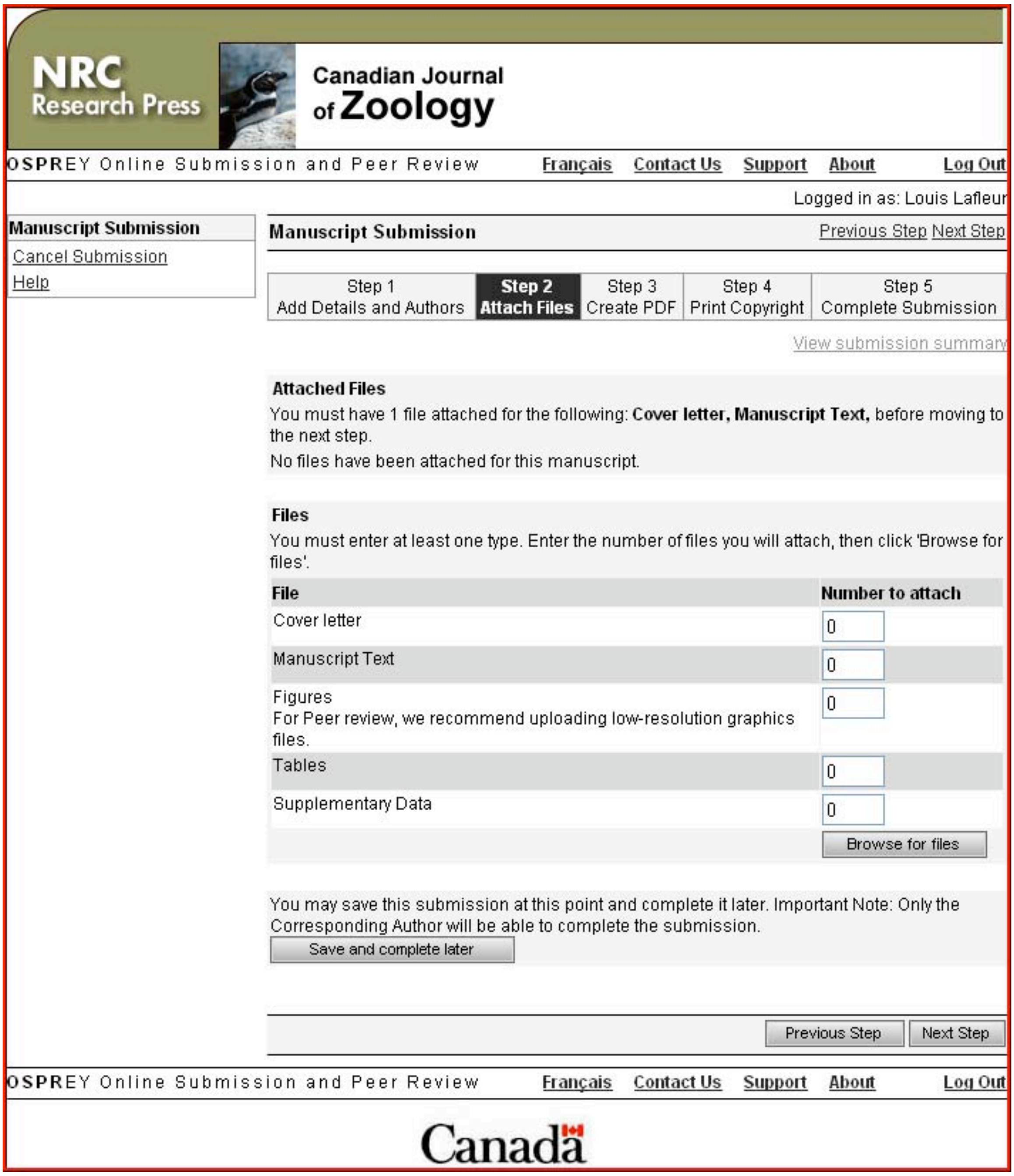
**Figure 3: Excerpt of Metadata exported from OSPREY**

**Training and Support**
Training was offered to all users prior to implementation with the exception of authors and reviewers. The length of training was dependent upon their user role. Editorial Office Assistants (who coordinate the peer review process for one journal and use the software extensively in their day-to-day work) received up to 4 days of training, while Editors and Associate Editors received 2 to 3 hours. A helpdesk is in place and is supported by one individual on a full-time basis with backup when required. The helpdesk communicates with users by phone or e-mail. The skills required include knowledge of business processes, browser functionality and PDF conversion, as well as an in-depth knowledge of the application and its configuration options and their use.

## 3 Results

OSPREY has been installed in Canada and Australia, where it is currently supporting a combined total of 32 journals. NRC Research Press began its implementation in 2004, and over several months its journals began to accept online submissions using OSPREY. OSPREY is also used by 5 client journals through a licensing agreement with NRC Research Press.

OSPREY’s interface is available in English and French and allows authors to upload manuscripts (Fig. 4), tables and figures, to create a single PDF file, and to check the status of their submissions. Upon the creation of the PDF file, authors are asked to review and approve the PDF file (Fig. 5).



Logged in as: Louis Lafleur

Manuscript Submission

**Attached Files**

You must have 1 file attached for the following: **Cover letter, Manuscript Text,** before moving to the next step.

No files have been attached for this manuscript.

**Files**

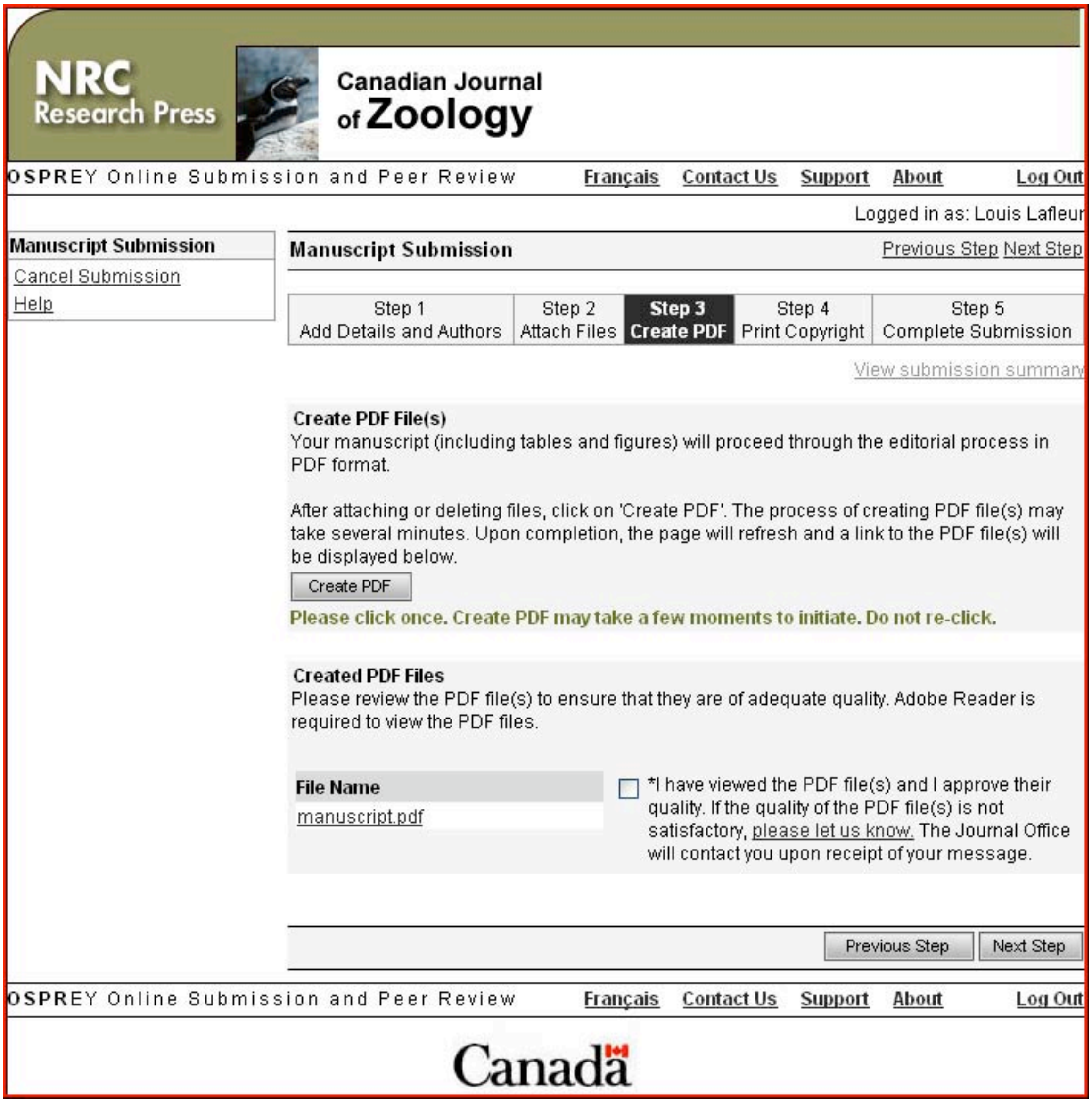
You must enter at least one type. Enter the number of files you will attach, then click 'Browse for files'.

| File | Number to attach |
|---|---|
| Cover letter | 0 |
| Manuscript Text | 0 |
| Figures<br>For Peer review, we recommend uploading low-resolution graphics files. | 0 |
| Tables | 0 |
| Supplementary Data | 0 |

You may save this submission at this point and complete it later. Important Note: Only the Corresponding Author will be able to complete the submission.

**Figure 4: Upload files**

Reviewers enter their reviews online or upload files. Editors select reviewers, vet the reviews and make recommendations or decisions on manuscripts, while editorial staff interacts with the system to verify submissions, send correspondence and assign roles and tasks.


NRC
Research Press
Canadian Journal of Zoology
OSPREY Online Submission and Peer Review
Français
Contact Us
Support
About
Log Out

Logged in as: Louis Lafleur

**Create PDF File(s)**

Your manuscript (including tables and figures) will proceed through the editorial process in PDF format.

After attaching or deleting files, click on 'Create PDF'. The process of creating PDF file(s) may take several minutes. Upon completion, the page will refresh and a link to the PDF file(s) will be displayed below.

Please click once. Create PDF may take a few moments to initiate. Do not re-click.

**Created PDF Files**

Please review the PDF file(s) to ensure that they are of adequate quality. Adobe Reader is required to view the PDF files.

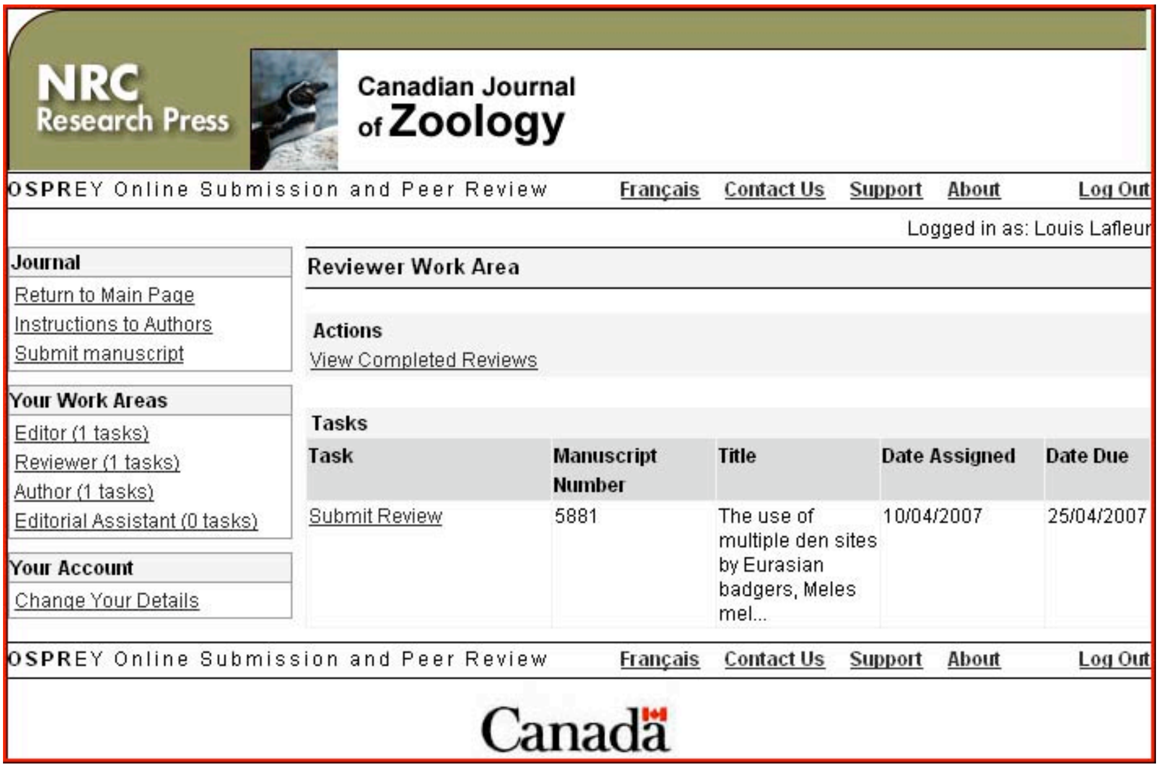
| File Name |
|---|
| manuscript.pdf |

□ *I have viewed the PDF file(s) and I approve their quality. If the quality of the PDF file(s) is not satisfactory, please let us know. The Journal Office will contact you upon receipt of your message.

**Figure 5: Create, Review and Approve PDF file**

The system is role-based, which allows journals to limit functions to appropriate users and restrict access to sensitive data. Users have a single sign-on; once logged in, they are presented with links to assigned tasks for each role [Fig. 6].

Each journal may be separately configured and many options are available including branding, workflow and selection of roles and tasks. To meet a journal's workflow requirements, each journal chooses tasks, and the order in which they occur, from a predefined list. In addition, OSPREY provides the flexibility to customize the term used to identify editorial staff roles (editor, editor-in-chief, associate editor, section editor, co-editor).



Logged in as: Louis Lafleur

| Task | Manuscript Number | Title | Date Assigned | Date Due |
|---|---|---|---|---|
| Submit Review | 5881 | The use of multiple den sites by Eurasian badgers, Meles mel... | 10/04/2007 | 25/04/2007 |

**Figure 6: Roles and Assigned Tasks**

## 3.1 Impact

Since the implementation of OSPREY, the number of submissions has increased significantly at NRC Research Press. Overall, they have risen over 32% (Fig. 7). Improved marketing and a rise in the Thompson ISI ranking of some journal titles have also contributed to the increased number of submissions. Subsequently, the total number of manuscripts accepted and rejected has also increased (Fig. 8).

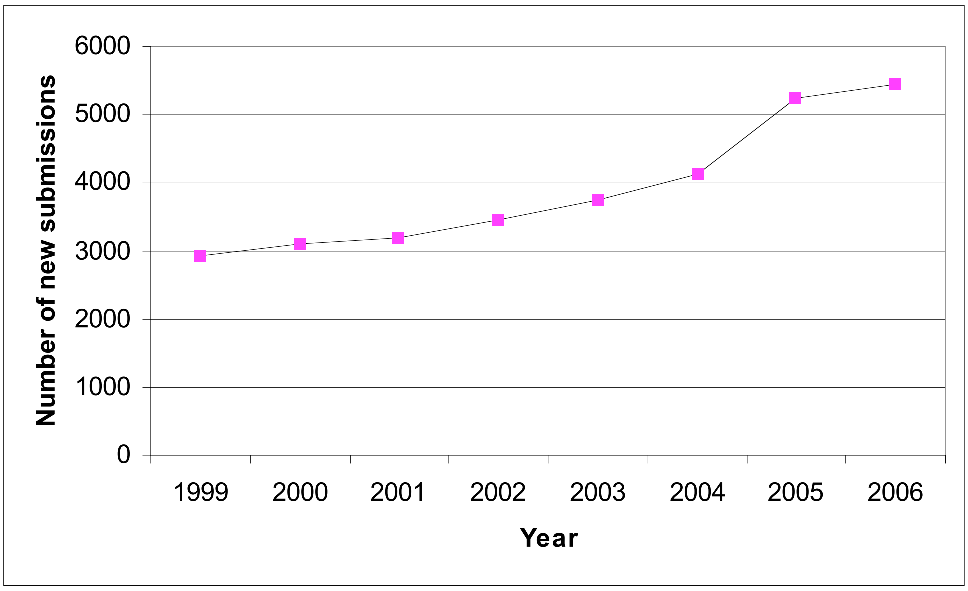


**Figure 7: Total number of manuscripts received at NRC Research Press**

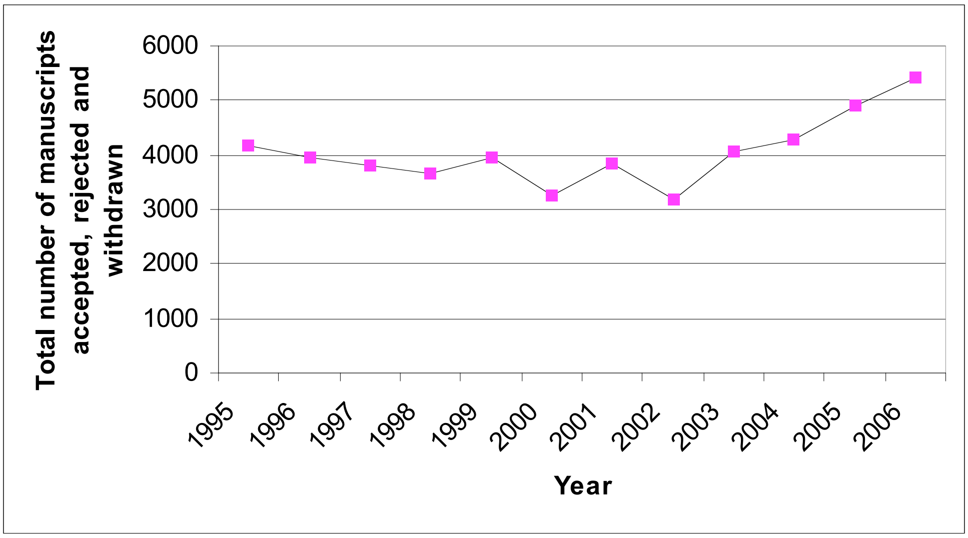


**Figure 8: Total number of manuscripts accepted, rejected and withdrawn at NRC Research Press**

Thanks to the use of an online system, there have been many changes in the way that authors, reviewers and editors work. Editorial staff now check PDF files for completeness and provide support to authors and reviewers using the system. Authors now receive an automatic acknowledgement of receipt of their submission immediately upon completion, and the paper is in the hands of the Editor faster. All correspondence is saved within the database for easy access. Prior to using an online system, acknowledgments would be sent by mail or e-mail and manuscripts would have to be delivered to the Editor by courier or expedited mail.

Reviewers are now sent an invitation to review a manuscript, as well as automatic reminders to submit their reviews. In the past, the manuscript would be sent to the reviewer by mail, courier or e-mail and a separate system would have to be in place to track the e-mails and any follow-up required. Reviewers can access manuscripts immediately and submit their reviews electronically, allowing Editors to have access to the reviews faster. Reviews and editorial decisions are tagged in XML and can be displayed in HTML or plain text format.

Editors now have access to all data required to process a manuscript from their desktop, regardless of their geographic location. They can now search for reviewers by expertise taxonomy keyword. Each journal has its own set of keywords, assigned to reviewers based on the reviewers' area of expertise. This functionality allows reviewers to be found quickly and their reviewer history, availability and performance, is available immediately.

One of the downsides of an electronic system is the learning curve. Editorial staff had to learn to work differently; instead of having stacks of mail on their desks, they now have an inbox full of e-mail. For convenience and backup, a paper record is still maintained in some cases, however; OSPREY is the primary repository for all manuscript data and correspondence.

## 3.2 Usability

One of our key design objectives was ease of use. We have received mixed reviews from users; some find the system very intuitive and others have difficulty uploading files during submission and finding links to files for download. Some users have also suggested that the number of clicks required to perform a function should be minimized. The usability issues will be addressed in the future.

## 3.3 Troubleshooting

For the architectural reasons presented previously, it was the right decision to make OSPREY distributed, accepting that the more distributed the system, the more complex the troubleshooting. OSPREY contains many components and troubleshooting can be very time-consuming. A recent investigation into a problem identified that certain types of corrupt source files could crash the conversion server. In this particular instance, the source file was not of a typical file type. Our testing procedures for conversion now include a thorough test of valid and corrupt files of all accepted file types.

# 4 Discussion

We recongized that there will be continuous maintenance and product enhancements when developing an in-house system. However; by building in-house we control the product lifecycle, features, priorities and release schedules. Maintenance includes such issues as supporting new versions of content-creation software (e.g., Microsoft Word), new web browsers, and updating the underlying software infrastructure (e.g., Apache web server), while product improvement enhances the application with new features and functions.

Online submission and peer review systems on the market today have been increasingly adding new functionality over the past few years. In other systems, editorial workflows are also customizable, parsing of references and linking to PubMed and CrossRef is now available, and some vendors are offering complete services from peer review to publication. Integrated database searching is also available in some products.[8]

## 4.1 Future Work

CISTI is currently implementing a workflow management system to manage the XML publishing process. Once this key component is in place, accepted manuscripts and metadata will flow seamlessly from OSPREY to the publishing system and will appear in the appropriate staff work areas automatically. Manuscript metadata and management data will be captured, and manuscripts will be forwarded to pre-editing. Upon publication, OSPREY will be updated with appropriate data (volume, issue, page number, date of publication).

The development and enhancement of OSPREY has opened the door to new opportunities. NRC Research Press currently has service agreements with 5 journals and plans to continue marketing OSPREY through its publishing services programme. A usability study is currently under way and will include interviews and an online survey of the user community. We will focus on ease of use and new functionality.

In addition, we are considering exploring the following functionality:

- parsing and tagging references to provide a link to abstract databases or full text (e.g. PubMed, CrossRef, and user organization link resolver)
- interfaces to allow authors to purchase paper and electronic reprints
- option for authors to identify papers for open access and supply payment or funding
- approving page proofs of accepted papers
- ability to access tasks directly from e-mail
- integration with external databases
- troubleshooting of common graphic file problems

OSPREY is a component-based system which offers us the flexibility to expand and enhance its functionality by changing components. Our current plans include replacing the existing conversion component. The next implementation will make use of web services instead of Java RMI.

# 5 Conclusion

There are many considerations when moving from a paper-based manual process to an online automated peer review and manuscript submission system or from one online system to another. When purchasing a commercial solution, or developing an in-house system, the impact on authors, reviewers and editorial staff must be considered and managed. Resources required to maintain a system are considerable, not only for software development but for upgrading hardware and software. Systems must be robust and flexible in their design to accommodate new requirements. Adequate training and user support must be put in place early in the project.

# References

[1] SCHUH, P. Integrating Agile Development in the Real World. Charles River Media, Inc of Hingham, Massachusetts, 2003.

[2] KURNIAWAN, B. How Servlet Containers Work [Online]. May 23, 2003. [Cited April 10, 2007]. Available from the World Wide Web at http://www.onjava.com/pub/a/onjava/2003/05/14/java_webserver.html

[3] Core J2EE Patterns - Data Access Object. [Online]. Sun Microsystems, Inc. [Cited April 10, 2007]. Available from the World Wide Web at http://java.sun.com/blueprints/corej2eepatterns/Patterns/DataAccessObject.html.

[4] HUNTER, J.; CRAWFORD, W. Java Servlet Programming. O'Reilly & Associates, Inc, 1998. pp. 207-231.

[5] An Overview of RMI Applications. [Online]. Sun Microsystems, Inc. [Cited April 10, 2007]. Available from the World Wide Web at http://java.sun.com/docs/books/tutorial/rmi/overview.html.

[6] Miktex 2.5. Release August, 2006. Available from the World Wide Web at http://www.miktex.org/.

[7] Adlib Express Server 3.8. Release February 15, 2007. Available from the World Wide Web at http://www.adlibsoftware.com/ExpressServer.aspx

[8] WARE, M. Online submission and peer review systems. A review of currently available systems and the experiences of authors, referees, editors and publishers. United Kingdom: Association of Learned and Professional Society Publishers, 2005.

# A Bachelor and Master Theses Portal: Specific Needs and Business Opportunities for the DoKS Repository Tool

*Rudi Baccarne*

Central Library, Katholieke Hogeschool Kempen, Geel, Belgium
e-mail: rudi.baccarne@khk.be

## Abstract

A few years ago a portal for bachelor and master theses from Flemish university colleges was established by means of the open source repository software DoKS. At present approximately 3500 theses from Flemish university colleges are available online. The growing use of the portal has led to a new communication stream that requires supervision and maintenance. Social software components amongst others are or will be integrated in the portal to give users a platform to perform tasks such as communicate, annotate and advertise. Although different local DoKS repositories and the concept of the DoKS application are similar to repositories and tools within the scientific community, the scope and the aim of a theses repository for university colleges are different. The main part of the database consists of applied research and the majority theses comprise trainee reports. Thus, in addition to students and instructors, the portal is attractive to key players in industry, non-profit institutions and private users with a particular interest in a theses subject. This paper examines the different opportunities and specific needs of a bachelor and master theses portal, illustrated by real life examples. Social software components can breath new life into former static text documents. Users can add comments, create blogs, add tables, illustrations and suchlike. Content sensitive advertisements enhance the content and usage of our theses records and create revenues that can be used to make new improvements. In addition we will discuss the need for new and strict procedures with regard to content control, copyright issues and embargos when a bulk collection of industry related theses are published online.

**Keywords:** print on demand; content sensitive advertising; Electronic Theses and Dissertations (ETD); social software

## 1 Introduction

In 2003 the library of the Katholieke Hogeschool Kempen (KHK) launched a portal [1] for electronic theses and curricula vitae of graduating students at Flemish university colleges. In order to create this portal a new software DoKS (Document and Knowledge Sharing) was built. The project is funded by the Institute for the Promotion of Innovation by Science and Technology in Flanders [2], private industry partners and non-profit organizations. One of the main reasons for developing a new software was the need for a system that could be highly customized by users and tailored to needs specific to Flemish university colleges. The need for a flexible way to add local metadata (awards, trainee posts, credit points, etc.) in addition to commonly agreed sets such as Dublin Core or ETD-MS was in particular a high priority for the different colleges that were interviewed during preparatory meetings.

The DoKS portals from different colleges are decentrally stored yet at the same time available through one interface via the OAI-PMH protocol. Apart from searching harvested metadata via the central OAI-harvester, the user can search the full text of Electronic theses and dissertations (ETDs) from different institutes by means of the Google custom search engine [3], the latter allowing specification of the websites to be searched.

The KHK portal [4] receives 1500 daily visitors and offers an almost daily feedback and thus provides an indication of its usefulness for the labour market, graduates and the industry amongst others. Although the concept is similar to ETD-repositories and tools from the scientific community, a theses repository of a university college has other target users in mind. From the start private industry partners and other organizations supported the development of the DoKS theses portal, their reasons for supporting the DoKS project varying in accordance with their core business or particular interest. (IT, recruitment, valorisation of knowledge, screening of publications on business related content, electronic publishing etc.). The business plan carried out in the framework of the DoKS project and feedback from users raised new ideas and commercial opportunities for consideration. In addition to the publication of ETDs, several new and sometimes secured services (curricula vitae, ratings, etc.) were added to the free service of rendering theses from Flemish university colleges available worldwide. The DoKS software is available by means of an open source license at Sourgeforge (http://sourceforge.net/projects/doksproject). Reports and manuals are available via a wiki (http://doks.khk.be/wiki/) and the project website (http://doks.khk.be).

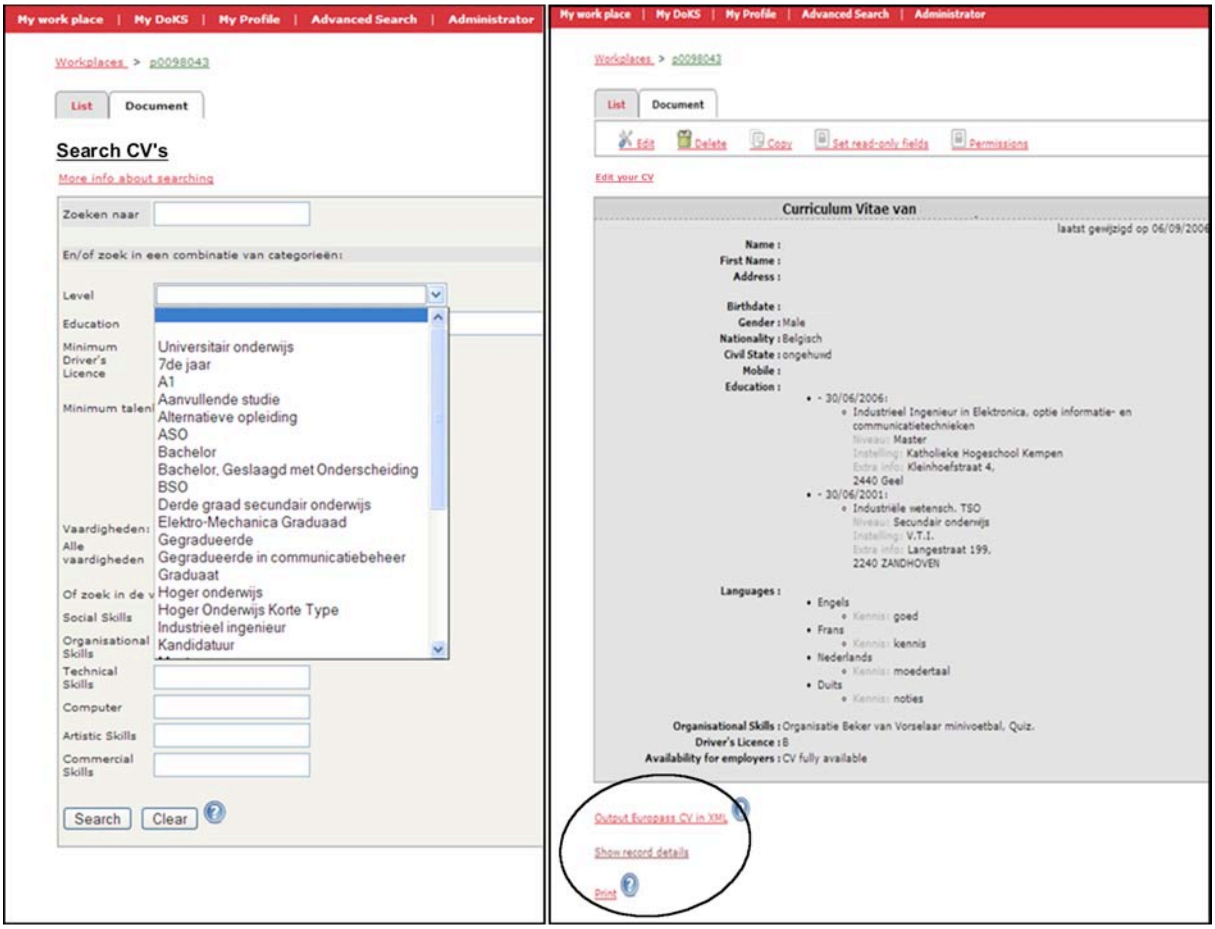
# 2 Specific Needs and Services for University Colleges

At international level, the focus of ETD-projects is on research theses from academic institutions. The main target audience for electronic theses projects is the research community. By making research theses more broadly available by means of open online repositories, researchers and research become more visible and as a consequence more widely cited. Advantages and value added services are generated for the researcher, his work and his place of work. [5, 6] Although university colleges are less focused on scientific research they show a growing interest in publishing electronic theses. From the beginning the DoKS repository tool was developed in accordance with the guidelines of the ETD community As a result the portal is interoperable and combinable to a larger extent. On the other hand, the increasing use of the site and the feedback gained has shown that a theses-portal with mainly bachelor and master theses has different needs and offers other opportunities to exploit.

## 2.1 Curricula vitae

In the framework of the DoKS project a business plan was carried out. Part of the business plan was a quantitative and a qualitative research of different target users of an ETD portal. One important conclusion drawn from the business plan and based on the findings of the user surveys was that there seemed to be considerable interest from the private sector in the use of the portal as a recruitment tool. As a consequence curricula vitae can be filed in a standardized way together with the ETD records. This renders the portal a simple and cheap alternative as a starting point in the search for new employees. The force of this system lies in the accuracy of the data it contains and specific search facilities to retrieve this data (see Figure 1). In addition it is possible to raise an alert when new CVs matching specific criteria are added. Entries to the system are made by students and/or institutes and the system is therefore more complete and unique than any other built *ad hoc* and by a third party. All companies interviewed were prepared to pay for such a service by means of a subscription or registration service.

Furthermore the business plan created for the DoKS project outlined the opportunities that could render the portal self maintaining. Once optimized and well positioned on the market the CV module should create enough revenues for the employment of a 50% employee for maintenance and administration of the system. At the moment different steps are being taken to convert this potential into reality. Different approaches to commercialize the CV module include partnerships, pay on demand, subscriptions and sponsorship to name but a few. It is clear that the graduating students also benefit from the CV module. The DoKS system automatically creates a CV for all graduating students based on data from student administration files. The student can complete this by adding extra data. The result is a CV that can be converted to a Europass CV by one mouse click. Statistics on the number of updated CVs confirm the students' interest. At the moment 1578 CV records are available belonging to KHK students who have graduated over the last two academic years. About 25 % of CV records were completed by students one month after the automatic creation of their CV and more than 70 % are updated before the end of an academic year. We believe that this enthusiasm is related tot the fact that the CVs are already formatted, half filled out and exportable to print and other formats (see Figure 1). In other words, the student does not have to make much effort to produce an attractive and easy-to-maintain CV. The idea of automatically generating many of the relevant metadata behind the scenes to avoid filling out lengthy electronic forms is relevant for different web environments. In the field of educational multimedia, for example, this is strongly expressed by Erik Duval in the slogan 'electronic forms must die' [7]. While going through the self-archiving wizard for their electronic thesis students must choose to what extent their CV will be available to employers. A majority choose the option 'CV fully available' thus allowing employers to contact the student directly. 'CV available without personal data (contact through DoKS website)' and 'CV not available' are two other options. Graduated students can access and update their CVs via a dedicated account for a limited period of two years.



Curriculum Vitae van

laatst gewijzigd op 06/09/2006

Name :
First Name :
Address :

Birthdate :
Gender : Male
Nationality : Belgisch
Civil State : ongehuwd
Mobile :
Education :

- - 30/06/2006:
  - Industrieel Ingenieur in Elektronica, optie informatie- en communicatietechnieken
    Niveau: Master
    Instelling: Katholieke Hogeschool Kempen
    Extra info: Kleinhoefstraat 4,
    2440 Geel
- - 30/06/2001:
  - Industriële wetensch. TSO
    Niveau: Secundair onderwijs
    Instelling: V.T.I.
    Extra info: Langestraat 199,
    2240 ZANDHOVEN

Languages :

- Engels
  - Kennis: goed
- Frans
  - Kennis: kennis
- Nederlands
  - Kennis: moedertaal
- Duits
  - Kennis: noties

Organisational Skills : Organisatie Beker van Vorselaar minivoetbal, Quiz.
Driver's Licence : B
Availability for employers : CV fully available

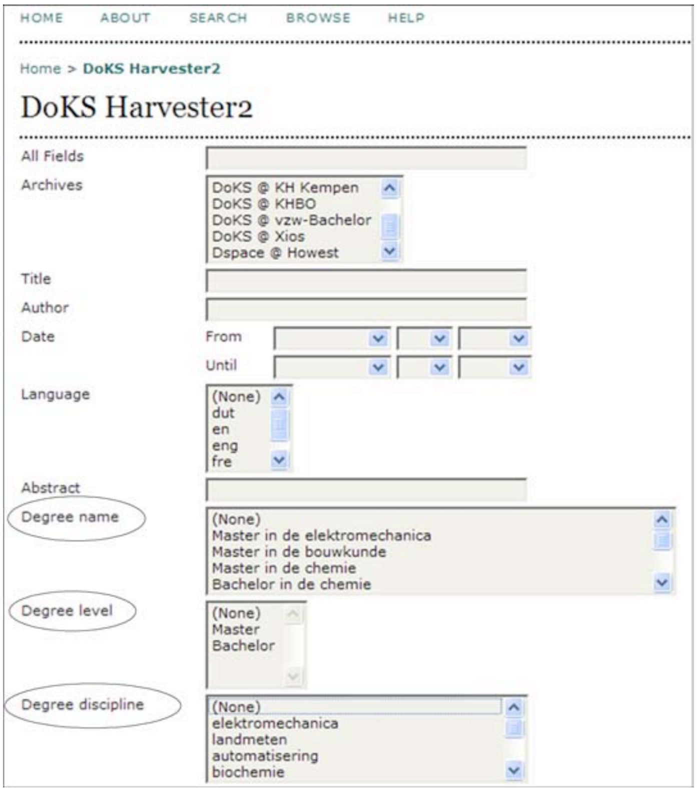
**Figure 1: Search interface for CVs and basic example CV**

In addition to the commercialization of the CV module, a good working DoKS portal can be exploited in several other ways. At the KHK the theses portal is linked to the Google AdSense program and creates revenues that are sufficient for replacement of hardware, new improvements, etc. Several other opportunities and partnerships with industry partners or non-profit organizations emerge once the portal is known to the different stakeholders. At the KHK this has led to partnerships with innovation agents, non profit partners and private industry partners that have a variety of reasons for supporting the maintenance of the portal [8, 9].

## 2.2 ETD-MS

ETD-MS is an interoperability metadata standard for electronic theses and dissertations [10]. The standard adds one element to the Dublin Core metadata elements, namely thesis.degree. This element has 4 qualifiers: thesis.degree.name, thesis.degree.level, thesis.degree.discipline, and thesis.degree.grantor. This standard is a result of the work established by the Networked Digital Library of Theses and Dissertations (NDLTD) which tries to coordinate the different worldwide initiatives. The aim of the standard is interoperability and is described as such on the NDLTD website [11] 'i.e., to make it possible to share information about ETDs. This will allow us to improve existing federated searches, create union databases, and provide greater consistency for researchers searching for theses and dissertations at different institutions.

To integrate bachelor and master theses in ETD union catalogues and repositories for scholarly communication we believe the level of education must be transparent and clearly distinguishable. This will help the end user to place the work in his context so he can judge it appropriately. The PKP-harvester software [12] we use for harvesting metadata from different local DoKS repositories did not however support ETD-MS. Therefore we recently created an ETD-MS plug-in [13] for this harvester. By using the plug-in users can perform searches on degrees and level of education to find graduates, their profiles and their learning outcomes.



DoKS Harvester2

All Fields

Archives: DoKS @ KH Kempen, DoKS @ KHBO, DoKS @ vzw-Bachelor, DoKS @ Xios, Dspace @ Howest

Title

Author

Date: From / Until

Language: (None), dut, en, eng, fre

Abstract

Degree name: (None), Master in de elektromechanica, Master in de bouwkunde, Master in de chemie, Bachelor in de chemie

Degree level: (None), Master, Bachelor

Degree discipline: (None), elektromechanica, landmeten, automatisering, biochemie

**Figure 2: ETD-MS plug-in for PKPHarvester2**

When other ETD programs consistently use the same standards, users can search records in a similar manner across records from different countries. The end user has an immediate knowledge of the type of the work, the level of the work, the educational program in which it was produced, the related degree and so on without necessarily knowing the language in which it was written.

## 2.3 Content (applied research, less academic, …)

At the KHK and by extension at most similar university colleges in Flanders nearly all theses comprise reports on work a student has done at a trainee post. As a consequence a thesis might contain confidential information. The industry partner where the trainee is placed has the option of requesting an embargo by means of strict procedures and dedicated forms. Although we expected to see a drastically increasing *a priori* demand for embargos once we started to publish ETDs, this seems not to be the case. Nevertheless the student must inform the trainee post about the online publication a clear demand for new embargos is seen once a thesis is online for a period of time. In some cases an embargo is requested because confidential information is published, but there seems to be several other reasons an industry partner does not want to see a trainee report published. This is often related to the high search engine ranking of our theses records - DoKS theses records are ranked higher than the web pages of the trainee posts - , old or false information on products is still available on the web, etc. In the future there are plans to give users a communication platform on which annotations can be made. In this way it is hoped that embargos can be avoided where the need for them goes beyond the publication of confidential information. On the other hand we see in literature [14] and also in practice a shift towards more transparency in domains (pharmaceutical industry, innovative IT companies) that were at first more resistant to the online publication of research data and material.

The power of the portal to serve the needs of innovation agents and intermediary organizations has resulted in the following collaborations:

**Flemish Chamber of Engineers (VIK)**
The Flemish Chamber of Engineers is developing an award program based on the DoKS repositories to stimulate entrepreneurship. The aim of this project is to filter theses with a high commercial or innovative character, especially those that have the potential to develop into enterprises. The idea stemmed from the fact that the annual number of new start-up businesses of an innovative nature in Belgium is very low compared with other countries [15]. Furthermore over the years the Chamber has kept files of new businesses that emerged from the basis of an innovative idea in a thesis;

**'Innovatiecentrum West-Vlaanderen' (West Flanders regional innovation centre)**
A study [16] carried out by the regional innovation centre of West Flanders pointed out that of all theses established in the context of a trainee post only a low percentage resulted in an economic surplus value for the firms involved. An analysis of the study however showed that the economic valorisation could be increased by taking measures such as recruitment of the student after graduation or extra guidance by the college. To achieve this the regional innovation centre allocates awards for students and valorisation budgets for the firms. The innovation centre urges University colleges from the region to set up a DoKS portal in order to improve and accelerate selection procedures for theses that would be considered for a valorisation trajectory;

**Indiegroup**
Indiegroup is an organization that develops software for the innovation market. Integrating innovative content from the theses of university colleges can create a surplus for their software 'Cognistreamer'. Cognistreamer is a platform for open innovation concepts. By means of RSS and XML crosswalks innovative theses projects from DoKS could be integrated and selectively disseminated via Cognistreamer to organizations that are working on related subjects.

## 2.4 Statistics

### 2.4.1 Daily Visitors

The statistics in Figure 3 are based on figures from Google Analytics and cover the last full 12 months the KHK DoKS portal was online. The extremely sudden peak on the 11th of January has a logical explanation. At that time it was noted that a majority of visitors downloaded the full text of a DoKS thesis directly via a Google result list bypassing the DoKS website. For several reasons we have now decided to use a URL rewrite mechanism so that users are always transferred to a DoKS thesis record from where they can download the full text of the document. First of all the figures for the use of our portal were seriously underestimated. First and foremost, users were downloading documents from the site without knowing they were reading a thesis document from a bachelor or master student at a Flemish university college.

In addition it is clear that there is stable use of the website which at the moment has an average of 1500 daily visitors. The trend is downwards during the weekend and holidays and use increases use during the periods the students are working on a thesis and need the portal to submit data and full text. The statistics from Figure 6 indicate that the use of the site is strongly related to the revenues created by Google Ads with the same steep increase from January 2007 onwards.

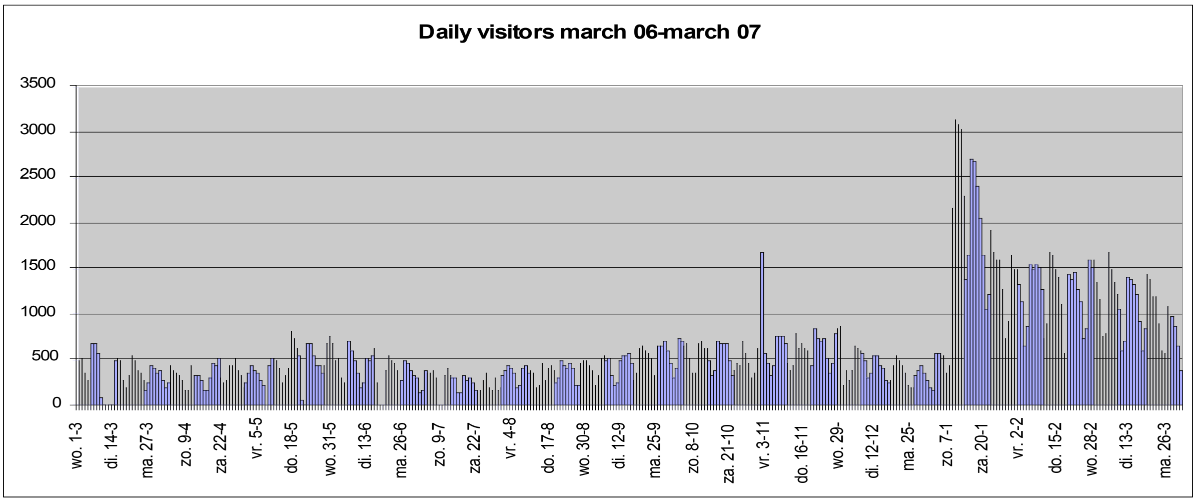


**Figure 3: Daily visitors DoKS@KHK**

### 2.4.2 Downloads

As shown in the graph below showing the number of theses downloads from the KHK-portal a ‘long tail’ curve emerges. This illustrates the wide and varied interest in theses content. The usage shown by the curve seems to be typical for e-business websites and indicates new economic mechanisms that are related to the internet. Theses that perhaps never came to light when they were stored physically at the library are, once online, consulted more than the most borrowed hardcopy theses from that same library. These new models and mechanisms are described by Chris Anderson with regard to e-business sites from the amusement industry (Amazon, Itunes, etc. ) [17].

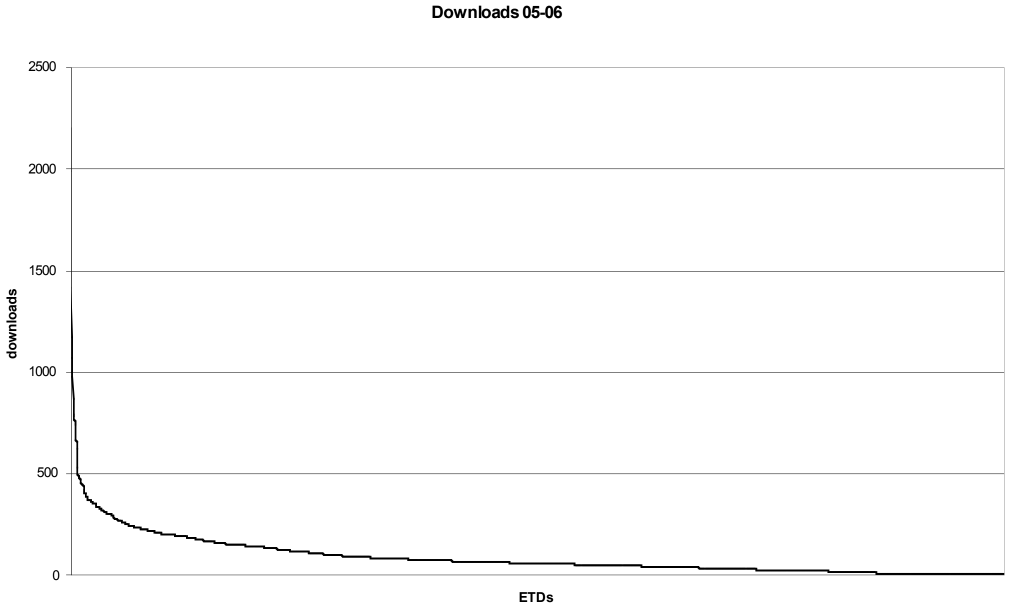


**Figure 4: KHK-ETD downloads 05-06**

## 2.5 User Feedback (categories)

When filled with a significant amount of content the DoKS repositories receive a high search engine ranking. As a result the number of daily visitors is significant. Among a variety of users, the feedback received in Flanders introduced several new business opportunities and interests. The feedback received can be categorized as follows:

- Job offers and offers for trainee posts
- Knowledge sharing
    - Collaboration proposals
    - Questions on thesis subject
    - Demand for annotations. Users want to comment on the content of a thesis and students want to add new views, opinions, corrections, etc.
- Editorial boards of journals
- Embargo requests from industry partners of the KHK
- Hardcopy requests (see section 3.3.5)
- Reporting on violations of the law with regards to:
    - professional confidentiality
    - copyright
    - privacy

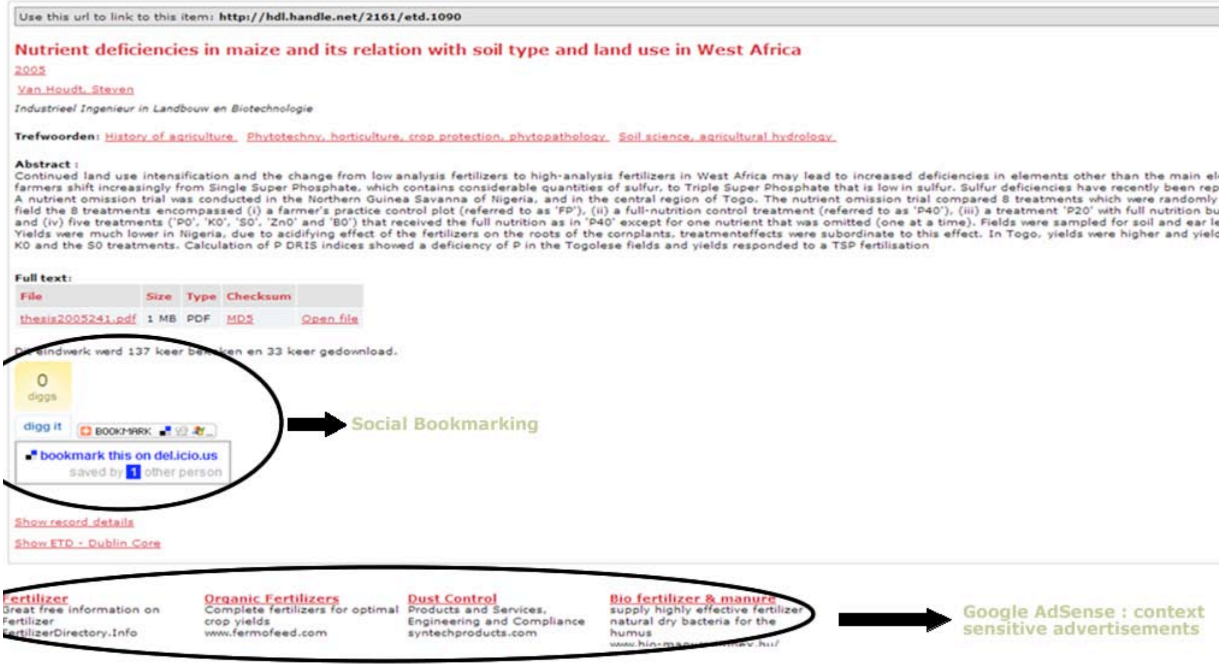
## 2.6 Social Networking and Business Opportunities

The use and feedback on the portal clearly indicated the need to add social networking tools. Plans are being made for the future to integrate the features of the KNOSOS [18] platform in DoKS. Users who want to add, blog, annotate or tag to name but a few will be allocated to the collaborative working space provided by

KNOSOS. In preparation of a structured approach to overcome different needs, the first experiments have been set up. The following paragraphs describe the way in which we have already addressed some user demands.

### 2.6.1 A New Splash Page

Students are nowadays familiar with new technologies (Internet, multimedia, publishing, web 2.0, etc.) but are not supported in the use of them in a traditional hardcopy print environment. By following new internet trends, DoKS is able to keep track of the way young undergraduate students use the internet. We believe this is a necessary condition to conserve the enthusiasm of our most important supplier of information, the students themselves. As a result the record splash page (Figure 5) has been drastically changed in favour of a more user-friendly interface.



Use this url to link to this item: **http://hdl.handle.net/2161/etd.1090**

**Nutrient deficiencies in maize and its relation with soil type and land use in West Africa**

2005

Van Houdt, Steven

*Industrieel Ingenieur in Landbouw en Biotechnologie*

**Trefwoorden:** History of agriculture, Phytotechny, horticulture, crop protection, phytopathology, Soil science, agricultural hydrology,

**Abstract :**
Continued land use intensification and the change from low analysis fertilizers to high-analysis fertilizers in West Africa may lead to increased deficiencies in elements other than the main el
farmers shift increasingly from Single Super Phosphate, which contains considerable quantities of sulfur, to Triple Super Phosphate that is low in sulfur. Sulfur deficiencies have recently been rep
A nutrient omission trial was conducted in the Northern Guinea Savanna of Nigeria, and in the central region of Togo. The nutrient omission trial compared 8 treatments which were randomly
field the 8 treatments encompassed (i) a farmer's practice control plot (referred to as 'FP'), (ii) a full-nutrition control treatment (referred to as 'P40'), (iii) a treatment 'P20' with full nutrition bu
and (iv) five treatments ('P0', 'K0', 'S0', 'Zn0' and 'B0') that received the full nutrition as in 'P40' except for one nutrient that was omitted (one at a time). Fields were sampled for soil and ear le
Yields were much lower in Nigeria, due to acidifying effect of the fertilizers on the roots of the cornplants, treatmenteffects were subordinate to this effect. In Togo, yields were higher and yielc
K0 and the S0 treatments. Calculation of P DRIS indices showed a deficiency of P in the Togolese fields and yields responded to a TSP fertilisation

**Full text:**

| File | Size | Type | Checksum | |
|---|---|---|---|---|
| thesis2005241.pdf | 1 MB | PDF | MD5 | Open file |

Dit eindwerk werd 137 keer bekeken en 33 keer gedownload.

Show record details

Show ETD - Dublin Core

**Fertilizer**
Great free information on Fertilizer
FertilizerDirectory.Info

**Organic Fertilizers**
Complete fertilizers for optimal crop yields
www.fermofeed.com

**Dust Control**
Products and Services, Engineering and Compliance
syntechproducts.com

**Bio fertilizer & manure**
supply highly effective fertilizer natural dry bacteria for the humus
www.bio-[illegible].hu/

**Figure 5: A thesis record splash page**

In the following paragraphs the benefits of the major adjustments, namely the integration of social bookmarking tools and context sensitive advertisements will be discussed.

### 2.6.2 Social Bookmarking

In the light of our current subject classification which is deemed insufficient [19], a new opportunity is presented by the use of folksonomies or a tagging system. Furthermore, an interactive way of supplying keywords or tags perfectly matches the broader aims of the DoKS project, namely, knowledge sharing and community building. At the moment social bookmarking tools are provided on the theses records. By means of the 'Delicious Tagometer' (see Figure 5) it is easy to find out which other people have tagged a particular thesis record. This will lead you to the bookmark pages of people with common interests. It also allows you to see how what resources other users have tagged on the same subject. A desired feature that until now has not been available is a way of aggregating tags from different users of our portal in a tag cloud. Once such a feature is available we can provide this aggregation of tags to our users.

### 2.6.3 Instant Messaging

Although a part of the metadata (author, department, degree title, address, etc.) is automatically imported from files received by the library from the Institute's general administration department, another part (title, abstract, language, volumes, contact details, number of desired copies, instructor, trainee post, trainee supervisor) must be submitted by the students via the DoKS repository software. In addition the full text must be submitted by a self-archiving approach. Given that on an annual basis as many as 800 students submit their thesis data, they have to follow strict procedures and guidelines. By giving students the opportunity to ask questions whilst submitting data, the administrators can assist students directly should they encounter problems. In DoKS this communication is provided by a Meebo [20] widget and is similar to Instant Messaging systems the use of which is very familiar to the students. Another attractive feature that comes with the integration of the Meebo widget is the ability to keep track of the number of concurrent users of your site.

### 2.6.4 Google AdSense

Once installed and filled with content a DoKS repository creates revenues via Google Adsense that can cover maintenance costs (upgrading hardware backup tapes) and/or new improvements. This is already the case for the repository of the Katholieke Hogeschool Kempen. The idea to start experimenting with Google Ads was based on a very practical mail question received from a user. The user had downloaded a thesis about laying out a private swimming pond. In the thesis prices for products were given which appeared to be much lower than the ones experienced by the user. Instead of contacting the student behind the particular theses we thought it might be easier to provide a direct path to suppliers of goods related to a thesis subject. Via Google AdSense relevant context sensitive ads are displayed on the pages containing theses records. The ads are related to the visitors' search and thus create a way to both monetize and enhance the theses records. We have currently been experimenting for almost a year with the system and evaluated that both benefits seems to be fulfilled. The ads are in most cases relevant and enhance our content.

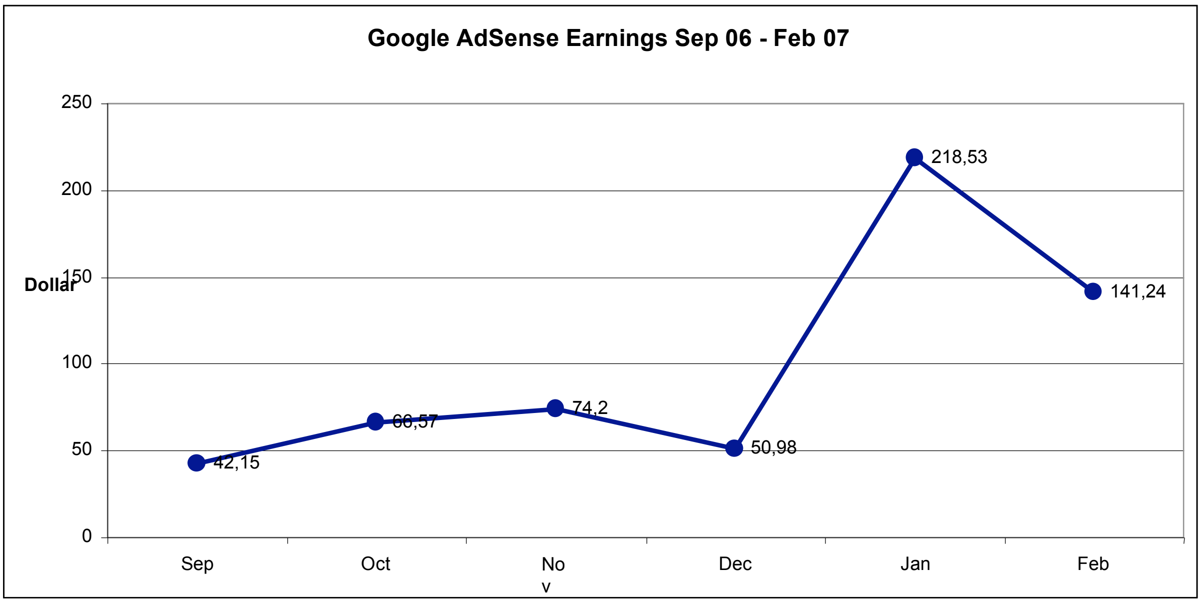
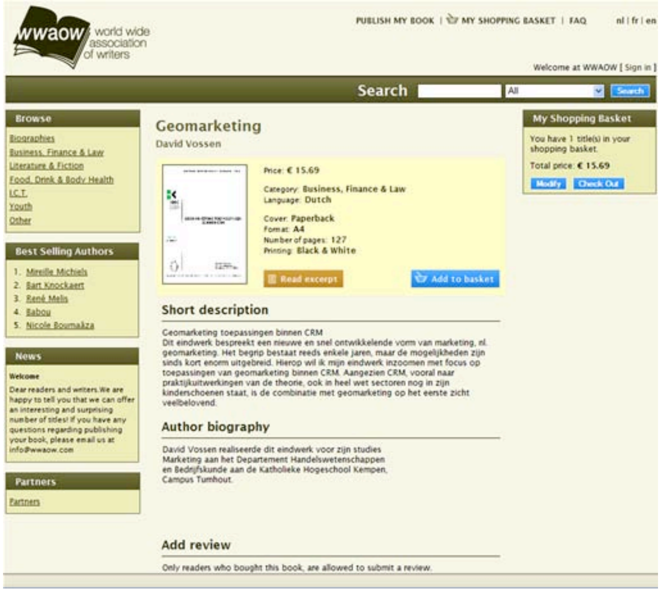


**Figure 6: Google AdSense earnings Sep 06 – Feb 07**

### 2.6.5 Print On Demand

With great surprise we noted a significant demand from our users to obtain a hardcopy version of the theses we published electronically. At first the intention was to deny these requests because it was thought that they would occur very occasionally and there were not the resources to give an appropriate answer. However more requests for hardcopies arose and by coincidence the DoKS portal caught the attention of a new player on the market of print on demand and self publishing, i.e. WWAOW (world wide association of writers) [21] This resulted very recently in a new collaboration and the first theses from different university colleges in Flanders are available via WWAOW (see Figure 7). Students are informed about this opportunity during the self-archiving procedure where they can choose whether they want to make their thesis available by means of the WWAOW website.



1. Mireille Michiels
2. Bart Knockaert
3. René Melis
4. Babou
5. Nicole Boumaâza

News

Welcome

Dear readers and writers.We are happy to tell you that we can offer an interesting and surprising number of titles! If you have any questions regarding publishing your book, please email us at info@wwaow.com

Geomarketing

David Vossen

Price: € 15.69

Category: Business, Finance & Law
Language: Dutch

Cover: Paperback
Format: A4
Number of pages: 127
Printing: Black & White

My Shopping Basket

You have 1 title(s) in your shopping basket.

Total price: € 15.69

Short description

Geomarketing toepassingen binnen CRM
Dit eindwerk bespreekt een nieuwe en snel ontwikkelende vorm van marketing, nl. geomarketing. Het begrip bestaat reeds enkele jaren, maar de mogelijkheden zijn sinds kort enorm uitgebreid. Hierop wil ik mijn eindwerk inzoomen met focus op toepassingen van geomarketing binnen CRM. Aangezien CRM, vooral naar praktijkuitwerkingen van de theorie, ook in heel wat sectoren nog in zijn kinderschoenen staat, is de combinatie met geomarketing op het eerste zicht veelbelovend.

Author biography

David Vossen realiseerde dit eindwerk voor zijn studies Marketing aan het Departement Handelswetenschappen en Bedrijfskunde aan de Katholieke Hogeschool Kempen, Campus Turnhout.

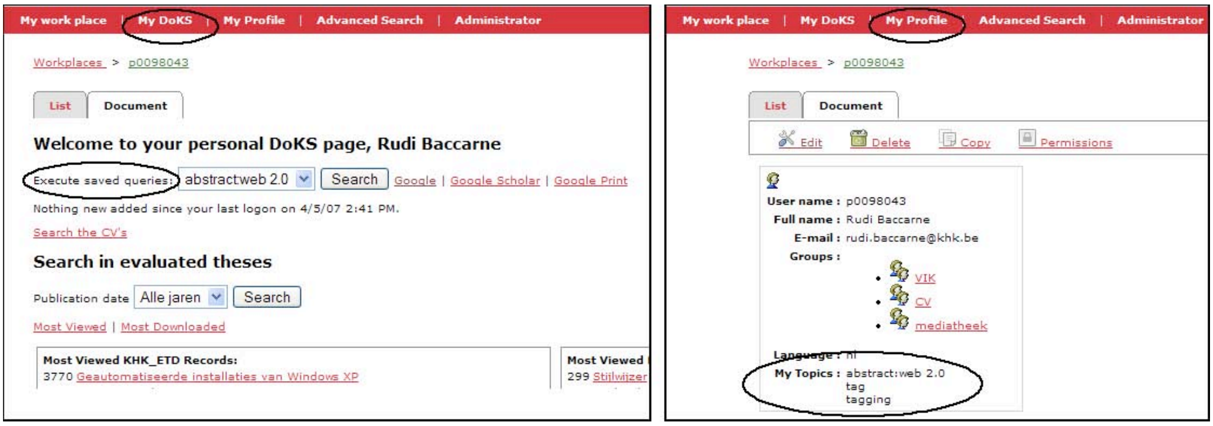
Add review

Only readers who bought this book, are allowed to submit a review.

**Figure 7: Print on demand via WWAOW**

### 2.6.6 Alerting via Persistent Query Mechanisms and RSS

All authenticated users have a personal profile page in which they can store keywords (My Topics). By means of a persistent query mechanism search queries can be saved and can be executed again. This technique is used to provide an alert system. Once logged in a personal homepage is displayed (MyDoKS). On this homepage there appears a list showing documents which are new since the last time the user logged in.


My work place | My DoKS | My Profile | Advanced Search | Administrator
Workplaces > p0098043
List
Document
Search
Google | Google Scholar | Google Print
Search the CV's
Search
Most Viewed | Most Downloaded
Edit
Delete
Copy
Permissions
VIK
CV
mediatheek

Welcome to your personal DoKS page, Rudi Baccarne

Execute saved queries: abstract:web 2.0

Nothing new added since your last logon on 4/5/07 2:41 PM.

Search in evaluated theses

Publication date Alle jaren

Most Viewed KHK_ETD Records:
3770 Geautomatiseerde installaties van Windows XP

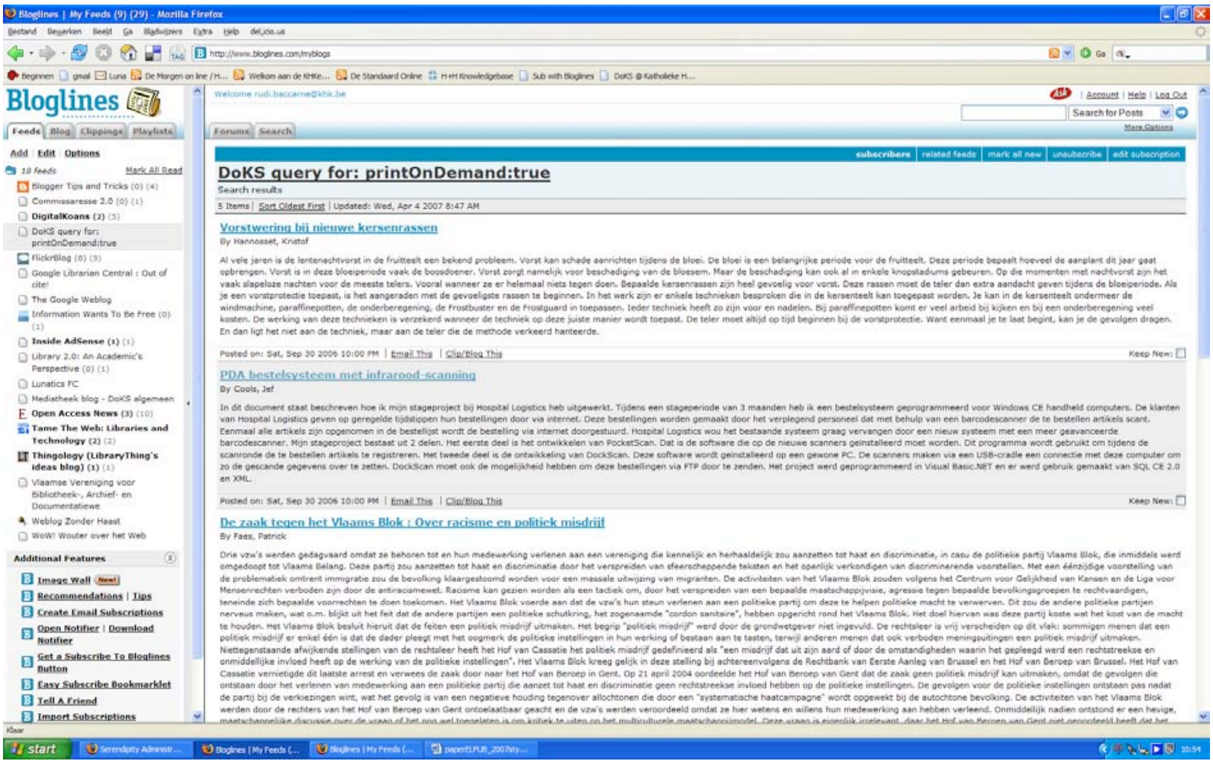
Most Viewed
299 Stijlwijzer

User name : p0098043
Full name : Rudi Baccarne
E-mail : rudi.baccarne@khk.be
Groups :
- VIK
- CV
- mediatheek

Language : nl
My Topics : abstract:web 2.0
tag
tagging

**Figure 8: DoKS personal homepage and profile page**

The built in RSS functionality can be used as an alert system as well. It is possible for example to subscribe to search queries via RSS, with the result that whenever a new item is published that reflects your query you will be alerted by you RSS reader. This RSS functionality is also a means to publish automatically updated lists. For example you can subscribe to a list of theses that are available by the print on demand system (see Figure 9), and new CVs to name but a few.


Bloglines | My Feeds (9) (29) - Mozilla Firefox
Bestand Bewerken Beeld Ga Bladwijzers Extra Help del.icio.us
http://www.bloglines.com/myblogs
Beginnen gmail Luna De Morgen on line / H... Welkom aan de KHKe... De Standaard Online H+H Knowledgebase Sub with Bloglines DoKS @ Katholieke H...
Bloglines
Feeds Blog Clippings Playlists
Add Edit Options
18 feeds Mark All Read
Blogger Tips and Tricks (0) (4)
Commissaresse 2.0 (0) (1)
DigitalKoans (2) (3)
DoKS query for: printOnDemand:true
FlickrBlog (0) (3)
Google Librarian Central : Out of cite!
The Google Weblog
Information Wants To Be Free (0) (1)
Inside AdSense (1) (1)
Library 2.0: An Academic's Perspective (0) (1)
Lunatics FC
Mediatheek blog - DoKS algemeen
Open Access News (3) (10)
Tame The Web: Libraries and Technology (2) (2)
Thingology (LibraryThing's ideas blog) (1) (1)
Vlaamse Vereniging voor Bibliotheek-, Archief- en Documentatiewe
Weblog Zonder Haast
WoW! Wouter over het Web
Additional Features
Image Wall New!
Recommendations | Tips
Create Email Subscriptions
Open Notifier | Download Notifier
Get a Subscribe To Bloglines Button
Easy Subscribe Bookmarklet
Tell A Friend
Import Subscriptions
Forums Search
Welcome rudi.baccarne@khk.be
Account | Help | Log Out
Search for Posts
More Options
subscribers | related feeds | mark all new | unsubscribe | edit subscription
Keep New: □
Klaar
start Serendipity Administr... Bloglines | My Feeds (... Bloglines | My Feeds (... paperELPUB_2007My...
10:54

DoKS query for: printOnDemand:true

Search results

5 Items | Sort Oldest First | Updated: Wed, Apr 4 2007 8:47 AM

Vorstwering bij nieuwe kersenrassen

By Hannosset, Kristof

Al vele jaren is de lentenachtvorst in de fruitteelt een bekend probleem. Vorst kan schade aanrichten tijdens de bloei. De bloei is een belangrijke periode voor de fruitteelt. Deze periode bepaalt hoeveel de aanplant dit jaar gaat opbrengen. Vorst is in deze bloeiperiode vaak de boosdoener. Vorst zorgt namelijk voor beschadiging van de bloesem. Maar de beschadiging kan ook al in enkele knopstadiums gebeuren. Op die momenten met nachtvorst zijn het vaak slapeloze nachten voor de meeste telers. Vooral wanneer ze er helemaal niets tegen doen. Bepaalde kersenrassen zijn heel gevoelig voor vorst. Deze rassen moet de teler dan extra aandacht geven tijdens de bloeiperiode. Als je een vorstprotectie toepast, is het aangeraden met de gevoeligste rassen te beginnen. In het werk zijn er enkele technieken besproken die in de kersenteelt kan toegepast worden. Je kan in de kersenteelt ondermeer de windmachine, paraffinepotten, de onderberegening, de Frostbuster en de Frostguard in toepassen. Ieder techniek heeft zo zijn voor en nadelen. Bij paraffinepotten komt er veel arbeid bij kijken en bij een onderberegening veel kosten. De werking van deze technieken is verzekerd wanneer de techniek op deze juiste manier wordt toepast. De teler moet altijd op tijd beginnen bij de vorstprotectie. Want eenmaal je te laat begint, kan je de gevolgen dragen. En dan ligt het niet aan de techniek, maar aan de teler die de methode verkeerd hanteerde.

Posted on: Sat, Sep 30 2006 10:00 PM | Email This | Clip/Blog This

PDA bestelsysteem met infrarood-scanning

By Cools, Jef

In dit document staat beschreven hoe ik mijn stageproject bij Hospital Logistics heb uitgewerkt. Tijdens een stageperiode van 3 maanden heb ik een bestelsysteem geprogrammeerd voor Windows CE handheld computers. De klanten van Hospital Logistics geven op geregelde tijdstippen hun bestellingen door via internet. Deze bestellingen worden gemaakt door het verplegend personeel dat met behulp van een barcodescanner de te bestellen artikels scant. Eenmaal alle artikels zijn opgenomen in de bestellijst wordt de bestelling via internet doorgestuurd. Hospital Logistics wou het bestaande systeem graag vervangen door een nieuw systeem met een meer geavanceerde barcodescanner. Mijn stageproject bestaat uit 2 delen. Het eerste deel is het ontwikkelen van PocketScan. Dat is de software die op de nieuwe scanners geinstalleerd moet worden. Dit programma wordt gebruikt om tijdens de scanronde de te bestellen artikels te registreren. Het tweede deel is de ontwikkeling van DockScan. Deze software wordt geinstalleerd op een gewone PC. De scanners maken via een USB-cradle een connectie met deze computer om zo de gescande gegevens over te zetten. DockScan moet ook de mogelijkheid hebben om deze bestellingen via FTP door te zenden. Het project werd geprogrammeerd in Visual Basic.NET en er werd gebruik gemaakt van SQL CE 2.0 en XML.

Posted on: Sat, Sep 30 2006 10:00 PM | Email This | Clip/Blog This

De zaak tegen het Vlaams Blok : Over racisme en politiek misdrijf

By Faes, Patrick

Drie vzw's werden gedagvaard omdat ze behoren tot en hun medewerking verlenen aan een vereniging die kennelijk en herhaaldelijk zou aanzetten tot haat en discriminatie, in casu de politieke partij Vlaams Blok, die inmiddels werd omgedoopt tot Vlaams Belang. Deze partij zou aanzetten tot haat en discriminatie door het verspreiden van sfeerscheppende teksten en het openlijk verkondigen van discriminerende voorstellen. Met een éénzijdige voorstelling van de problematiek omtrent immigratie zou de bevolking klaargestoomd worden voor een massale uitwijzing van migranten. De activiteiten van het Vlaams Blok zouden volgens het Centrum voor Gelijkheid van Kansen en de Liga voor Mensenrechten verboden zijn door de antiracismewet. Racisme kan gezien worden als een tactiek om, door het verspreiden van een bepaalde maatschappijvisie, agressie tegen bepaalde bevolkingsgroepen te rechtvaardigen, teneinde zich bepaalde voorrechten te doen toekomen. Het Vlaams Blok voerde aan dat de vzw's hun steun verlenen aan een politieke partij om deze te helpen politieke macht te verwerven. Dit zou de andere politieke partijen nerveus maken, wat o.m. blijkt uit het feit dat de andere partijen een politieke schutkring, het zogenaamde "cordon sanitaire", hebben opgericht rond het Vlaams Blok. Het doel hiervan was deze partij koste wat het kost van de macht te houden. Het Vlaams Blok besluit hieruit dat de feiten een politiek misdrijf uitmaken. Het begrip "politiek misdrijf" werd door de grondwetgever niet ingevuld. De rechtsleer is vrij verscheiden op dit vlak: sommigen menen dat een politiek misdrijf er enkel één is dat de dader pleegt met het oogmerk de politieke instellingen in hun werking of bestaan aan te tasten, terwijl anderen menen dat ook verboden meningsuitingen een politiek misdrijf uitmaken. Niettegenstaande afwijkende stellingen van de rechtsleer heeft het Hof van Cassatie het politiek misdrijf gedefinieerd als "een misdrijf dat uit zijn aard of door de omstandigheden waarin het gepleegd werd een rechtstreekse en onmiddellijke invloed heeft op de werking van de politieke instellingen". Het Vlaams Blok kreeg gelijk in deze stelling bij achtereenvolgens de Rechtbank van Eerste Aanleg van Brussel en het Hof van Beroep van Brussel. Het Hof van Cassatie vernietigde dit laatste arrest en verwees de zaak door naar het Hof van Beroep in Gent. Op 21 april 2004 oordeelde het Hof van Beroep van Gent dat de zaak geen politiek misdrijf kan uitmaken, omdat de gevolgen die ontstaan door het verlenen van medewerking aan een politieke partij die aanzet tot haat en discriminatie geen rechtstreekse invloed hebben op de politieke instellingen. De gevolgen voor de politieke instellingen ontstaan pas nadat de partij bij de verkiezingen wint, wat het gevolg is van een negatieve houding tegenover allochtonen die door een "systematische haatcampagne" wordt opgewekt bij de autochtone bevolking. De activiteiten van het Vlaams Blok werden door de rechters van het Hof van Beroep van Gent ontoelaatbaar geacht en de vzw's werden veroordeeld omdat ze hier wetens en willens hun medewerking aan hebben verleend. Onmiddellijk nadien ontstond er een hevige, maatschappelijke discussie over de vraag of het nog wel toegelaten is om kritiek te uiten op het multiculturele maatschappijmodel. Deze vraag is eigenlijk irrelevant, daar het Hof van Beroep van Gent niet geoordeeld heeft dat het

**Figure 9: Example of a DoKS RSS search query subscription via Bloglines**

# 4 Conclusion

Apart from being more visible to the scholarly community as well as to the labour market, students and university colleges will profit in the long term by contributing to the portal in several ways. The submitted work will have to meet certain conditions before being accepted for publication. Students learn about digital publishing and structured authoring. They have to deal with choices between different file and image formats, reducing file size and structured authoring, to name but a few. The wider availability of the work creates a mentality change amongst students and lecturers towards different phases in the electronic publication chain (citing, references, copyright, technical implications of electronic publishing, etc.). In the long term this will lead to better quality in electronic documents whilst at the same time students with a rather resistant attitude to computers, internet and the like will be introduced to the internet and electronic publishing.

When filled with a significant amount of content the DoKS repositories receive a high search engine ranking. As a result the number of daily visitors is significant. Among a variety of users, the feedback received at the KHK, has introduced several new business opportunities and interests. In this sense a successful repository can be seen as a powerful public relations tool. Because DoKS supplies a java and JavaScript-like scripting engine (Beanshell) for task automation, complex work flows, specialized import/export, etc. formerly manual processes such as MARC-export for the library, collection of abstracts and titles, publishing and so on are automated. Furthermore at the KHK DoKS is used to support services such as the employment agency and the research department. As a result the tool is highly appreciated by the users of the institution. Other benefits of integrating student scholarship in institutional repositories are discussed in many blogs and publications. A collection of similar and other arguments from which the quotation below is extracted is listed in the 'Law Librarian Blog' from Carol A. Parker [22]. 'The students' scholarship would attain visibility on a scale never before seen, and the students would enjoy the benefit of informing the subsequent work of others. Plagiarism should not be an issue because most institutional repositories are indexed at the full-text level, meaning that a simple Google search would quickly identify an existing paper that was later used without proper attribution. …

Digital collections of student work can also be used for publicity and outreach, especially with alumni. Many schools already inform alumni of recent faculty publications; alumni could also be informed of student scholarship published in repositories. Making student scholarship available in digital collections provides students with a connection to their schools after graduation.

Law schools would also be sending the message that they take student scholarship seriously. Knowing that their work will also be subject to scrutiny beyond the four walls of their professors' offices would give law students added incentive to produce better scholarship.'

## References

[1] ETD portal for Flemish University colleges. See http://www.doks.be

[2] http://www.iwt.be

[3] See http://www.google.com/coop/

[4] http://doks2.khk.be/eindwerk

[5] The Open Citation Project - Reference Linking and Citation Analysis for Open Archives. *The effect of open access and downloads ('hits') on citation impact: a bibliography of studies*. Retrieved 5 April 2007 http://opcit.eprints.org/oacitation-biblio.html

[6] PIWOWAR, H. A.; DAY, R. S.; FRIDSMA, D. B., (2007). *Sharing Detailed Research Data Is Associated with Increased Citation Rate*, PLoS ONE, March 21, 2007 Retrieved 5 April 2007 http://www.plosone.org/article/fetchArticle.action?articleURI=info:doi/10.1371/journal.pone.0000308

[7] DUVAL, E. (2004). We're on the road to…, in ED-MEDIA 2004. Lugano, Switzerland. Retrieved 6 april 2007 http://www.cs.kuleuven.ac.be/~hmdb/publications/files/pdfversion/41316.pdf

[8] For an overview of DoKS partners see: http://www.doks.be/partners.htm

[9] BACCARNE, R (2005). *From central administration of hardcopy Bachelor- and Master theses towards a decentralized ETD-system with value added services* / Conference Paper, ETD2005, the 8th International Electronic Theses and Dissertations Symposium. The Scientia, University of New South Wales, Sydney, Australia, 28-30 september 2005. Retrieved 5 April 2007 http://adt.caul.edu.au/etd2005/papers/045Baccarne.pdf

[10] For a full description of ETD-MS see: http://www.ndltd.org/standards/metadata/current.html

[11] http://www.ndltd.org

[12] http://pkp.sfu.ca/harvester_download

[13] The plug-in can be downloaded at: http://doks.khk.be/do/record/Get?dispatch=view&recordId=SDoc413ebf1711bbc5e40111bc0f15560001

[14] TAPSCOTT, D; WILLIAMS, A. D. , *The new Science of Sharing* In: Business Week, March 2, 2007 Retrieved 29 march 2007 http://www.businessweek.com/innovate/content/mar2007/id20070302_219704.htm

[15] DE CLERCQ, D.; MANIGART, S.; CLARYSSE, B.; CRIJNS, H.; DE SUTTER, M.; VERZELE, F. *Global Entrepreneurship Monitor: Executive report for Belgium and Wallonia, Vlerick Leuven Gent Management School*, 2003, 95 p. Retrieved 5 april, 2007 http://www.gemconsortium.org/document.asp?id=265

[16] VANNESTE, P.; BLOMME, E.; DESAEGER, A.; GRYMONPREZ, P. (red.), *Kennisvalorisatie als opstap naar innovatiebij KMO's en kleine non-profit organisaties*, VZW Kortrijks Ondernemerscentrum, Kortrijk, 2006, 53p.

[17] ANDERSON, C. *The Long Tail: Why the Future of Business is Selling Less of More*. New York : Hyperion, 2006.

[18] http://www.knosos.be

[19] For an essay on the insufficiency of traditional categorization methods for the electronic world, See: Clay Shirky. *Ontology is overrated: Categories, links, and tags*, 2005. Retrieved 10 April 2007 http://www.shirky.com/writings/ontology overrated.html.

[20] http://www.meebome.com/

[21] http://www.wwaow.com

[22] PARKER, C. A.[blog] http://lawprofessors.typepad.com/law_librarian_blog/2007/02/institutional_r.html

# The FAO Open Archive: Enhancing Access to FAO Publications Using International Standards and Exchange Protocols

*Claudia Nicolai; Imma Subirats; Stephen Katz*

Food and Agriculture Organization of the United Nations
Viale delle Terme di Caracalla 1, 00153 Rome, Italy
e-mail: Claudia.Nicolai@fao.org; Imma.Subirats@fao.org; Stephen.Katz@fao.org

## Abstract

Since 1998, the Food and Agriculture Organization of the United Nations (FAO) has been publishing its electronic publications in the FAO Corporate Document Repository (CDR). The electronic publishing workflow is maintained by the Electronic Information Management System (EIMS). The EIMS-CDR holds more than 38 500 documents and is the gateway to FAO's publications. The EIMS-CDR coexists with the FAODOC – the online catalogue for documents produced by FAO. FAODOC catalogues and indexes both electronic and printed documents while the EIMS-CDR manages full text documents and a minimal set of metadata. This paper discusses the merger of the EIMS-CDR and the FAODOC into a unique FAO Open Archive based on the integration of the electronic publishing and the bibliographic cataloguing requirements. The FAO Open Archive will be the foundation for the collection, management, maintenance and timely dissemination of material published by FAO. To improve the effectiveness of the proposed repository, it is necessary to streamline the current electronic publishing workflow. The merger of the EIMS-CDR and the FAODOC will strengthen FAO's role as a knowledge dissemination organization. Especially, as one of the principal tasks of the FAO is to efficiently collect and disseminate information regarding food, nutrition, agriculture, fisheries and forestry.

**Keywords:** open access; open archive initiative; interoperability; digital repositories; data content standards

## 1 Introduction

The Food and Agriculture Organization of the United Nations (FAO) has more than 50 years of experience in the production and the dissemination of information, both through its headquarters-based regular programme and through field projects. The collection, analysis, interpretation and dissemination of information relating to nutrition, food and agriculture are FAO's main functions [1]. The World Wide Web has proven to be a powerful means for FAO to disseminate multilingual information.

In this context, FAO was an early implementer of:

1. an online catalogue for documents produced by FAO (FAODOC, Figure 1), a multilingual online catalogue which contains bibliographic metadata of FAO electronic and printed documents [2];
2. the Electronic Information Management System (EIMS), a workflow management tool and database which manages the publication of electronic documents and multimedia resources on FAO's Web sites [3]; and
3. the Corporate Document Repository (CDR, Figure 2), a corporate output interface for FAO full text electronic publications stored in the EIMS [4, 5].

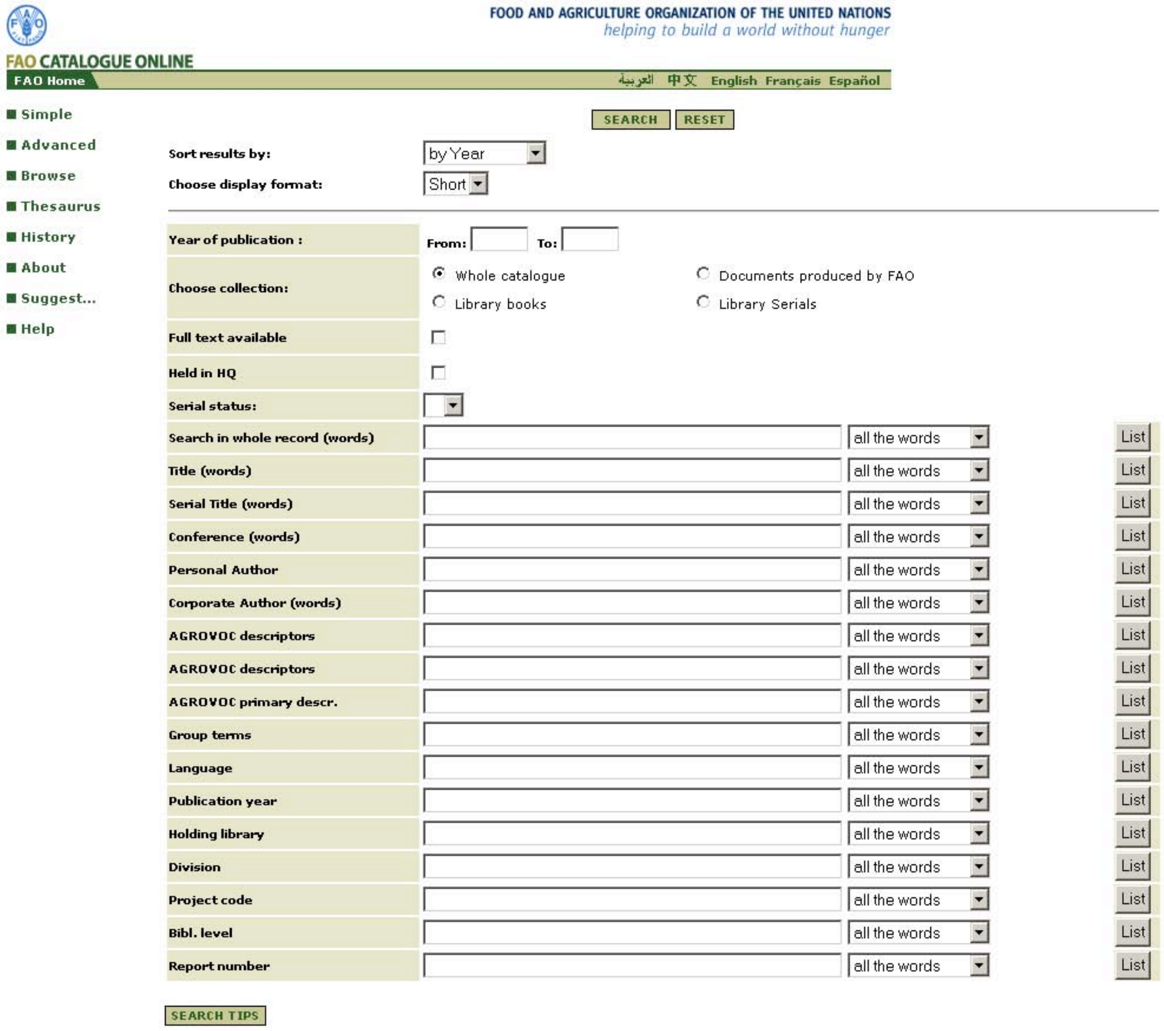
The FAODOC is a multilingual, online catalogue of documents and publications produced by FAO since 1945. The system uses UNESCO's CDS/ISIS software [6]. More than 160 000 documents have currently been catalogued. Since its inception, the FAODOC has focused on the production of high quality bibliographic records.

The FAO Web site was released in 1995 and the first electronic publishing workflow (through EIMS) was initiated in 1998. Currently, more than 38 550 resources (full text documents and multimedia items) are managed by the EIMS (Table 1). Photos, videos and audio are accessible through different systems on the FAO Web site. The CDR was conceived as the online digital library of FAO electronic documents and publications, as well as selected non-FAO material. At present, more than 23 000 full text documents are available through the CDR.

| Resource type | Number of Records |
|---|---|
| full text documents | 23 000 |
| photos | 8 500 |
| videos | 6 300 |
| audio | 750 |
| Total | 38 550 |

**Table 1: Resources at FAO (as at 10 April 2007)**

For each system described above, the objectives are different. The FAODOC focuses on the cataloguing of FAO documents. The EIMS deals with electronic publishing, especially the management at the full text level (rather than the description of documents). The CDR focuses on the dissemination of FAO documents archived through the EIMS. In 2003, a link between both databases was created, linking the FAODOC records to the full text documents archived in EIMS-CDR.


FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS
helping to build a world without hunger
FAO CATALOGUE ONLINE
FAO Home
العربية 中文 English Français Español
Simple
Advanced
Browse
Thesaurus
History
About
Suggest...
Help
SEARCH
RESET
SEARCH TIPS

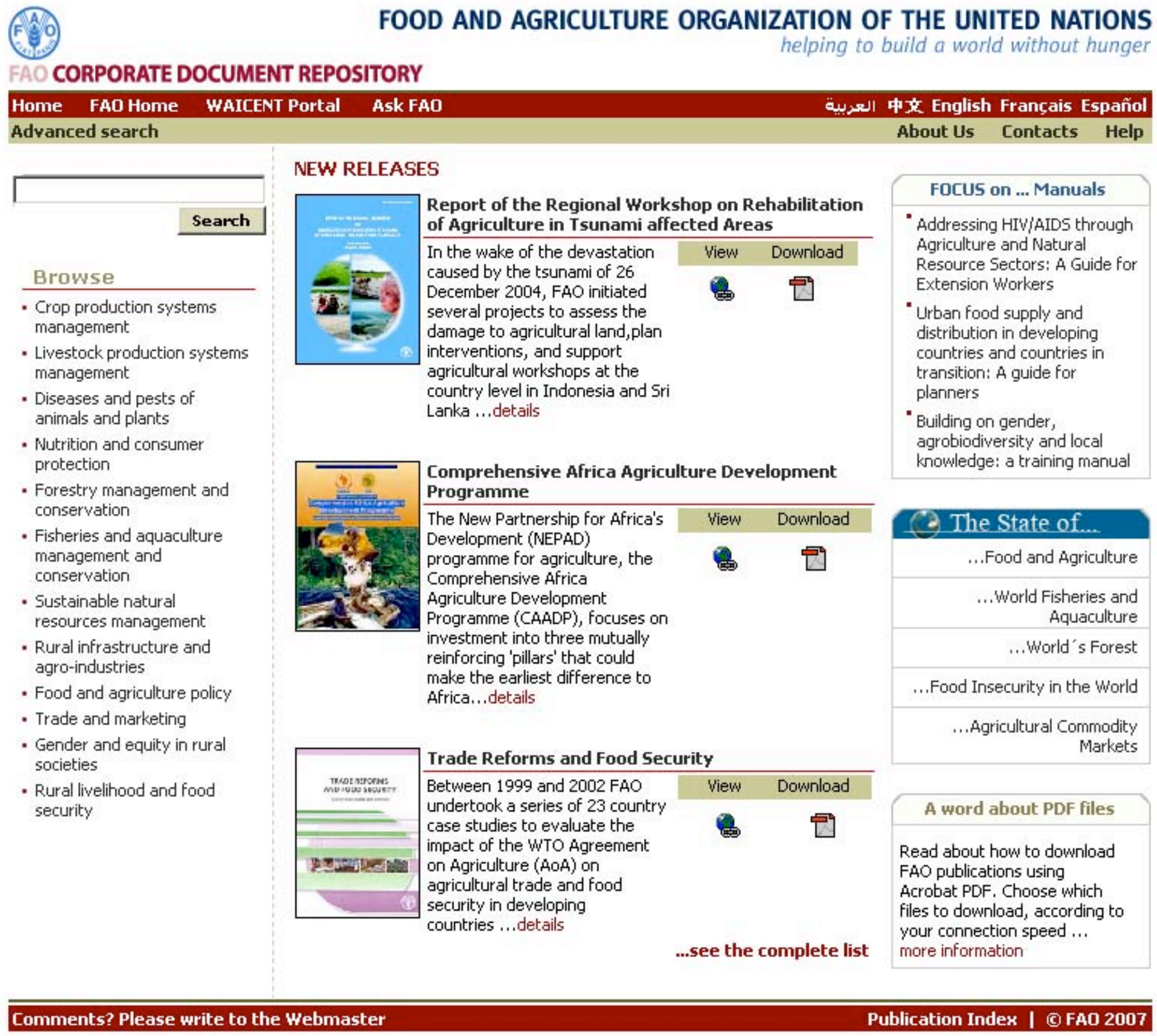
| | | | |
|---|---|---|---|
| Sort results by: | by Year | | |
| Choose display format: | Short | | |
| Year of publication : | From: To: | | |
| Choose collection: | ● Whole catalogue<br>○ Library books<br>○ Documents produced by FAO<br>○ Library Serials | | |
| Full text available | □ | | |
| Held in HQ | □ | | |
| Serial status: | | | |
| Search in whole record (words) | | all the words | List |
| Title (words) | | all the words | List |
| Serial Title (words) | | all the words | List |
| Conference (words) | | all the words | List |
| Personal Author | | all the words | List |
| Corporate Author (words) | | all the words | List |
| AGROVOC descriptors | | all the words | List |
| AGROVOC descriptors | | all the words | List |
| AGROVOC primary descr. | | all the words | List |
| Group terms | | all the words | List |
| Language | | all the words | List |
| Publication year | | all the words | List |
| Holding library | | all the words | List |
| Division | | all the words | List |
| Project code | | all the words | List |
| Bibl. level | | all the words | List |
| Report number | | all the words | List |

**Figure 1: FAODOC user interface**

This paper describes the process of merging the EIMS-CDR and the FAODOC and the creation of the FAO Open Archive. The result will be one unique sustainable digital repository offering a solid foundation for the collection, management, maintenance and timely dissemination of material published by FAO. To improve the effectiveness of the proposed repository, it will be necessary to streamline the existing electronic publishing workflow and to integrate the current functions into new modules. The FAO Open Archive is based on three key elements:

1. a metadata set based on international description guidelines and format;
2. a workflow procedure that guarantees the processing of all documents published by FAO; and
3. a system architecture based on cataloguing and electronic publishing.

This paper is divided into the following sections: Section 2 presents the current situation for the EIMS-CDR and the FAODOC; Section 3 details the objectives of the FAO Open Archive; Section 4 describes the workflow procedures, the new architecture, the compliance to International Standards for Bibliographical Description (ISBD) [7] and metadata sharing with other systems; and Section 5 is the conclusion and the next steps in implementing the FAO Open Archive.


FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS
helping to build a world without hunger
FAO CORPORATE DOCUMENT REPOSITORY
Home
FAO Home
WAICENT Portal
Ask FAO
العربية
中文
English
Français
Español
Advanced search
About Us
Contacts
Help
Search
Browse

- Crop production systems management
- Livestock production systems management
- Diseases and pests of animals and plants
- Nutrition and consumer protection
- Forestry management and conservation
- Fisheries and aquaculture management and conservation
- Sustainable natural resources management
- Rural infrastructure and agro-industries
- Food and agriculture policy
- Trade and marketing
- Gender and equity in rural societies
- Rural livelihood and food security

NEW RELEASES

**Report of the Regional Workshop on Rehabilitation of Agriculture in Tsunami affected Areas**

In the wake of the devastation caused by the tsunami of 26 December 2004, FAO initiated several projects to assess the damage to agricultural land,plan interventions, and support agricultural workshops at the country level in Indonesia and Sri Lanka ...details

**Comprehensive Africa Agriculture Development Programme**

The New Partnership for Africa's Development (NEPAD) programme for agriculture, the Comprehensive Africa Agriculture Development Programme (CAADP), focuses on investment into three mutually reinforcing 'pillars' that could make the earliest difference to Africa...details

**Trade Reforms and Food Security**

Between 1999 and 2002 FAO undertook a series of 23 country case studies to evaluate the impact of the WTO Agreement on Agriculture (AoA) on agricultural trade and food security in developing countries ...details

- Addressing HIV/AIDS through Agriculture and Natural Resource Sectors: A Guide for Extension Workers
- Urban food supply and distribution in developing countries and countries in transition: A guide for planners
- Building on gender, agrobiodiversity and local knowledge: a training manual

The State of...
...Food and Agriculture
...World Fisheries and Aquaculture
...World´s Forest
...Food Insecurity in the World
...Agricultural Commodity Markets
A word about PDF files

Read about how to download FAO publications using Acrobat PDF. Choose which files to download, according to your connection speed ... more information

Comments? Please write to the Webmaster
Publication Index | © FAO 2007

**Figure 2: CDR user interface**

## 2 Objectives

The objective of the FAO Open Archive is to create a unique sustainable digital repository for the dissemination of FAO publications and simultaneously, enhance interoperability with other information systems. The FAO Open Archive will guarantee efficient electronic publishing and metadata management, the effective dissemination of FAO information resources and the preservation of the Organization’s institutional memory.

# 3 Current Situation for EIMS-CDR and FAODOC

FAODOC has been managing all bibliographic information for FAO documents and publications for over 30 years (since 1976). Since 1998, FAO established a workflow to manage the electronic publishing and dissemination of FAO full text documents through the EIMS-CDR [8]. The EIMS-CDR and the FAODOC workflows, actors and content are described below.

## 3.1 EIMS-CDR, the Electronic Publishing and Digital Repository

There are four different user profiles in the EIMS-CDR workflow:

- originator – the person within the FAO unit responsible for providing the source files and/or the printed copy of the publication;
- data owner – the FAO unit responsible for the content of the publication;
- focal point – the person responsible in EIMS-CDR for managing requests from FAO units [9]; and
- liaison officer – the person within a FAO unit who ensures that publications are made available online. The liaison officer is the link between the originator and the focal point.

Detailed guidelines of the EIMS-CDR workflow are available to all FAO users and EIMS-CDR administrators. Following is a brief description of standard workflow steps:

1. The originator provides source files to the external printing unit. When the publication is printed, the external printing unit provides the focal point with the source files, the PDF version and the hard copy. In some cases files are provided by the originator;
2. The data owner creates and locates a record in EIMS;
3. The data owner notifies the focal point of the record and the uploaded files;
4. The focal point completes the record. Conversion to HTML or PDF is handled by focal points or outsourced to an external company. When conversion is completed, the focal point notifies the data owner of the test URL for reviewing the publication;
5. The data owner reviews the publication and either approves it or requests changes, by notifying the focal point;
6. The focal point reviews the final publication, publishes it and notifies the data owner of the public URL. If no conversion is required, the focal point prepares an HTML table of contents that links to the low-resolution PDF files and notifies the data owner of the public URL (in some cases only PDF files are published without the associated HTML pages).

Publications are made available in various electronic formats:

- Full HTML version; HTML loads quickly and is easier to read on-screen. ~14 000 records;
- Full PDF version; PDF is better for printing and downloading a local copy. ~2 200 records;
- Full HTML version and PDF version. ~6 500 records; and
- HTML table of contents linked to Full PDF version. ~500 records

## 3.2 FAODOC, the Online Catalogue

The FAODOC cataloguing process involves various actors:

- originator – the person within the FAO unit responsible for delivering to FAODOC the hard copy of the publications and/or the full text documents to be published in EIMS-CDR;
- EIMS-CDR focal point – the person who notifies the FAODOC cataloguer of a new record in EIMS-CDR, so they link the FAODOC record to the EIMS-CDR full text document; and
- cataloguer – the person who selects and catalogues the publications (hard copies and full text documents from EIMS-CDR).

The FAODOC manages the cataloguing of document and the dissemination of bibliographic information through an Online Public Access Catalogue (OPAC). There are procedures for the exchange of information between the FAODOC and the document producers, but there is no specific electronic tool to manage the reception of documents, as exists in the EIMS-CDR workflow. The lack of any workflow management system makes it difficult to control the reception and cataloguing of documents.

## 3.3 Main Differences between EIMS-CDR and FAODOC

The process of merging the two existing databases is a challenging task, as each has a different structure and workflow procedure. The first step towards the FAO Open Archive was to determine the similarities and differences between the EIMS-CDR and the FAODOC.

### 3.3.1 Software Overview

The EIMS-CDR was developed by FAO to manage the electronic publishing workflow. The CDR and the EIMS both run on a Microsoft Windows platform with an Oracle 9 database server. The software uses Microsoft's ASP programming language (Active Server Pages), with some ad hoc modules and functionalities developed in ASP.Net (the successor to ASP). The EIMS architecture results from the interaction of several modules that manage different aspects of the overall workflow. All modules interact with a single database that stores the records' descriptive metadata and detailed workflow information.

The FAODOC uses CDS/ISIS, a software package for information storage and retrieval – developed, maintained and disseminated by UNESCO. It is freely available for non-commercial purposes. The customization of data input and output interfaces occurred in Poland at the Institute for Computer and Information Engineering and at FAO.

### 3.3.2 Metadata Structure

CDS/ISIS manages a database whose main content is text, while the EIMS-CDR uses a relational Oracle database. The structure and logic of the two databases are completely different. However, these differences are not a barrier for the merger into a new single relational database.

Both systems use a very similar set of metadata fields to describe documents. The FAODOC contains detailed document information, while the EIMS-CDR provides fewer details on the actual document, but stores much information related to the actors, workflow and full text management. The mapping of the EIMS-CDR and the FAODOC databases has already occurred. It was not a complicated procedure, as both systems use a similar metadata field set. The compliance of both databases to the Dublin Core metadata standard and the AGRIS AP [10] at export level, facilitated the mapping. Only those fields required for the EIMS-CDR workflow have been added to those that already exist in the FAODOC.

### 3.3.3 Database Content

The EIMS-CDR and the FAODOC currently use FAO cataloguing guidelines. The decision to adopt international cataloguing standards was taken to guarantee interoperability with other digital repositories.

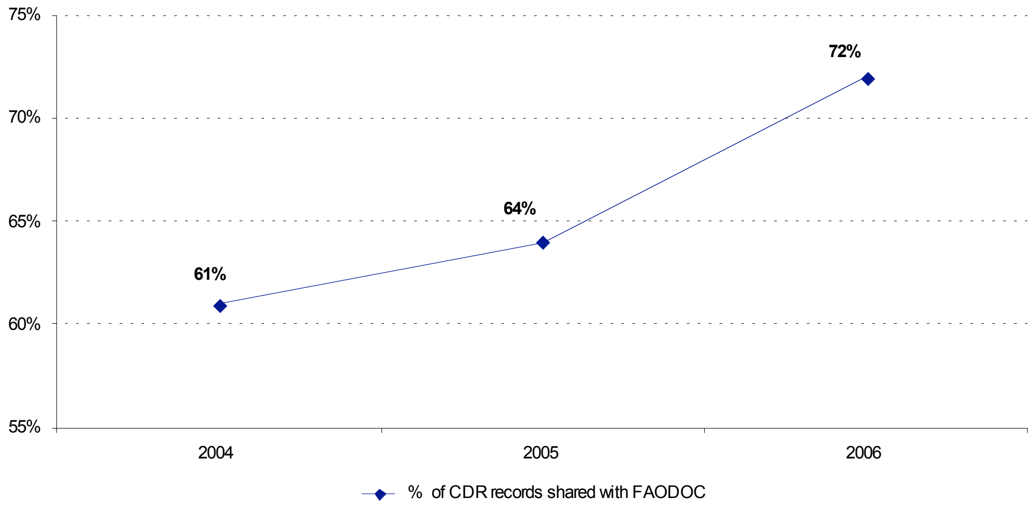


**Figure 3: Percentage of the EIMS-CDR records catalogued in the FAODOC**

In the EIMS-CDR, each record corresponds to one document (e.g., a book or a meeting report). The FAODOC catalogues documents and their analytics (e.g., a document is considered a book and the analytics are its chapters). Therefore, a book can have more than one record. The one-to-many relationship of records will be taken into consideration when merging data from the two databases.

The content of the two databases partially overlap, resulting in duplicate bibliographic records. The percentage of the EIMS-CDR full text documents linked from the FAODOC has increased over time (Figure 3): 72 percent of all records created in 2006 in the EIMS-CDR have been linked to from the FAODOC. This implies a duplication of effort (at metadata management level) and jeopardizes the dissemination and the maintenance of the FAO's institutional memory.

# 4 The Approach to Create the FAO Open Archive

The FAO Open Archive is based on the integration of the electronic publishing and the bibliographic cataloguing requirements. This merger requires the analysis of current workflows to detect similar procedures and reorganise them into a single coherent workflow. This process should focus on:

1. system architecture;
2. workflow procedure;
3. compliance with international data content standards; and
4. exposing metadata in a standardized way.

## 4.1 The New System Architecture

The architecture of the FAO Open Archive should integrate all features that are currently managed through the EIMS-CDR and the FAODOC. The FAODOC only manages the cataloguing process, but the FAO Open Archive must include the facility to deal with the reception of documents workflow, and improve the cataloguing module. The electronic publishing system is structured as a modular system where each module deals with a specific aspect of document publication. This approach will remain in the new architecture, integrated with new functionalities.

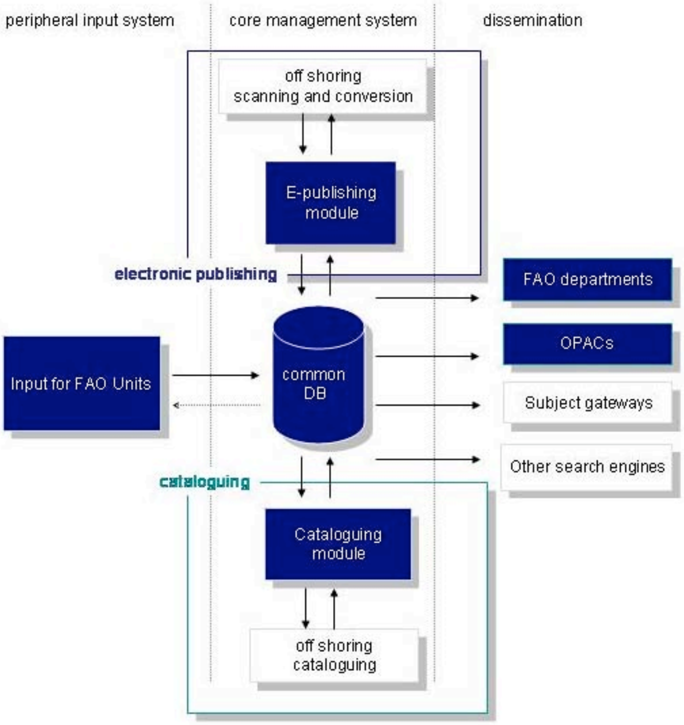


**Figure 4: FAO Open Archive architecture**

The FAO Open Archive architecture is detailed in Figure 4. The following elements define the architecture of the system:

1. integrated workflow; from left to right, the flow of information starts from the peripheral input system elements, passes through the core of the management system and to the dissemination interfaces;
2. common database; and
3. management of the two main functions of the FAO Open Archive; electronic publishing and cataloguing.

The objective of the system architecture is to manage all aspects of the electronic document life cycle. Electronic publishing and cataloguing will be managed through the same system and share the same database, e.g., from the document's creation, to its cataloguing, indexing and conversion to a suitable electronic format, to its dissemination on the Web.

**Input for FAO units**. This module will be used for data input and will be developed based on the current EIMS. FAO units now have individually customized EIMS interfaces. Each customization involves a basic internal workflow that can vary from one-step to multiple-step approval. FAO units are responsible for the introduction (and minimal description of documents) into the electronic publishing workflow. In the FAO Open Archive, FAO units will continue to provide data through a user-friendly system describing the document with a minimal set of metadata. With the FAO Open Archive, electronic publishing and cataloguing will share a common data entry point. The records that the FAO Open Archive will manage includes documents and multimedia files (photos, videos and audio) and non-FAO material (publications written in collaboration with FAO, yet FAO does not hold the copyright).

**Electronic publishing**. FAO will continue to publish documents online in electronic format. They will be managed through two modules:

- core module for electronic publishing – this module will be used to review the information from FAO units, based on EIMS, and to manage the conversion of full text documents into electronic formats (HTML, PDF, etc.); and
- scanning requests managing module – this module will be directly connected to the core module for electronic publishing and will be used to keep track of the work assigned to internal resources or of the work orders sent for scanning and/or conversion to external companies.

**Cataloguing.** FAO will offshore the cataloguing, using the minimal set of metadata and the full text provided by the FAO units. FAO cataloguers will check and validate the offshored records in order to guarantee the quality of the bibliographic description for the full text documents. Cataloguing will also be managed through two modules:

- core module for cataloguing – this module will be used to select records to be offshored for cataloguing and indexing and to check metadata quality. It will be used exclusively by cataloguers to manage the information to be released into the Open Archive; and
- cataloguing offshoring module – this module will be directly connected to the core module for cataloguing and will be used to manage the XML exports of data to be catalogued by external companies and to manage import and validation of offshored records.

## 4.2 Workflow Procedures

As well as the architecture, the workflow of the FAO Open Archive must integrate two main activities that so far have been conducted separately: electronic publishing and cataloguing. Figure 5 shows a top-down representation of the new workflow:

1. FAO units initiate a record by inserting a minimal set of metadata into the data input module. Only minimal information is requested to initiate a record: author, title, year and job number (a FAO unique identifier). The system verifies whether the job number exists in the database. A simple validation workflow within the peripheral input system will ensure that the records inserted are eligible for publication in the FAO Open Archive.

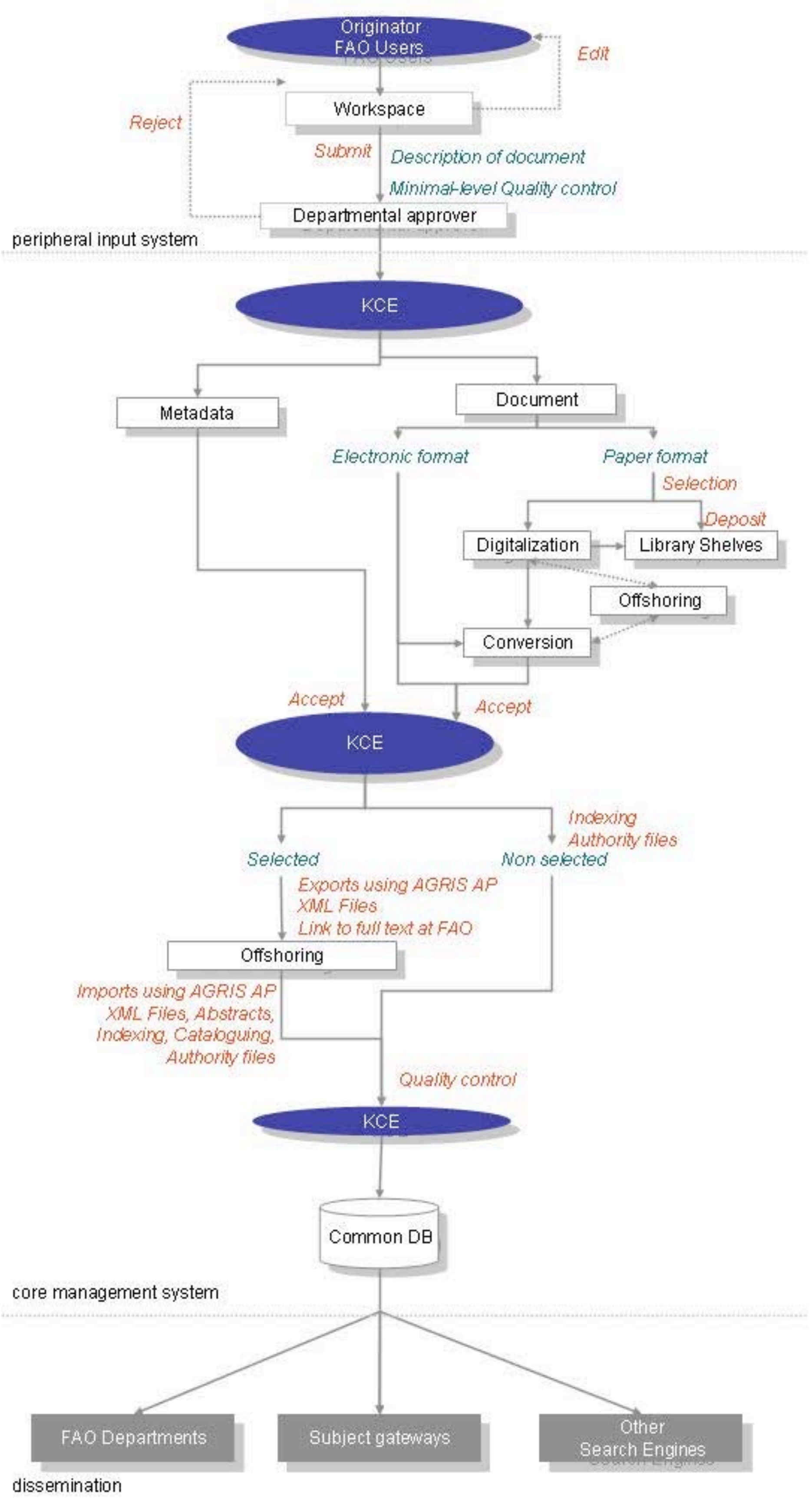


**Figure 5: FAO Open Archive Workflow**

2. The electronic publishing administration and the cataloguing administration are notified of the addition of a new record. They can take action simultaneously on the full text and the metadata of the records.

    2.1. If the document received is already in electronic format it requires validation and conversion to the most suitable format. This task can be carried out in-house or can be offshored. If the document needs digitalization then it is offshored for scanning.

    2.2. Using the minimal set of metadata in the system and the link to the full texts, the documents are catalogued and indexed by FAO and/or external cataloguers. The records that are selected for offshoring are exported using XML. When exported records are received from the external company they are imported into the system, checked and validated.

3. Validated records are disseminated through FAO Web sites. Moreover, search engines, services providers and digital libraries will harvest the records' metadata enhancing access to FAO documents.

## 4.3 Compliance with International Data Content Standards, ISBD

During the past few years, ISBD [11] has been identified as the standard most suitable for FAO. In April 2006, a study of the impact of changing FAO cataloguing rules recommended the adoption of ISBD rules:

> "*... recommend that FAO adopt the ISBD rules and build a system that will send and accept queries according to the OpenURL standard. In this way, FAO will build a system that will work with (interoperate with) other catalogues, while making FAO documents far more accessible to users. FAO, OCLC and other databases can create OpenURLs based on records that follow international guidelines and in this way, create an interoperable system* [12]".

ISBD rules are rigorous and exact. ISBD is based on the principles of adequate identification, searchability and consistency so that:

1. no two different documents can be confused with each other; and
2. the many details comprising a description, are presented in a uniform manner so that they can be interpreted without unnecessary ambiguity [13].

By applying the ISBD rules, FAO will not only enhance the international exchange of FAO records, but will also assist in the interpretation of records across languages, because ISBD records can be interpreted on a first level (identification of elements) by users of every language. This is because of the fixed order of ISBD records. Finally, ISBD is independent of any metadata format. In conclusion, ISBD rules are simple, exact, widely used and supported by the International Federation of Library Associations and Institutes (IFLA). ISBD will facilitate the interoperability with other institutions and/or services providers, as it is an international standard followed by many of the world's major libraries and bibliographic institutions.

One of the biggest challenges will be the handling of the legacy data; old records require re-cataloguing, e.g., titles need to be transcribed according to ISBD rules. A possible solution could be to import bibliographic records from databases that have already catalogued FAO documents, ignoring fields that are not relevant to FAO's needs and adding specific information already existing in FAO records, e.g., AGROVOC Thesaurus [14] descriptors. However, the legacy data can be updated, prioritizing those records which have the full text available and/or are accessed on a regular basis. The introduction of an additional code to distinguish old from new ISBD records is required.

The FAO units will introduce a minimal-level description based on ISBD and the offshored and FAO cataloguers could then bring the records to full ISBD level.

## 4.4 Exposing Metadata in a Standardised Way

This is a very important issue, and it has been addressed successfully by the Open Archives Initiative (OAI). Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) is a simple protocol that allows data providers to expose their metadata for harvesting to services providers. The FAO Open Archive will be OAI compliant, so the FAO metadata can be harvested by any services providers and/or digital libraries.

The concept of OAI-PMH can be applied to a wide range of digital materials, e.g. images, audio or videos. It is mandatory to expose metadata as Dublin Core. It is important to note that the protocol enables multiple metadata formats. These alternative forms of metadata can be as rich as is necessary to describe content. During the last few years, FAO has made an intensive effort to promote the exchange of high-quality metadata within the AGRIS Network, an international initiative based on a collaborative network of institutions in agriculture and related subjects. The AGRIS AP is a metadata format that facilitates sharing of metadata across different information systems. It is a metadata schema which uses elements from metadata standards such as Dublin Core (DC), Australian Government Locator Service Metadata (AGLS) [15] and Agricultural Metadata Element Set (AgMES) [16] namespaces. The standard enhances the quality of the description of agricultural information resources, enabling greater processing possibilities by service providers. The AGRIS AP has proved to be a successful initiative, and as a result, the FAO Open Archive will be fully compliant with the AGRIS AP at export level.

In conclusion, exposing metadata will:

1. improve the retrieval of FAO documents from a large number of sources (e.g., portals, aggregators and services providers);
2. allow aggregators to detect FAO documents and thereby help to disseminate them; and
3. enhance the visibility and awareness of FAO's available resources.

# 5 Conclusions and Next Steps

This paper illustrates the first phase for the creation of the FAO Open Archive, focussing on finding a strategy to solve:

1. the duplication of efforts in creating and managing metadata; and
2. the lack of integration of electronic publishing and cataloguing.

The relevant findings from this first phase are:

- The FAODOC and the EIMS-CDR will use a common database and a workflow supported by a workflow management system. FAO will supply FAO bibliographic metadata together with the full text.
- The conversion of the FAODOC and the EIMS-CDR to the FAO Open Archive will facilitate the data input and maintenance of information. The FAO units will continue to be involved in the metadata creation process.
- The use of ISBD rules will simplify the creation of metadata. The legacy data will be updated to ISBD standards, prioritizing those records, which a) are accessed on a regular basis, and b) have the full text available to improve the effectiveness of the OpenURL protocol.
- The visibility and dissemination of FAO documents will be maximized by exposing content through OAI-PMH. The FAO Open Archive should have the ability to transfer and use information in a uniform and efficient manner across multiple organisations and information technology systems.

The creation of the FAO Open Archive will strengthen FAO's role as a knowledge dissemination organization. The following phase is related to the software implementation. The integration of open source software into FAO Open archive is still under evaluation.

# Acknowledgements

We would like to thank Anne Aubert, Johannes Keizer, Giorgio Lanzarone, Romolo Tassone and Jim Weinheimer for their valuable contributions.

# Notes and References

[1] FAO Constitution, Article I. http://www.fao.org/docrep/x1800e/x1800e01.htm#1 Last accessed in April 2007.

[2] Catalogue for Documents produced by FAO (FAODOC) http://www4.fao.org/faobib/index.html Last accessed in April 2007.

[3] Electronic Information Management Services (EIMS). http://www.fao.org/eims/ Last accessed in April 2007.

[4] Corporate Document Repository (CDR) http://www.fao.org/documents/ Last accessed in April 2007.

[5] The Knowledge Exchange & Capacity Building Division (KCE) of FAO is the responsible for all the above mentioned systems.

[6] AGRIS/CARIS Centre of Information Management for international agricultural research http://www.fao.org/Agris/ Last accessed in April 2007.

[7] International Standards for Bibliographic Description (ISBDs http://www.ifla.org/VI/3/nd1/isbdlist.htm Last accessed in April 2007.

[8] SALOKHE, G.; PASTORE, A.; RICHARDS, B.; WEATHERLEY, S.; AUBERT, A.; KEIZER, J.; NADEAU, A.; KATZ, S.; RUDGARD, S.; MANGSTL; ANTON. *FAO's role in Information Management and Dissemination – Challenges, Innovation, Success, Lessons Learned*. 2005. ftp://ftp.fao.org/docrep/fao/008/af238e/af238e00.pdf Last accessed in April 2007.

[9] This task involves the scanning and conversion of documents, corrections, modifications and the publication of HTML/PDF files.

[10] The AGRIS Application Profile for the International Information System on Agricultural Sciences and Technology Guidelines on Best Practices for Information Object Description http://www.fao.org/docrep/008/ae909e/ae909e00.htm Last accessed in April 2007.

[11] In 1969 the International Federation of Library Associations and Institutes (IFLA) created a general framework for the creation of standards to regularize the form and content of bibliographic descriptions (Byrum, J.D., "The Birth and Re-birth of the ISBDs: Process and Procedures for Creating and Revising the International Standard BibIiographic Descriptions". *IFLA journal*, Vol. 27, No. 1, 2001). The work resulted in the ISBD rules which specify the requirements for the description and identification of the most common types of resources that are likely to appear in library collections.

[12] WEINHEIMER, J. (2006). *Consequences of changing FAO cataloguing rules & format with ISBD/AACR2/MARC21: a report for the Food and Agriculture Organization of the United Nations*. Internal report.

[13] COETZEE, H. (2005). *Do we still need bibliographic standards in computer systems?* http://www.liasa.org.za/interest_groups/igbis/papers/IGBIS_WSJul04_Bib_Stds_Helena_Coetzee.doc Last accessed in April 2007.

[14] AGROVOC is a multilingual structured and controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains. http://www.fao.org/aims/ag_intro.htm

[15] AGLS Metadata Standard http://www.naa.gov.au/recordkeeping/gov_online/agls/summary.html Last accessed in April 2007.

[16] Agricultural Metadata Element Set (AgMES) http://www.fao.org/aims/intro_meta.jsp Last accessed in April 2007.

# Five Years on – The Impact of the Budapest Open Access Initiative

*Melissa R. Hagemann*

Information Program, Open Society Institute
400 West 59th Street, New York, NY 10019, USA
e-mail: mhagemann@sorosny.org

## Abstract

Open Access was first defined by the Budapest Open Access Initiative following a meeting organized by the Open Society Institute/Soros foundations. The subsequent Open Access movement has had a large impact on the scholarly communications system. This is seen in the growing number of Open Access journals and institutional and subject-based repositories which have developed over the past five years. The reaction of publishers to the movement has been mixed with individual publishers (both commercial and non-profit) experimenting with the model while large publishers' associations have generally shown resistance. However, the movement continues to gain strength as research funding agencies adopt Open Access mandates to the research they support.

**Keywords:** open access; Budapest open access initiative; Open Society Institute

## 1 Introduction

February 14, 2007 marked the fifth anniversary of the release of the Budapest Open Access Initiative (BOAI), which offered the first definition of Open Access [1]. This paper examines the impact of the BOAI over the past five years. Background information on the role of the Open Society Institute (OSI)/Soros foundations will be provided, followed by an examination of key objective measurements for analyzing the impact of the BOAI.

In 2001 OSI's Information Program began to follow the developments of several projects which shared the ultimate goal of making peer-reviewed scholarly content freely available online. Among these projects were arXiv.org, the preprints archive for Physics, Mathematics, Computer Science and Quantitative Biology and the Public Library of Science's petition, which called on researchers not to submit their articles to any publisher which did not allow articles to be freely available after six months. OSI organized a meeting in Budapest in December 2001 which brought together a group of leaders who were exploring alternative publishing models. During the meeting it was decided to link the blossoming repository (or self-archiving) movement with Open Access journal publishing. Thus the BOAI defined these as two complementary strategies for achieving Open Access. The simultaneous promotion of the two strategies has proven to be highly productive. Ultimately to succeed, both strategies rely on mandating Open Access to publicly funded research.

Following the release of the BOAI, OSI's Information Program pledged $3 million to support Open Access initiatives. While OSI initially intended to spend these funds over a three year period, we realized that the transition to Open Access will require a longer time commitment on the part of OSI and more funding than initially pledged. This paper documents both the impact of OSI's direct funding of the principles outlined in the BOAI, as well as broader policy and funding discussions which followed the release of the BOAI.

## 2 Methodology

Key objective measurements for evaluating the impact of the BOAI include:

- a review of meetings which have followed the BOAI;
- the number of Open Access journals and institutional and subject-based repositories which have developed in the past five years;
- the number of sites which link to the BOAI as well as to some of the Open Access projects which OSI has funded;
- a review of the response of publishers to the Open Access movement;
- an examination of the major declarations and funders' policies regarding Open Access which have followed the BOAI.

## 3 The Development of a Movement

Having defined Open Access, the BOAI inspired lively debates among publishers, academics, librarians, and funders (both governmental and private) regarding the future of scholarly communication. Much of OSI's funding in the past five years has been dedicated to meetings, conferences and workshops which introduce the concept of Open Access. As of January 2007, OSI has provided $441,300 in funding to support over 40 meetings to introduce and promote Open Access throughout the world.

In addition to supporting meetings on Open Access, OSI has funded projects which directly advocate for Open Access. Examples of these are the Open Access News blog, which is written by Peter Suber. Open Access News has come to be regarded as the main source for information on the Open Access movement and this can be seen in the over 5,400 sites which link to it. OSI also supports some of SPARC's (the Scholarly Publishing and Academic Resources Coalition) work to advocate for Open Access. SPARC has developed the Alliance for Taxpayer Access, an organization representing taxpayers, patients, physicians, researchers and institutions that support Open Access to taxpayer-funded research.

Seeing the need to facilitate the discovery and use of Open Access journals and repositories, OSI funded the development of the Directory of Open Access Journals (www.doaj.org) and the Directory of Open Access Repositories (www.opendoar.org). The DOAJ was developed by Lund University Libraries and as of April 2007 lists 2,622 Open Access journals, an increase of over 2,300 since its launch in 2003.

To complement the DOAJ, OSI brought together a group of funders to support the development of the Directory of Open Access Repositories by the University of Nottingham and Lund University Libraries. Currently OpenDOAR lists 853 institutional repositories and 15,400 sites link to the OpenDOAR. As of April 2007, 522 sites link to the BOAI. In particular, organizations often link to the BOAI in reference to defining Open Access.

Beyond the Open Access meetings and projects which OSI has funded, the discussion regarding Open Access has been broadened since 2002 to include national, international and institutional funders. In 2003, the Howard Hughes Medical Institute (HHMI) and the Max Planck Society both held meetings which addressed Open Access from a funder's perspective. The HHMI meeting produced the Bethesda Statement [2] (the meeting was held at HHMI's headquarters in Bethesda, Maryland) and the Max Planck conference developed the Berlin Declaration [3]. Both the Bethesda Statement and the Berlin Declaration provide definitions of Open Access which focus on the role of funders. Thus adding the Budapest definition to this mix, many refer to the "BBB" definition of Open Access.

## 4 Publishers' Reaction to Open Access

The BOAI received stiff criticism from publishers' associations when it was announced in February 2002. Sally Morris of the Association of Learned and Professional Society Publishers (ALPSP) said: "We are convinced all of our scholarly communities will be ill-served by an initiative which promotes systematic institutional archiving of journal content without having in place a viable alternative model to fund the publication of that content. This can only serve to undermine the formal publishing process which these communities value. She warned against those who would 'give it all away first and then start worrying later" [4].

However by the fall of 2002, ALPSP and OSI held a joint workshop in London which described the Open Access publishing model. This was the first in a series of three ALPSP/OSI workshops. By the third workshop, Martin Richardson of Oxford University Press (OUP) described how OUP was experimenting with the hybrid model of Open Access. Through the hybrid model publishers offer authors the choice of paying the article processing fee and having their article made freely available online, or they can elect not to pay and then only journal subscribers will have access to the article. This model seems attractive to authors, as by electing to have their article made freely available through Open Access, it has the potential to reach a larger audience. When OUP adopted the hybrid model for their *Journal of Nucleic Acids*, they found that a high percentage of authors elected to pay the article processing fee. Based upon this response, OUP converted the journal to full Open Access [5]. The hybrid model offers publishers of traditional subscription-based journals a way to experiment with Open Access and allow the pace of change to be dictated by the authors themselves. Jan Velterop, former publisher of BioMed Central and currently the Director of Open Access at Springer, described how the hybrid model can work for publishers wishing to experiment with Open Access in his *Guide to Open Access Publishing and Scholarly Societies* [6] commissioned by OSI. Within Springer, Velterop leads the Springer Open Choice Program which allows authors who submit their articles to all Springer journals to choose the hybrid model of

Open Access. Through Springer Open Choice, authors are allowed to retrain their copyright. Springer has adopted the Creative Commons Attribution License 2.0 as the Springer Open Choice License [7].

In addition to subscription-based journals which are converting to Open Access, there are many new Open Access journals which have been developed. Today the largest commercial Open Access publisher is BioMed Central which publishes over 175 titles. SciELO (the Scientific Electronic Library Online), based in Brazil, publishes over 200 Open Access titles and is supported by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), in partnership with BIREME (the Latin American and Caribbean Center on Health Sciences Information). Hindawi Publishing, based in Cairo, publishes over 60 titles among a wide range of fields including Engineering, Life Sciences, Mathematics, and Physical Sciences. Most importantly, Hindawi Publishing has shown that the article processing fee business model is sustainable. As Paul Peters, Head of Business Development at Hindawi explained: "Based on our experience as a publisher of both subscription-based journals and author-pays open access journals, I would not only argue that the author-pays publishing model is sustainable, but also that it has many economic advantages over the subscription model. Even though our open access journal collection is only a few years old, we have already achieved profitability for the collection as a whole. Moreover, using a business model based on publication charges has enabled us to expand our publishing program in a much more sustainable way than we were able to using a subscription model" [8].

The Public Library of Science (PLoS), launched by Nobel Laureate Harold Varmus, Mike Eisen, and Pat Brown, has demonstrated that Open Access journals can compete with the top subscription-based journals in terms of producing high quality journals. *PLoS Biology* is ranked as the most highly cited general biology journal with an impact factor of 14.7 [9]. And PLoS is pushing the boundaries of the traditional concept of a journal with their new PLoS ONE which represents cutting edge innovation which could fundamentally change how research is communicated.

While individual publishers are experimenting with Open Access, some of the publishers' associations continue to strongly oppose it. This was highlighted in January 2007 when *Nature* revealed that the Association of American Publishers (AAP) had hired a high profile public relations firm, Dezenhall Resources headed by Eric Dezenhall, to attack the Open Access movement [10]. Dezenhall's corporate clients have reportedly included Enron and Exxon Mobil. "Dezenhall told the association's professional and scholarly publishing division, he could help – in part by simplyfing the industry's message to a few key phrases that even a busy senator could grasp. Phrases like: 'public access equals government censorship,' and 'government is seeking to nationalize science and be a publisher.' The publishers liked what they heard" [11].

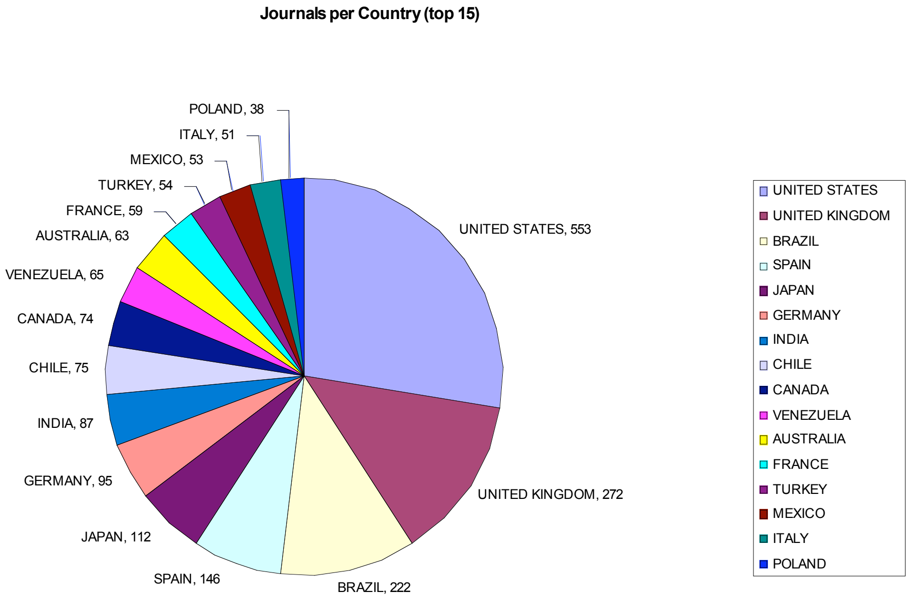


**Figure 1: Journals per Country in DOAJ (top 15)**

At the grassroots level, the reaction of publishers as well as users of research material to the Open Access movement can be seen by looking at statistics from the DOAJ. As previously mentioned 2,300 titles have been added to the DOAJ since its launch in 2003, thus 2,300 journals have either converted to or been launched as Open Access journals during this time. The importance of the DOAJ can be seen by the fact that 17,800 sites link to it. The DOAJ also receive over 5 million visits per month (although this figure does include robots).

By examining the countries where journals in the DOAJ are based (see Figure 1) it is clear that while many journals are based in the United States and the United Kingdon, Open Access has been adopted by publishers throughout the world, including those based in many developing countries.

| | **February 2007** | **December 2006** |
|---|---|---|
| 1 | 1582079: 37.32%: .com (Commercial) | 1234542: 37.58%: .com (Commercial) |
| 2 | 381159: 18.10%: [unresolved numerical addresses] | 365680: 23.31%: [unresolved numerical addresses] |
| 3 | 145734: 6.36%: .net (Networks) | 189425: 9.08%: .net (Networks) |
| 4 | 227688: 6.21%: .za (South Africa) | 160363: 5.77%: .za (South Africa) |
| 5 | 73271: 3.85%: .edu (USA Higher Education) | 50301: 2.39%: .br (Brazil) |
| 6 | 93426: 3.61%: .org (Non Profits) | 42332: 2.03%: .edu (USA Higher Education) |
| 7 | 8678: 3.28%: .bg (Bulgaria) | 16593: 1.89%: .org (Non Profits) |
| 8 | 77620: 3.15%: .ch (Switzerland) | 44513: 1.88%: .ch (Switzerland) |
| 9 | 55465: 1.75%: .br (Brazil) | 34474: 1.58%: .uk (United Kingdom) |
| 10 | 44309: 1.68%: .de (Germany) | 15913: 1.44%: .de (Germany) |
| 11 | 38913: 1.38%: .uk (United Kingdom) | 24092: 0.84%: .fr (France) |
| 12 | 22442: 0.80%: .ca (Canada) | 15505: 0.71%: .it (Italy) |
| 13 | 23282: 0.73%: .fr (France) | 9933: 0.61%: .in (India) |
| 14 | 18516: 0.73%: .my (Malaysia) | 8485: 0.59%: .jp (Japan) |
| 15 | 15142: 0.61%: .se (Sweden) | 10227: 0.54%: .tr (Turkey) |
| 16 | 11221: 0.61%: .in (India) | 9997: 0.50%: .ca (Canada) |
| 17 | 14596: 0.57%: .mx (Mexico) | 8553: 0.47%: .pl (Poland) |
| 18 | 15897: 0.54%: .it (Italy) | 9999: 0.45%: .se (Sweden) |
| 19 | 7332: 0.45%: .jp (Japan) | 10889: 0.44%: .gr (Greece) |
| 20 | 10317: 0.43%: .nl (Netherlands) | 8329: 0.41%: .mx (Mexico) |
| 21 | 10485: 0.41%: .tr (Turkey) | 8584: 0.40%: .be (Belgium) |
| 22 | 4753: 0.38%: .dk (Denmark) | 6726: 0.40%: .es (Spain) |
| 23 | 9422: 0.38%: .au (Australia) | 7104: 0.39%: .nl (Netherlands) |
| 24 | 7384: 0.36%: .es (Spain) | 7727: 0.36%: .pt (Portugal) |
| 25 | 7944: 0.36%: .pl (Poland) | 6114: 0.33%: .fi (Finland) |

**Figure 2: Hits to DOAJ based upon country (top25).**

The high global appeal of Open Access journals is also seen by examining the hits to the DOAJ based upon country (see Figure 2). While the wealthy research countries are represented, many developing countries also make the top 25, thus demonstrating that Open Access journals have been promoted widely and deely to the global research community. The high percentage of hits coming from unresolved domains could be due to the fact that many users in developing countries access the DOAJ through internet cafes and other third party access points.

| | **2004 Feb** | **2004 Nov** | **2005 Feb** | **2005 Nov** | **2006 Feb** | **2006 Nov** | **2007 Feb** |
|---|---|---|---|---|---|---|---|
| **Successful requests:** | 264,931 | 1,318,720 | 1,225,736 | 1,945,841 | 2,632,710 | 2,607,935 | 3,062,684 |
| **Redirected requests** | 57,660 | 513,306 | 395,886 | 328,585 | 525,862 | 1,745,736 | 2,318,193 |
| **Distinct files requested:** | 33,016 | 171,181 | 272,397 | 280,800 | 487,478 | 738,879 | 776,702 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Distinct hosts served:** | 33,107 | 120,320 | 81,189 | 171,378 | 138,900 | 231,663 | 175,055 |
| **Data transferred MB:** | 1,570 | 12,960 | 11,440 | 20,420 | 27,330 | 23,900 | 25,990 |
| **Link to journal** | | | | | | 750,677 | 836,151 |
| ***Explanation:*** | | | | | | | |
| **Successful requests** | Each time a user prompts the server to show a file it is a request | | | | | | |
| **Redirected requests** | A redirected request can be either a redirection within DOAJ (i.e. from a bibliographic record to an abstract) or from DOAJ to an external server (i.e. from an abstract in DOAJ to the full-text on a publisher's site). | | | | | | |
| **Distinct files requested:** | Indicates how many different files in the DOAJ have been requested during one month. | | | | | | |
| **Distinct hosts served:** | Indicates how many different registered IP-addresses have consulted the DOAJ during one month. | | | | | | |
| **Data transferred MB:** | Indicates how much data has been transferred (downloaded) from DOAJ during one month. Take in consideration that one metadata record is only a very small number of bytes - 1000 Megabytes= 1 Gigabyte. | | | | | | |
| **Link to journal** | Indicates how many times users have used the DOAJ to go to an abstract or full-text on the publishers sites during one month. | | | | | | |

**Figure 3: DOAJ requests**

And finally, the high use of the Open Acess journals in the DOAJ (see Figure 3) is seen in the growing number of requests which the DOAJ has received over the past three years.

## 5 Mandating Open Access – The Role of the Funding Agencies

The role of the research funders within the Open Access movement is extremely important. By 2003 the Open Access movement was advocating for Open Access to research supported by both governmental and private research funders. The research funders have begun to adopt mandates for Open Access (or Public Access in the case of government-supported research as this research is supported by tax dollars). The message that research funders (and taxpayers) are essentially paying twice for the same information has resonated with funders. In the case of government funded research, the public supports the research itself through grants from the federal research agencies and then the public (through libraries, hospitals, etc.) must purchase the journals in which the publicly funded research is published to access the research results.

In 2003 the Wellcome Trust published an economic analysis of scientific publishing [12]. Based upon this report, the Trust decided to pursue an Open Access policy for the research which it funds. This ultimately led to the Trust becoming the first funder to mandate Open Access to all of the research it funds in September 2006 [13].

The Science and Technology Committee of the House of Commons launched an Inquiry into the state of Scientific Publishing in 2004. Its final report concluded that "the current model of scientific publishing is unsatisfactory" and "recommends that the Research Councils and other Government funders mandate their funded researchers to deposit a copy of all articles in repositories." [14]. Although the report was released in 2004 it took some time for the Research Councils in the UK to adopt policies mandating Open Access. Today five out of the seven Research Councils [15] mandate Open Access to the research which they fund. Of particular significance, among the five which mandate Open Access is the Medical Research Council. This coupled with the mandate from the Wellcome Trust insures that the bulk of medical research funded in the UK will be available through Open Access.

A 2006 study by the European Commission on the Economic and Technical Evolution of the Scientific Publication Markets of Europe recommended public access to publicly-funded results [16]. This study was discussed at a meeting organized by the Commission on Scientific Publishing in the European Research Area in February 2007 in Brussels. As a result of the meeting, the Commission will now include the costs of Open

Access publishing as an eligible cost in Community funded projects and will begin discussions with the European Parliament and the Council regarding mandating Open Access [17].

In the U.S., the Federal Research Public Access Act (FRPAA) will be re-introduced this spring. FRPAA would mandate Public Access to research funded by the eleven largest government departments and agencies (i.e. National Institute of Health, National Science Foundation, Department of Energy, etc.). FRPAA would require that every federal agency with an annual research budget of $100 million or more implement a public access policy which would require researchers who receive full or partial support from the agency to deposit a copy of their article in a stable digital repository maintained by that agency or in another suitable repository that permits free public access, interoperability, and long-term preservation no later than six months after the article has been published in a peer-reviewed journal. This would be a huge improvement over the current NIH Public Access Policy which "requests" NIH funded authors to deposit a copy of their article in PubMed Central and has seen a very low compliance rate on the part of the authors [18].

Funding agencies in developing and transition countries are also considering mandating Open Access to the research which they fund. In Ukraine, a Parliamentary Inquiry on Harmonization of Governmental Educational Policies was launched in December 2005 and concluded that the Ministry of Education and Science should encourage the development of Open Access resources in science, technology and education with Open Access a condition of state funded research. Subsequently, an Open Access Working Group was formed in Ukraine with representatives of the Parliamentary Committee on Science and Education, the State Fund for Fundamental Research, the Scientific and Publishing Council of the National Academy of Science of Ukraine, the Ministry of Science and Education, the National Library of Ukraine, the State Department of Intellectual Property, the Kyiv public administration, and the International Renaissance Foundation (Soros Foundation–Ukraine) [19]. In South Africa, the South African National Research Foundation has pledged to support all costs associated with their grantees publishing in Open Access journals. And the Library of the Chinese Academy of Sciences held the first Open Access meeting in China in June 2005 and is working with other government funding bodies to support Open Access.

## 6 Lessons Learned

As mentioned earlier, OSI initially pledged $3 million to support the Open Access movement when the BOAI was launched in 2002. Since then OSI has seen that the transition to Open Access will require a longer time committment on our part and more funding than initiatlly pledged. In 2002 it was hoped that other foundations would join in supporting Open Access. With the exception of the Gordon and Betty Moore Foundation and the Sandler Family Supporting Foundation which have provided generous support to PLoS, other foundations have not embraced Open Access, although some of the leading American foundations provide substantial support to other open content issues such as Intellectual Property Rights reform and the development of open source software. More philianthropic support directed at advocating for the adoption of Open Access mandates by govenment and research funding institutions would be extremely helpful in countering the lobbying efforts of the large publishers' associations.

From OSI's experience with the BOAI it is clear that it was important to first define Open Access and develop specific strategies for achieving it. This allowed the key stakeholders to develop communities and subsequently a movement to support Open Access. This could serve as an example for the development of other movements around open content issues, such as open educational resources.

## 7 Directions for the Future

While the developments over the past five years are encouraging, much still remains to be done for Open Access to meet its full potential. Among the top priorities for the movement are:

1. Mandates from governments/funding agencies: Europe appears to be leading the way in terms of adopting significant mandates with the leadership of the Wellcome Trust and the five Research Council in the UK which have adopted mandates. In the U.S., while the FRPAA will be re-introduced this year in the Senate, strong opposition to it, led by AAP, poses a real obstacle to its adoption and increased support for public access advocacy will be needed;

2. Mandates from universities for deposit of material in repositories: In addition to developing repositories, more universities must adopt mandates for the deposit of all research written by

those affiliated with the university in the institutional repositories. This will require continued advocacy at many levels of the university administration and faculty,

3. The development of more Open Access journals: Some estimate that there are 24,000 peer-reviewed journals, thus this would mean that just over 10% are Open Access if one considers that the DOAJ lists 2,622 Open Access titles. More Open Access journals must be developed so that authors can have a choice to publish in an Open Access journal as opposed to a subscription-based journal. In addition to the numbers, it is important that the quality of the Open Access journals is high so that authors will elect to publish in them,

4. Continued unity of the Open Access movement: The Open Access movement (the Open Access publishing and the self-archiving/repositories communities) must remain united behind the common goal of making peer-reviewed content freely available and not allow differing mandates directed at journals or repositories to divide the movement.

## 8 Conclusion

The impact of the BOAI is clearly seen when one considers that before the meeting in Budapest, there was not even a term or definition for Open Access. Now Open Access is being debated by governments and publishers and mandated by funding bodies and universities. Much still remains to be achieved, but it is clear that Open Access has permanently changed the field of scholarly communication.

## Acknowledgements

I would like to thank Leslie Chan, Program Supervisor for the Joint Program in New Media Studies and the International Studies Program at the University of Toronto at Scarborough for his guidance in the development of this paper as well as Lars Björnshauge, Director of Lund University Libraries, for providing a wealth of data on the DOAJ.

## Notes and References

[1] The BOAI defines Open Access as the free availability of peer-reviewed literature on the public internet, permitting any user to read, download, copy, distribute, print, search, or link to the full texts of the articles. See http://www.soros.org/openaccess/.

[2] Bethesda Statement on Open Access Publishing: http://www.earlham.edu/~peters/fos/bethesda.htm.

[3] Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities: http://oa.mpg.de/openaccess-berlin/berlindeclaration.html.

[4] DODD, D. Access to research should be open to all, say many in the scientific community. Try telling that to publishers, *Information World Review*, April 15, 2002, p. 9.

[5] Oxford University Press Release. Oxford Journals takes bold step towards free access to research. 26 June, 2004. http://www.oxfordjournals.org/our_journals/nar/narpressjun04.pdf.

[6] VELTEROP, J.M. *Open Access Publishing and Scholarly Societies*. July 2005. http://www.soros.org/openaccess/scholarly_guide.shtml.

[7] Springer Open Choice License: http://www.springer.com/dal/home/open+choice?SGWID=1-40359-12-161193-0&teaserId=55557&CENTER_ID=115382.

[8] See Paul Peters comments in the *Nature Newsblog*, June 21, 2006: http://blogs.nature.com/news/blog/2006/06/openaccess_journal_hits_rocky.html.

[9] See overview of PLoS Journals: http://www.plos.org/journals/index.html.

[10] GILES, J. PR's 'pit bull' takes on open access. *Nature*, Vol. 445/25 January 2007, p. 347.

[11] WEISS, R. Pubilshing Group Hires 'Pit Bull of PR'. *Washington Post*, January 26, 2007.

[12] Economic Analysis of Scientific Publishing. Commissioned by the Wellcome Trust, January 2003. http://www.wellcome.ac.uk/assets/wtd003182.pdf.

[13] Wellcome Trust position statement in support of open and unrestricted access to published research. Last updated 14 Marc h 2007. http://www.wellcome.ac.uk/doc_WTD002766.html.

[14] Scientific Publications: Free for all? *Select Committee on Science and Technology, Tenth Report*, 7 July 2004. http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39914.htm.

[15] See SHERPA's Juliet site on research funders' open access policies. www.sherpa.ac.uk/juliet.

[16] DEVROEY, J.P.; DUJARDIN, M.; VANDOOREN, F. *Study on the economic and technical evolution of the scientific publications markets in Europe*. Commissioned by DG-Research, European Commission, January 2006. http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf.

[17] Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee on scientific information in the digital age: access, dissemination and preservation. COM(2007) 56 final, 14 February 2007. http://ec.europa.eu/research/science-society/document_library/pdf_06/communication-022007_en.pdf.

[18] See The Alliance for Taxpayer Access site: FRPAA: http://www.taxpayeraccess.org/frpaa/index.html#issue

[19] See Access to Knowledge, Ukraine site. Open Access Working Group formed in Ukraine: http://www.a2k.org.ua/news.php?id=1172&lng=en

# Openness in Higher Education:<br>Open Source, Open Standards, Open Access

*Brian Kelly[1]; Scott Wilson[2]; Randy Metcalfe[3]*

[1] UKOLN, University of Bath, Bath, BA2 7AY, United Kingdom
e-mail: b.kelly@ukoln.ac.uk
[2] CETIS, University of Bolton, Deane Road, Bolton BL3 5AB, United Kingdom
email: scott.bradley.wilson@gmail.com
[3] OSS Watch, University of Oxford, 13 Banbury Road, Oxford OX2 6NN, United Kingdom
e-mail: randolph.metcalfe@oucs.ox.ac.uk

## Abstract

For national advisory services in the UK (UKOLN, CETIS, and OSS Watch), varieties of openness (open source software, open standards, and open access to research publications and data) present an interesting challenge. Higher education is often keen to embrace openness, including new tools such as blogs and wikis for students and staff. For advisory services, the goal is to achieve the best solution for any individual institution's needs, balancing its enthusiasm with its own internal constraints and long term commitments. For example, open standards are a genuine good, but they may fail to gain market acceptance. Rushing headlong to standardize on open standards may not be the best approach. Instead a healthy dose of pragmatism is required. Similarly, open source software is an excellent choice when it best meets the needs of an institution, but not perhaps without reference to those needs. Providing open access to data owned by museums sounds like the right thing to do, but progress towards open access needs to also consider the sustainability plan for the service. Regrettably institutional policies and practices may not be in step with the possibilities that present themselves. Often a period of reflection on the implications of such activity is what is needed. Advisory services can help to provide this reflective moment. UKOLN, for example, has developed of a Quality Assurance (QA) model for making use of open standards. Originally developed to support the Joint Information Systems Committee's (JISC) digital library development programmes, it has subsequently been extended across other programmes areas. Another example is provided by OSS Watch's contribution to the development of JISC's own policy on open source software for its projects and services. The JISC policy does not mandate the use of open source, but instead guides development projects through a series of steps dealing with IPR issues, code management, and community development, which serve to enhance any JISC-funded project that takes up an open source development methodology. CETIS has provided a range of services to support community awareness and capability to make effective decisions about open standards in e-learning, and has informed the JISC policy and practices in relation to open standards in e-learning development. Again, rather than a mandate, the policy requires development projects to become involved in a community of practice relevant to their domain where there is a contextualised understanding of open standards.

**Keywords:** open standards; open source; open access; quality assurance; advisory services

## 1 Introduction

The Joint Information Systems Committee (JISC) of the UK education funding councils has been engaged in a long-running process of engaging with the concept of 'openness' in educational technology and digital content. This engagement has moved through several phases, from initial evangelism into today's more pragmatic stance, and effected through the agency of three services:

- UKOLN has been charged with the development of the JISC information environment, formerly known as the Distributed National Electronic Resource (DNER), the UK education sector framework for the distribution of published digital content;
- CETIS is the Centre For Educational Technology & Interoperability Standards, and has responsibility for the development of open standards to support e-learning;
- OSSWatch provides advice and guidance on the use of open-source technologies in education.

Together these three services offer the JISC support at the policy and strategy level on the three strands of 'openness' in technology discourse - namely, open content, open standards, and open source. In each area the emergence of widespread use of social software and distributed systems (the 'Web 2.0' phenomena) has provided a disruption affecting each service and its strategy on 'openness'.

## 2 Definition of Open Standards

In a paper on open standards it is important to have a clear definition of the meaning of the term. In practice, however, it can be difficult to reach an agreed definition. Rather than attempting to produce a formal definition the following list of the characteristics of open standards is given:

- The development of open standards is the responsibility of a trusted neutral organisation;
- The responsibility for the ongoing maintenance and development of open standards is taken by a trusted neutral organisation;
- Involvement in the development of open standards is open to all;
- There are no discriminatory barriers to use of open standards;
- Access to open standards is available to all, without any financial barriers.

It should be noted, however, that such characteristics do not necessarily apply to all organisations with a responsibility for open standards. For example within organisations such as W3C (the World Wide Web Consortium) discussions on areas in which standardisation will occur are decided by member organisations who have paid the required membership fee. Similarly the initial discussions and agreements on the preferred approaches to the standardisation work may be determined by such member organisations. Also standards produced by organisation such as the BSI (British Standards Institution) are not necessarily available free-of-charge.

## 3 Why Use Open Standards?

Open standards are important in the development of networked services for several reasons. They aim to:

**Support interoperability:** Interoperability is often critical to those creating digital services. There will be a need to ensure that services and data can be used not only within a correct environment, but also across other digital services and across other application areas. A prime purpose of open standards is to provide such interoperability.

**Maximise access:** Cultural heritage services normally seek to maximise access to their resources and services. Ideally access will not be limited by constraints such as the device used by the end user; their physical location; their location on the network; etc. or personal factors such as disabilities;

**Provide application- and device-independence:** The dangers of lock-in to particular applications or hardware platforms are widely acknowledged;

**Ensure architectural integrity:** Unlike proprietary solutions, for which the development and intended usage is likely to be constrained by commercial and competitive factors, open standards which are developed within a wider context can help to ensure architectural integrity across a wide range of scenarios;

**Provide long-term access to resources and services:** Long term access to scholarly resources and cultural heritage resources is of particular importance for public sector organisations.

The authors of this paper feel that an understanding of such benefits is widely accepted within the development community. What, therefore, are the barriers to an implementation of a vision based on this approach?

## 4 The Complexities of Open Standards

The reality is that despite the widespread acceptance of the importance of open standards and the feeling among some that use of open standards should be mandatory in the development of networked services in practice, many organisations fail to implement open standards in their provision of access to digital resources. This may be due to several factors:

**Disagreements over the Meaning**: There are many complex issues involved when selecting and encouraging use of open standards. Firstly there are disagreements over the definition of open standards. For example Java, Flash and PDF are considered by some to be open standards, although they are, in fact, owned by Sun, Macromedia and Adobe, respectively, who, despite documenting the formats and perhaps having open processes for the evolution of the formats, still have the rights to change the licence conditions governing their use (perhaps due to changes in the business environment, company takeovers, etc.) Similarly there are questions regarding the governance of apparent open standards, with the control of RSS 1.0 and RSS 2.0 providing an interesting example; this lightweight but powerful syndication format for Web context has a complex history plagued by disagreements over governance and the roadmap for future developments;

**Difficulties in Mandating and Enforcing Compliance**: There are also issues with the mandating of open standards. For example: What exactly does ‘must’ mean? When told you must comply with HTML standards a developer working on a project might first ask what if I don't? Then what if nobody does? They might also ask what if I use PDF instead of HTML? There is a need to clarify the meaning of must and for an understandable, realistic and reasonable compliance regime;

**Failure in the Market Place**: It also needs to be recognised that open standards do not always succeed in gaining acceptance in the market place: they are often regarded as too complex to be deployed and the user community may be content to use existing closed solutions and reluctant to make the investment needed to make changes to existing working practices;

**Failure to Satisfy User Needs and Expectations**: There is a danger that a development approach over-emphasises the importance of open standards to the detriment of the end user and the end user’s needs and expectations. It is often tempting to look only at the benefits of open standards for the developer or the provider of a service. We can see the temptation to develop a service based on a rich standard which can address a wide variety of use case scenarios. The danger would be that the end user rejects the service in preference to a simpler one.

Despite such reservations, in reality many IT development programmes are successful. The success may be based on the deployment of agreed and well-defined open standards. However in other cases development work may adopt a more pragmatic approach, making use of mature open standards, but having a more flexible approach to newer standards, for which there has been no time to reflect on the strengths and weaknesses and the experiences gained in their use.

## 5 Experiences in the UK

The Joint Information Systems Committee (JISC) (http://www.jisc.ac.uk/) who provide leadership in the innovative use of Information and Communications Technology to support education and research in the UK, have traditionally based their funding of development programmes around the use of open standards. Technical development for JISC’s eLib programme, which was launched in 1996, was based on a standards document (eLib, 1996). The document formed the basis of a revised standards document which was produced to support JISC’s Distributed National Electronic Resource (DNER) programme (which was later renamed the JISC Information Environment). Standards document (JISC, 2001). This work in turn influenced the NOF-digitise Technical Standards document (NOF, 2001) which was used by the national NOF-digitisation programme, which was responsible for digitisation projects across the cultural heritage sector.

The authors have been involved in providing technical advice and a support infrastructure for JISC-funded development programmes.

**Experiences of the QA Focus Project**

Although projects funded by the eLib programme were expected to comply with the eLib standards document, in practice compliance was never formally checked. It was probably sensible at the time (the mid 1990s) to avoid mandating a formal technical architecture and corresponding open standards – that could easily had led to mandating use of Gopher! In those early days of the Web, we were seeing rapid developments in the variety of services which were being provided on the Web and many new open standards being developed. However over time, and as the Web matured and the rate of innovation slowed, there was an increasing realisation of the need to provide a more stable environment for technical developments and the corresponding need to address the issue of compliance.

In 2000 JISC funded the QA Focus project (http://www.ukoln.ac.uk/qa-focus/) to develop a quality assurance framework, which would help ensure that future projects would comply with standards and recommendations and deploy best practices (Kelly, 2003). The project's aim was to develop a quality assurance (QA) methodology which would help to ensure that projects funded by JISC digital library programmes were functional, widely accessible and interoperable; to provide support materials to accompany the QA framework and to help to embed the QA methodology in projects' working practices. Liaison with a number of projects provided feedback on the current approach to use of standards. The feedback indicated: (a) a lack of awareness of the standards document; (b) difficulties in seeing how the standards could be applied to projects' particular needs; (c) concerns that the standards would change during the project lifetime; (d) lack of technical expertise and time to implement appropriate standards; (e) concerns that standards may not be sufficiently mature to be used; (f) concerns that the mainstream browsers may not support appropriate standards and (g) concerns that projects were not always starting from scratch but may be building on existing work and in such cases it would be difficult to deploy appropriate standards. Many of these were legitimate concerns, which needed to be addressed in future programmes.

This feedback was very valuable and provided a counter-balance to views which suggested the need for a heavyweight compliance regime which forced projects to comply fully with a technical architecture and corresponding open standards. The feedback led to the development of a contextual framework which is described later.

## 6 Open Standards: The CETIS Experience

In the late 1990s CETIS began life as the UK IMS Centre, a project funded by JISC to engage in the new IMS (instructional management systems) specification consortium. IMS began developing a series of specifications for XML data and content interoperability for elearning following the emerging paradigm of 'Learning Objects'. CETIS engaged in the development of the specifications, while also engaging with the the UK education community to disseminate information about open standards, promoting a message that placed open standards as the key mechanism for preventing vendor lock-in and supporting long-term sustainability for the newly emerging 'Virtual Learning Environment' technology sector. As the sector developed, CETIS expanded to engage in a wide range of open standards work at a UK, European, and international level.

This message proved very attractive for policy-makers, who were keen to find a new procurement strategy following the unpopularity of the 'single primary vendor' approach that had been used previously within the schools sector, but still needed to provide some form of strategic co-ordination to prevent resources being wasted. Open standards seemed an ideal tool for this policy task: standards could be mandated such that the choice of systems were restricted to those that could conform; these conforming systems could then be more easily replaced by institutions using the interoperability effected by open standards if they were no longer the optimal choice. This style of procurement policy was adopted in various ways by the Learning and Skills Council, JISC, BECTa, and the DfES, and continues to be the key approach of agencies in the UK education sector to this day through initiatives such as the e-Framework[1], the BECTa Learning Platform Framework, and DfES Information Standards Board.

While the overall message has been an attractive one at the policy level, the experience of open standards at a practical level has proved less clear-cut. In particular, the intended effect of interoperability and reduced opportunity for vendor lock-in has not always been well served by the means of open standards. There have been influences from the political, business and technology context of the development and application of open

[1] See http://www.e-framework.org

standards that in some cases have served to either reduce or completely reverse the effect of standards on interoperability.

The process of standardisation can be a difficult one for those concerned. For example, the specification process itself was being driven largely by the vendors themselves, for whom it may be argued the interests are not served best by the agenda of open standards. A good example of this is the first attempt by IMS at a standards framework for Learning Management Systems. This was implemented by the company now known as BlackBoard as a 'reference implementation' of the APIs defined by IMS. However, this reference implementation formed the basis of a product (the BlackBoard LMS) that competed with the other consortium offerings, resulting in the collapse of the first standards agreement.

IMS reorganised its efforts and offered a second set of standards based on XML document transfer rather than system APIs. These new standards had their own problems, however. Many of the new specifications offered little real interoperability as practically all aspects of the specification had become optional to accommodate the diverse capabilities of consortium members. Customers attempting to use the specifications to interoperate systems found that their vendors had implemented incompatible subsets of the specification that resulted in data and content transfer requiring costly manual transformation; the very thing standards had sought to eliminate. In response a number of application profiles were developed, the most well-known today being SCORM[2], to improve interoperability for particular purposes.

In some cases interoperability in practice did not match customer expectations. For example, the early implementations of IMS Content Packaging, the specification for open transfer of content by e-learning systems, used an approach one of the authors of this paper calls the 'white screen of lock-in' approach. This involves inserting between the open content manifest, and open (typically HTML) content a layer of proprietary XML metadata containing instructions to a specific system on how to load the content. Other systems importing the content see the table of contents, but as users click on items in the table all they see is a blank screen as the system renders the proprietary metadata instead of the content. This approach was used by both WebCT and Blackboard in their initial implementations; it may be the case here that neither company expected the specifications to be actually used for interoperability purposes, but simply wanted to assert 'conformance'. At this point in the development of the market it is also highly likely that most customers had just taken delivery of systems and were probably not very interested in ensuring they had a clear exit strategy, and were quite happy to take a conformance statement as sufficient evidence of goodwill in terms of future interoperability.

The issue of standards conformance and compliance has been a difficult one within the e-learning community, particularly with the number of competing application profiles developed. The general approach CETIS took was to take the pragmatic step of inviting vendors to demonstrate working interoperability with other partners within a closed environment, giving developers the opportunity to identify and fix issues before exposing interoperability problems to customers. An alternative approach was to take a more rigourous approach to the definition of application profiles with the intent of producing formal conformance tests, which was the subject of the TELCERT project. CETIS was also involved in the development of the RELOAD[3] tool to implement the IMS content specifications in a rigourous fashion to help users overcome interoperability issues. Today, many institutions use RELOAD to fix errors in standards-conformant content, or convert between incompatible implementations.

There has also been the claim from many smaller vendors that the standards developed by consortia (which often require annual membership fees for access) are themselves a form of lock-in. By releasing complex specifications that are difficult to implement, a barrier to entry is raised that only the largest vendors can afford to cross. This accusation has been levelled at a range of standards, most notably the Web Services specification stack promoted by Microsoft, Sun and BEA, which has swelled to an enormous volume of standards weighing in at thousands of pages. Whether this is a result of deliberate conspiracy or a rather monolithic development approach is moot; the overall effect has been that some developers have found WS-* excessively cumbersome and instead embraced various forms of simpler web services based on HTTP and XML (e.g. REST[4]) using simple proprietary API definitions. These proprietary lightweight APIs are the basis of many of the services considered part of "Web 2.0", such as del.icio.us, Blogger, and Google. It should be noted, however, that in another case, IMS QTI, smaller vendors were actually more able to implement a complex specification than the major vendors, so the argument that standards can be raised as a barrier to entry needs to be looked at critically.

[2] Sharable Content Object Reference Model. See http://www.adlnet.org/

[3] Reusable Learning Object Authoring and Delivery. See http://www.reload.ac.uk

[4] Representation State Transfer. An architectural model for web resources. See Fielding, 2000.

Another twist in the open standards story has been the issue of patents and IPR claims. While open standards are generally thought of as being free to use, this is conditional on the licensing of appropriate patents by contributing companies and the copyright policy of the standards organisation. In two recent cases, this has resulted in the 'encumbering' of open standards with patent issues. The first case involved the company ContentGuard, who were granted US patents for a range of technologies concerned with Digital Rights Management (DRM). ContentGuard actively engaged in the standardisation process through IEEE, developing the Open Digital Rights language (ODRL) in competition with their own XML Rights Management Language (XrML). However, they did this knowing that whichever technology customers used, they would still have to pay a license fee to ContentGuard, even if they chose to use the 'open' standard. The ContentGuard DRM patent situation has been the ongoing subject of legal disputes and commercial negotiations (Rosenblatt, 2005).

The second case involved the infamous '44 claims' of the Blackboard patent (see Feldstein, 2006, and Geist, 2006), which covers many of the features of modern e-learning systems, many of which were implemented by Blackboard at its inception as a result of implementing the first IMS specifications. Ironically, this then created the situation where vendors and open-source projects were then unsure whether adopting IMS specifications would also result in patent infringement. The patent issue, combined with the merger of Blackboard and WebCT into a single dominant vendor, have increased the pressure on institutions to create an exit strategy from their existing platform. Open standards should have made this far easier to accomplish this type of technology switch, which will be costly to implement for many institutions involving a large amount of content and data migration.

The use of patents as bargaining power, leverage, and influencer in open standards has been considered in other sectors, for example, Henrik Glimstedt's work on analysing the open standards process within the mobile telephony market (Glimstedt, 2001 & 2000). However, in educational technology patents in open standards have only recently become an important factor as a result of the Blackboard case.

While standards are a technology artifact, the process of constructing a standards involves an interplay of political and economic motives and is not simply a quest for an optimal technical solution. Where efforts on a particular axis are stalled or meet with opposition, a common tactic is for the proponents to find a new venue to pursue standardisation goals; a useful analysis of how the standards process involves the interplay of personal and organisation motives is given in zur Muehlen et al (2005) in their description of the evolution of open standards for workflow, and how various standards bodies have engaged in a sort of dance with various key players moving between organisations to pursue particular goals. In e-learning a similar interplay has been seen with new standards organisation proposed or created in response to the changing political or business context, such as HEKATE and LETSI. Krechmer (2005) set out a set of criteria for openness in standards, covering the areas of participation (open meetings, consensus, due process), dissemination (open IPR, open change, open documents) and usage (one world, open interface, ongoing support) which in practice are hard to reconcile with the practices of standardisation as seen in the organisations CETIS works with. While most specification bodeis have due process, an open IPR policy of some sort, and a one world (i.e. single international standard rather than regionalisation) approach, most do not support open meetings and instead favour a membership payment model. IMS, for example, decided in 2006 to delay releasing draft documents for public scrutiny to provide a competitive advantage for subscribing members; while understandable in terms of marketing membership fees, this violates Krechmer's 'Open documents' principle. Taken together, Krechmer's principles, applied in practice, show there is a great deal of interpretation possible for the meaning of 'open' in an 'open standard'.

To date, a substantial part of the effort of CETIS has been influencing the prevention of unnecessary or conflicting standards rather than the creation of desirable standards. An example of the type of case where standards prevention is necessary is where standardisation is initiated very early in the development of a technology, in a situation where adoption of a standard would genuinely impact on innovation (this is unusual; mostly, the opposite is true, as standards unlock opportunities to innovate). While early standardisation can be very tempting as a 'land grab' technique by pioneers in new types of applications, it can ultimately be damaging to the healthy diversity of solutions on offer as it sets, rather than an interoperability specification, a de jure dominant design which prevents entry into the market by alternatives (Abernathy and Utterbeck, 1978). The e-learning standards area was dominated early on by what Baskin, Krechmer and Sherif (1998) call anticipatory standards: "standards that must be created before widespread acceptance of the device or services", rather than responsive or participatory standards. This can be interpreted as "whoever defines the standard designs the future", and provides a temptation to develop standards prematurely.

While there are known caveats and issues in the area of open standards, there have also been some remarkable successes achieved as a result; understanding the critical success factors involved in open standards is an ongoing effort by CETIS. Tim Bray, one of the original developers of XML, considers that the number of successful XML-based standards is very small, and that 5 critical standards (HTML, DocBook, ODF, UBL, and Atom) form the core of achievement in XML standards to date (Bray, 2006). In other sectors, such as mobile telephony, there has been a considerable body of research on the standards process and its contribution to the mobile telephony market (see, e.g., Glimstedt, 2000, 2001; Pfannes provides an excellent overview of sources).

This complex story has informed the evolution of the approach to open standards taken by CETIS, which since procurement as a JISC service in 2006 (as the JISC-CETIS Service) has moved away from promoting adoption of open standards in a fairly unambiguous way to explicitly supporting a more complex message on interoperability. While the goal of interoperability has remained the same, and is at the heart of the strategy of the JISC-CETIS Service, the means by which interoperability is achieved is now seen by JISC-CETIS as having a number of strands and strategies, only some of which involve the use of open standards. The new multi-faceted approach sees a role for a range of technology interventions to achieve interoperability:

- adoption of open standards to exchange data and content;
- adoption of common infrastructure, such as the emergence of de-facto common libraries and open-source platforms;
- common implementation patterns and conventions that make it easier to engineer interoperating solutions;
- post-hoc interoperability achieved using latent semantic analysis and other techniques to analyse proprietary systems and their data;
- proprietary but publicly-documented interfaces;
- open processes and communications that support a dialogue about interoperability;
- adoption of emerging standards and patterns from communities of practice.

Some of these new strands have been added to the JISC-CETIS strategy as a result of observing the development of working interoperability within Web 2.0, where the standards process has, if anything, been even more convoluted and compromised than in the education sector (there are, for example, somewhere from 7 to 9 known variants of 'RSS'; see Pilgrim, 2004). The interoperability that has been achieved using the basic approach of 'Simple, Sloppy, and Scalable' (as Google's Adam Bosworth puts it; see Steinberg, 2005) has been highly successful and enabled large numbers of new services and initiatives. By contrast, the e-learning sector has seen a long period of consolidation with relatively little innovation but increasing costs. In some cases it may be argued that the prevalence of open standards may have actually reduced practical interoperability; for example, the existence of learning object specifications such as SCORM and IEEE Learning Object Metadata, and their place in mandated conformance and procurement regimes, may have negatively impacted the uptake of content syndication formats (RSS, Atom) in education.

This broader approach to interoperability seems to offer a much greater prospect of lasting impact than a purely standards-based approach, as it enables JISC-CETIS to engage with a wider range of communities and stakeholders and to try different strategies to meet particular needs. For example, it allows JISC-CETIS to engage with open-source initiatives such as Moodle and LAMS in a more balanced way in terms of their overall impact and value, rather than keeping a standards-conformance scorecard as a simplistic measure of positive impact. It also offers a more pragmatic basis to look at the role of Web 2.0 services, and wholly proprietary developments such as Second Life, in the evolving picture of e-learning technology.

The CETIS/JISC-CETIS experience represents an evolution in the organisation's understanding of the concept of 'openness' in terms of interoperability as an interplay of many factors. The net result of this new pragmatism is to focus attention on the desired state and the role in which 'openness', in various forms, contributes to progression towards it. Rather than recommending that organisations mandate open standards and enforce their conformance, JISC-CETIS instead encourages interoperability conversations and convergence on common approaches, backed up by simple functional evaluations of interoperability in practice.

# 7 Open Source: The OSSWatch Experience

Early evidence demonstrated that open source software was used in UK higher and further education institutions in advance of any advisory service being set up by JISC[5]. Much as expected, OSS Watch's initial scoping study in 2003 revealed a mixed economy. No institution was maintaining an exclusively proprietary nor exclusively open source environment. That raised a number of questions for a new advisory service.

Why were institutions turning to open source solutions? Institutional policy in this area notoriously lags behind practice. OSS Watch's 2006 survey, for example, found that less than 25% of institutions had any mention of open source in their IT policies (OSS-Watch, 2006). Yet more than 75% investigated open source solutions at every viable opportunity.

The top three reasons that institutions gave for considering open source software (in 2003) were: interoperability, cost, and security. Interoperability, in particular, was a surprise. However, in retrospect it seems clear that the tendency for open source software to conform to open standards was already beginning to reap benefits with the infrastructure IT stack. This connection between open source and open standards needed elaboration if unbiased advice and guidance was to be provided to universities and colleges in the UK.

One challenge that we face initially is purely definitional. What is 'open source software'? There are competing useful guides. The earliest and more philosophically driven movement is free software movement, led by the Free Software Foundation and its Free Software Definition (Free Software Foundation, 2005). A related but less ideologically motivated option is the Open Source Initiative's Open Source Definition (OSI, 2006). The latter is, of course, based on the Debian Free Software Guidelines (Debian Project, 2004). In addition to these there are numerous other more local variations. However, since OSS Watch was established by its funders as an open source software advisory service, it seemed most sensible to accept the OSI's definition of open source software. Thus in numerous places on the OSS Watch site you will find a clear statement that, “For OSS Watch open source software is always software released under an Open Source Initiative (OSI) certified licence” (OSS-Watch, 2005).

A clear and consistent statement of what open source software is, however, does not require suppression of alternate characterisations of free software. OSS Watch regularly makes reference especially to the Free Software Definition and encourages institutions to become familiar with the differences in language and intent between the significant groups in this space. Universities and colleges engaging with free and open source software in a sensible fashion cannot be shielded from the complexities of their own engagement. On the other hand, dealing with these complexities head on can alleviate some of the anxiety they may generate for those less certain of their grounding here.

In some respects open source software is better placed, definitionally, than open standards. There appears to be universal agreement that the Open Source Initiative is the maintainer of the Open Source Definition, even if some vendors do not feel bound by the need to pursue OSI certification for licences they describe as “open source” under which they release some or all of their software. For a time this practice can weaken the clarity that an advisory service can provide in its advice and guidance. Fortunately, the open source community is such that most high-profile vendors flouting the norms of the “open source” appellation find the negative public relations it generates to be counter-productive. Recently a number of such companies have reformed their practices and can now be acknowledged as open source companies even by OSS Watch[6].

Whether a project is using an OSI-certified licence is important. It underwrites what can usefully be said about its licensing conditions in the absence of additional paid legal advice. Institutions involved in procurement exercises are not typically interested in software that requires additional legal advice to know what can and cannot be done with it or how it can be further developed, in the case where the source code is provided. This might be one explanation for the slow take up in the UK of the Bodington virtual learning environment (VLE) as against Moodle. Although Bodington was “open sourced” by its home development institution, the University of Leeds, the licence placed upon it was not OSI-certified. This created a challenge since it could not be proclaimed as open source software by those with a strict adherence to OSI-certification as the key marker of open source software. It took some years for the Bodington community to sort this licensing issue satisfactorily (OSS-Watch,

[5] See OSS Watch Scoping Study, 2003, http://www.oss-watch.ac.uk/studies/scoping/

[6] Notable here is Alfresco's move to a GNU GPL release of its principal codebase (see http://www.alfresco.com/legal/licensing/whitepaper/). A smaller example, but one prominent in the university web content management market is Squiz.net's MySource Matrix (see http://matrix.squiz.net/evaluations/licence/choosing-gpl-or-ssv).

2006b). In the interim Moodle, which is released under the GNU GPL, was able to increase its market share in UK further education colleges to approximately 56% (OSS-Watch, 2006).

However, although an OSI-certified licence is important, it is not the sole determining of software suitability. OSS Watch therefore avoids making specific software recommendations. Instead the principal task is to help universities and colleges understand legal, social, technical and economic issues that arise when they engage with free and open source software. The goal is not the promotion of open source software for its own sake. Indeed, for OSS Watch the choice of proprietary or open source solutions is immaterial. What matters is that institutions have the resources to think through their procurement, deployment, or development IT concerns in a sensible and rational fashion. The best solution for any single institution will depend upon local conditions and individual needs.

This pragmatic approach to advice and guidance is consistent with that employed by UKOLN in its work on standards. It is also a guiding principle in the JISC Policy on Open source software for JISC projects and services (JISC, 2005). This policy is based on the UK government policy in this area and should be seen as an implementation of that policy[7]. Neither the government policy nor the JISC policy mandate open source software for deployment or open source licensing for release of development outputs. Rather, both policies draw attention to open source as one possible exploitation route for software which has been developed with government funds. The JISC policy goes further, providing useful guidance notes for those projects wishing to take up an open source development methodology (see, e.g., Raymond, 1997, and Fogel, 2005).

OSS Watch works closely with JISC-funded development project to aid their understanding of open source development methodologies. Since the JISC policy essentially urges projects to "get their IPR house in order at the earliest possible time", early consultation meetings using involve discussions around licence choice. Again a pragmatic approach rises to the top. Licence choice for software development project can be a fraught affair. The tendency to simply choose the licence you have heard most often mentioned is disconcerting. Without presuming to provide legal advice, OSS Watch helps projects think through the options available to them. In the end the choice will remain entirely in their hands, but issues such as compatibility with other code, potential for developing a community around the project, and an initial long term sustainability plan will certainly be explored.

## 8 A Contextual Approach

We have described some of the limitations of open standards and the feedback we have received from those seeking to make use of open standards in their development work. We have also described the experience of using open source. However, this need not mean an abandonment of a commitment to seek to exploit the benefits of open standards or open source. Nor should it mean imposing a stricter regime for ensuring compliance. Experience has made it clear that there is a need to adopt a culture, which is supportive of use of open standards and open source but provides flexibility to cater for the difficulties in achieving this.

This culture and approach is based on:

- A contextual model which recognizes the diversity and complexities of the technical, development and funding environments;
- A process of learning and refinement from patterns of successful and unsuccessful experiences;
- A support infrastructure based on openness, such as use of Creative Commons to encourage take-up of support materials and address the maintenance and sustainability of such resources.

It is apparent that there is a need to recognise the contextual nature to this problem; i.e. there is not a universal solution, but we should try to recognise local, regional and cultural factors, which will inform the selection and use of open standards.

Over time, in response to the problems outlined, the authors and others have developed a layered approach towards open standards intended for use in development work (Kelly, 2005). This approach is illustrated in Figure 1.

[7] See http://www.govtalk.gov.uk/policydocs/policydocs_document.asp?docnum=905

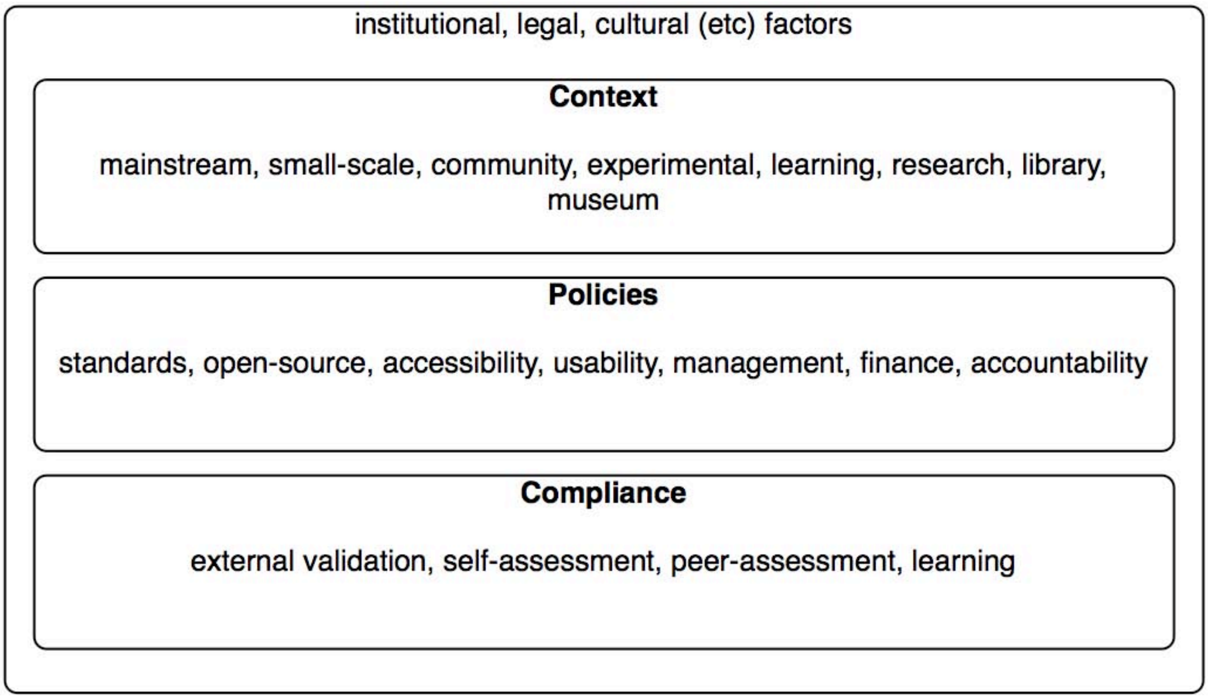

institutional, legal, cultural (etc) factors
Context
mainstream, small-scale, community, experimental, learning, research, library, museum
Policies
standards, open-source, accessibility, usability, management, finance, accountability
Compliance
external validation, self-assessment, peer-assessment, learning

**Figure 1: A Layered Approach to Use of Standards and Open Source**

This approach uses the following layers:

> **Contextual Layer:** This reflects the context in which the standards or open source software are being used. Large, well-funded organisations may choose to mandate strict use of open standards in order to build large, well-integrated systems which are intended for long term use. For a smaller organisation, perhaps reliant on volunteer effort with uncertain long-term viability, a simpler approach may be more appropriate, perhaps making use of proprietary solutions;
>
> **Policy Layer:** This provides an annotated description (or catalogue) of relevant policies in a range of areas, including open standards, open source, accessibility and accountability. The areas will include descriptions of standards, the ownership, maturity, risk assessment, etc. It summarises the strengths and weaknesses of the standards;
>
> **Compliance Layer:** This describes mechanisms to ensure that development work complies with the requirements defined within the particular context. For large, public funded programmes there could be a formal monitoring process carried out by external auditors. In other contexts, projects may be expected to carry out their own self-assessment, or take part in peer-assessment with related projects. In such cases, the findings could be simply used internally within the project, or, alternatively, significant deviations from best practices could be required to be reported to the funding body.

It should be noted that, although it is possible to deploy this three-layered approach within a funding programme or community, there will be a need to recognise external factors, over which there may be no direct control. This may include legal factors, wider organisational factors (for example there are differences between higher and further education, museums, libraries and archives), cultural factors, and available funding and resources etc.

It is also important to note that the contextual approach is not intended to provide an excuse to continue to make use of proprietary solutions which may fail to provide the required interoperability. Rather the approach seeks to ensure that a pragmatic approach is taken and that lessons can be learnt from the experiences gained. In order to ensure that the experiences are shared across the development community (and more widely) it will be important to ensure that systematic procedures are in place to ensure that the experiences are properly recorded and that such experiences are widely disseminated.

A requirement that funded projects should document their decisions on the selection of standards, open source licenses, and open source software, and provide reports based on their experiences in their use will help to ensure that such information is recorded in a systematic way, providing this information in an open and easily accessed fashion will help ensure that such information can be widely disseminated. The use of a Wiki, with RSS to allow the content to be syndicated and news of changes to the information, can help to support this.

After the selection and deployment of standards there will be a need to ensure that the standards are being used in an appropriate fashion. One means of ensuring that this happens is the use of a quality assurance framework. A similar approach may also be suitable, with minor modifications, for the selection of open source software, and open source licenses for development outputs.

## 9 Supporting a Contextual Approach

The provision and implementation of a model which provides a pragmatic approach to the selection and use of standards will not guarantee that appropriate decisions are made and that the selected standards are deployed in the most appropriate fashion. There also needs to be a support infrastructure in place which ensures that technical managers, implementers, designers and others involved in research and development activities are able to make technical decisions which are appropriate for the intended purpose.

A support model which is being developed is illustrated in Figure 2.

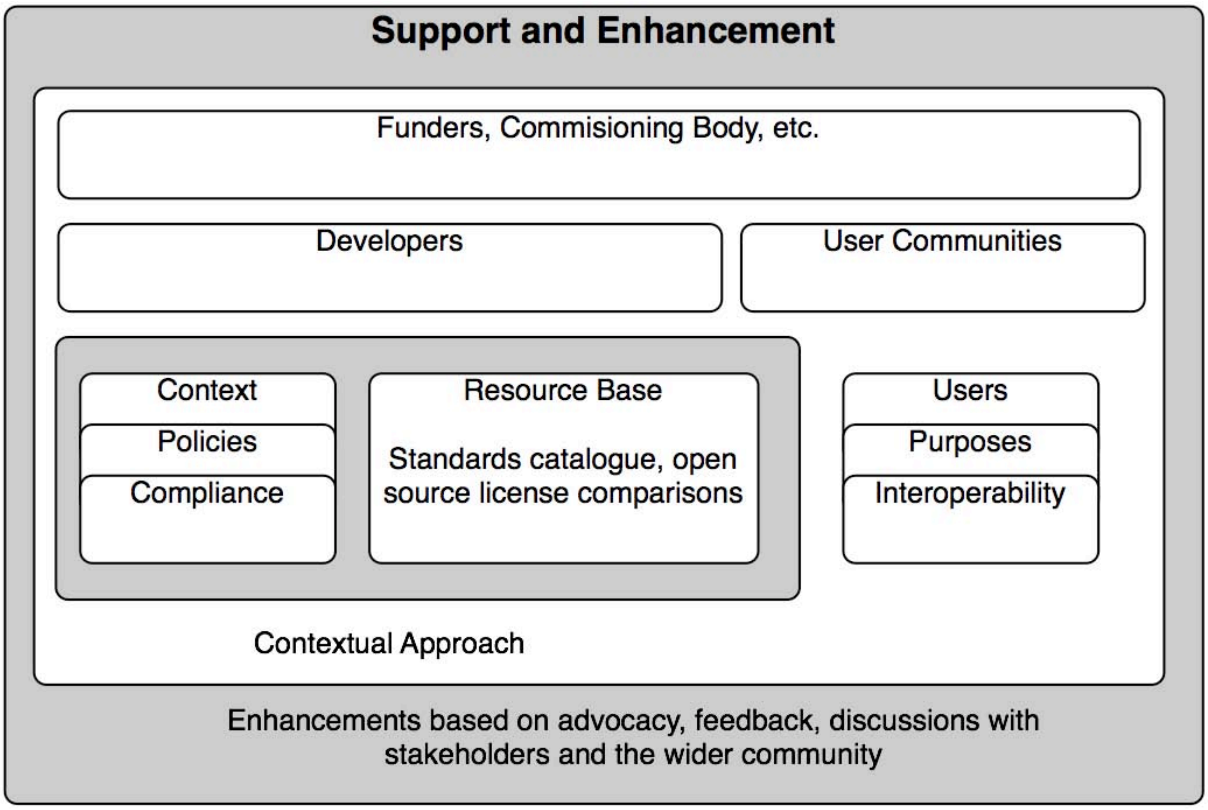

Support and Enhancement
Funders, Commisioning Body, etc.
Developers
User Communities
Context
Policies
Compliance
Resource Base
Standards catalogue, open source license comparisons
Users
Purposes
Interoperability
Contextual Approach
Enhancements based on advocacy, feedback, discussions with stakeholders and the wider community

**Figure 2: Support Model For Use of Standards**

This support model is based on the following features:

**The contextual model**: This is described elsewhere in this paper. It should be noted that the contextual model primarily intended for use by the development community. The end user community need not be aware of the contextual model that was used as part of the development process;

**User engagement**: Engagement with the user community will be essential to ensure the sustainability of the approach – it needs to be remembered that the development approach is not an end in itself, but a means for satisfying the needs of the user community.

There are several user communities involved in development activities. The *development community* will typically focus on areas related to the standards, development approach and related areas. The *user community*, in contrast, will often be disinterested in such issues, concerned primarily with use of a service which functions effectively. Although developers should be aware of the needs to address end user needs, it may be difficult to achieve this goal. It should therefore be a requirement of the *funding body* or organisation which has sponsored development work to ensure that mechanisms are put in place which will ensure that the approaches taken in development will ensure that the needs of the user community are satisfied. In the e-learning space, JISC-CETIS provides a range of Special Interest Groups (SIGs) that have a focus within a particular domain or context, where there is an effort on the part of the organisation to bring together developers and users to promote a better contextual awareness of the role of open standards.

Mechanisms for ensuring the development work is successful in meeting user needs may include:

> **Advocacy**: There will be a need for the development community to promote the advantages of the preferred approaches to development. This could include promoting the advantages of use of open standards. Such advocacy needs to be tailored for the intended target audience, with other developers and end users requiring different approaches;
>
> **Feedback**: A wide range of feedback will be required. For example, developers will need to provide detailed feedback on the contents of the resource base, funding agencies on the contextual model and implementation experiences, and end users on the end user service;
>
> **Engagement**: A passive feedback mechanism is unlikely to provide useful feedback. A more effective approach would be to provide more engaging mechanisms that act not only as a one-way transfer of information, but provide richer two-way discussions;
>
> **Refinement**: The feedback and engagement processes should help to refine those areas in which deficiencies have been identified. This could include over-simplistic or over-complex approaches to the development model.

## 10 Towards a Contextual Approach to Open Access?

So far in this paper we have looked in some detail at the experiences of advisory services in the adoption and use of open standards and open source software, and how this has lead to the development of a contextual approach and support services to assist developers, agencies and users. How applicable is this work to the promotion of open access?

As with open source and open standards, open access is again clearly a “good thing” in principle, that in practice requires an understanding of the context of use, the policy framework within which the organisation operates, and an understanding of the measures that can be used to assess whether open access – or, perhaps, more accurately, the benefits intended to be realised using open access – have actually been achieved in practice. For example, the “green” and “gold” open access options (Harnad, 2004) could be treated in a similar way to the various approaches to open source licensing, and to choices of open standards. The contextual model would offer a resource base providing detailed information on each approach, a connection to the policy context (e.g. mandates), access to communities where experience has already been gathered on use, and a set of measures for conformance, such as community peer review and availability of outcomes for public scutiny.

A support strategy for open access may use similar mechanisms to those for open source and open access, including advocacy, feedback, and refinement of the resource base in light of user experience and the active engagement and support of a community of use.

In the areas of open standards and open source we introduced the idea of transparency in the decision making process as part of the strategy for a pragmatic approach to adoption. In the case of open access, this would mean organisations and projects publicly documenting their decision on which open access strategy to adopt, or whether not to adopt an open access approach.

As noted earlier, it is also important to note that the contextual approach is not intended to provide an excuse to continue to not support open access. Rather the approach seeks to ensure that a pragmatic approach is taken and that lessons can be learnt from the experiences gained. For example, where existing open access strategies do not meet the requirements of particular contexts, and how new or hybrid strategies can be identified that better suit those contexts.

## 11 Conclusions

This paper has argued that what is needed is a more contextual approach to the open standards. It could be argued that what we need is not a list of open standards or open source licenses, or open access approaches but an *process for adopting open approaches* which is based on a desire to exploit the potential benefits of open standards, open source and open access, tempered by a degree of flexibility to ensure that the importance of satisfying end users needs and requirements is not lost and that over-complex solutions are avoided. This process could adopt the contextual approach documented in this paper and watch patterns of usage.

## References

[1] ABERNATHY, W. J.; UTTERBACK, J. M. (1978), Patterns of Industrial Restructuring. *Technology Review, 80 (7)*, 1-9.

[2] BASKIN, E.; KRECHMER, K.; SHERIF, M.H., (1998). The Six Dimensions Of Standards: Contribution Towards A Theory Of Standardization. In Louis Lefebvre, A., Mason, R., and Khalil, T. (1998, eds.), *Management of Technology, Sustainable Development and Eco-Efficiency*, Elsevier Press, Amsterdam, p. 53.

[3] BRAY, T., (2006). Don't Invent XML Languages. *Ongoing(weblog)*. Retrieved from http://www.tbray.org/ongoing/When/200x/2006/01/08/No-New-XML-Languages

[4] Debian Project, (2004). *Debian Free Software Guidelines, v1.1*. Retrieved from http://www.debian.org/social_contract#guidelines

[5] eLib (1996), *eLib Standards Guidelines*. Retrieved from http://www.ukoln.ac.uk/services/elib/papers/other/standards/

[6] FELDSTEIN, M., (2006). The Blackboard Patent Claims in Plain English. *e-Literate(weblog)*. Retrieved from http://mfeldstein.com/the_blackboard_patent_claims_in_plain_english/

[7] FIELDING, R.T., (2000). *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, UC Irvine, 2000

[8] FOGEL, K., (2005). *Producing Open Source Software*. Retrieved from http://producingoss.com/

[9] Free Software Foundation, (2005). *The Free Software Definition*. Retrieved from http://www.fsf.org/licensing/essays/free-sw.html

[10] GEIST, M.,(2006). Patent battle over teaching tools. *BBC News, August 14, 2006*. Retrieved from http://news.bbc.co.uk/2/hi/technology/4790485.stm

[11] GLIMSTEDT, H., (2000). Politics of Open Standards, Modular Innovation, and the Geography of Strategic Patenting in GSM and UMTS Technologies. *Division of Innovation, Lund Institute of Technology*. Retrieved from http://www.innovation.lth.se/files/Glimstedt%20April%2014th.pdf

[12] GLIMSTEDT, H., (2001). Competitive Dynamics Of Technological Standardization: The Case Of Third Generation Cellular Communications. *Industry and Innovation, 8(1), 49-78. Routledge.*

[13] HARNAD, S.; BRODY, T.; VALLIERES, F.; CARR, L.; HITCHCOCK, S.; GINGRAS, Y; OPPENHEIM, C.; STAMERJOHANNS, H.; HILF, E., (2004). "The green and the gold roads to Open Access," *Nature Web Focus*, http://www.nature.com/nature/focus/accessdebate/21.html

[14] JISC, (2001), *Standards and Guidelines to Build a National Resource*. Retrieved from http://www.jisc.ac.uk/fundingopportunities/projman_standards.aspx

[15] JISC, (2005), *Policy on Open source software for JISC projects and services*. Retrieved from http://www.jisc.ac.uk/fundingopportunities/open_source_policy.aspx

[16] Kelly, B., Guy, M., and James, H., (2003). Developing A Quality Culture For Digital Library Programmes. *Informatica 27(3) Oct. 2003*.

[17] KELLY, B.; RUSSELL, R.; JOHNSTON, P.; DUNNING, A.; HOLLINS, P.; PHIPPS, L. (2005). *A Standards Framework For Digital Library Programmes*. ichim05 Conference Proceedings. Retrieved from http://www.ukoln.ac.uk/web-focus/papers/ichim05/

[18] KRECHMER, K. (2005). Open Standards Requirements. *The International Journal of IT Standards and Standardization Research, 4(1), January - June 2006.*

[19] ZUR MUEHLEN, M.; NICKERSON, J.V.; SWENSON, K.D. (2005). Developing Web Services Choreography Standards - The Case of REST vs. SOAP. *Decision Support Systems 40 (2005) 1, pp. 9-29.*

[20] NOF (2001), *NOF-digitise Technical Standards And Guidelines*. Retrieved from http://www.mla.gov.uk/resources/assets/T/technicalstandardsv5_pdf_7959.pdf

[21] OSI, (2006). *The Open Source Definition*. Retrieved from http://www.opensource.org/docs/osd

[22] OSS-Watch, (2005). *What is open source software?* Retrieved from http://www.oss-watch.ac.uk/resources/opensourcesoftware.xml

[23] OSS-Watch, (2006). *OSS Watch 2006 Survey*. Retrieved from http://www.oss-watch.ac.uk/studies/survey2006/execsummary.xml

[24] OSS-Watch, (2006b). *Bodington released under Apache License v2.0*. Retrieved from http://www.oss-watch.ac.uk/resources/bodington-open.xml

[25] PFANNES, P. (2002). *Strategic Levers in Standardization Processes in the Mobile Communication Industry*. Thesis, Center for Digital Technology and Management, Munich

[26] PILGRIM, M. (2005). *The myth of RSS compatibility*. DiveIntoMark (weblog). Retrieved from http://diveintomark.org/archives/2004/02/04/incompatible-rss

[27] RAYMOND, E., (1997). *The Cathedral and the Bazaar*. Retrieved from http://www.catb.org/~esr/writings/cathedral-bazaar/

[28] ROSENBLATT, B., (2005). Opposition Mounts to OMA DRM Patent Licensing Scheme. *DRM Watch, April 4, 2005*. Retrieved from http://www.drmwatch.com/standards/article.php/3495026

[29] STEINBERG, D.H., (2005). Bosworth's Web of Data. *O'Reilly Network, 2005*. Retrieved from http://www.onlamp.com/pub/a/onlamp/2005/04/22/bosworth.html

# Peer-to-Peer Networks as a Distribution and Publishing Model

*Jorn De Boever*

Centre for Usability Research, Department of Communication Science, K.U. Leuven
Parkstraat 45 (b 3605), 3000 Leuven, Belgium
e-mail: jorn.deboever@soc.kuleuven.be

## Abstract

Content publishing and distribution often occurs in a costly and inefficient manner via client/server networks. Client/server models exhibit negative network externalities in that each additional user causes additional costs by increasingly congesting the system through the consumption of scarce resources. In an era of increasing demand for and size of content, the traditional client/server model produces evidence of its restrictions in terms of cost efficiency and scalability. Content providers – such as publishers, the media industry and users – are exploring new distribution or publishing models that might address the flaws of client/server models. An increasing amount of user generated content, open access and open content initiatives offer content for free, in spite of the fact that the distribution and storing of this content is not free of charge. This reasoning explains the importance of examining innovative distribution models that possibly provide answers to the shortcomings of client/server systems. In some cases, peer-to-peer systems might provide solutions for the flaws of client/server models in that they are characterized by cost efficiency and scalability. The facts that users spend more time online, have an increasing amount of resources (e.g. bandwidth, CPU cycles, content, and storage capacity) at their disposal, store and consume more content and bandwidth, is the basis of the viability of peer-to-peer systems. Peer-to-peer is still associated with illegal copyright infringing activities, although there are several companies exploring new ways for legal and secure content distribution through peer-to-peer networks. In this paper, we try to offer a broad analysis of the opportunities and challenges of several peer-to-peer applications and architectures. We further elaborate criteria in order to understand when the implementation of a peer-to-peer system might be appropriate. These criteria go further than merely technical criteria in that they include social criteria as well, which are as important as the technical ones. If peer-to-peer systems turn out to be a success for content publishing, it may lead to new business models that change the way content is distributed.

**Keywords:** peer-to-peer; e-publishing; content distribution; classification

## 1 Introduction

Mass content distribution through the internet often occurs in an inefficient en costly manner. This problem, caused by the limited scalability and high costs of the client/server model, leads the internet to be still a medium of mainly texts and images.

The internet has been marked by a vivid evolution of commercialization during the last decades in that it has become a medium for the masses and it involves more than just information. During the nineties, the internet consisted mainly of client/server models which are uncomplicated methods to manage and control the distribution of content. Throughout the past years, several evolutions have emerged that enticed consumers into wanting more than purely text and images. Several aspects – such as the widespread penetration of broadband internet, higher reliability of connections, the evolution of compression technology, more storage capacity, more CPU power and a large amount of content residing on the personal computers of end-users – changed the way users consume the internet. Internet users are spending more time online and exchange more information and files. The combination of these aspects resulted in an increasing demand for multimedia content that contains e.g. audio and video. In other words besides text and images, people nowadays consume larger, more bandwidth consuming content such as audio and video as well. This shift towards larger content makes it difficult for publishers to gain profit via a client/server model. Several measurement studies of peer-to-peer networks provide evidence for the large and increasing amount of large files such as video that is being shared [1, 2]. Furthermore, users have become more active in a sense that they dispose of an increasing amount of digital tools to create and publish content themselves in a relatively easy way. We observe that users have more content stored at their hard disks that is not accessible to other users that might be interested in this content.

Although compression technologies already offer some possibilities, we still observe multimedia content on the internet to be limited and if available it often turns out to be of poor quality. The widespread penetration of the internet causes content providers to explore new distribution platforms that provide solutions for the disadvantages of the current models. Publishers, the media industry and end users are exploring systems and platforms to publish and distribute online services and content. On the one hand, publishers and media companies attach great importance to examining new innovative models to distribute their digital content in a scalable and cost efficient way. They consider this as a necessity because in the current client/server model, every new consumer implies additional costs. On the other hand, users are generating more content themselves and want accessible systems to publish their content. Hosting user generated content at little or no costs for the users often involves high expenses for the hosts. It is therefore necessary to examine new models to distribute user generated content at little costs.

Peer-to-peer is often associated with illegal file sharing because of the popularity of networks such as Napster, Gnutella, KaZaA and BitTorrent. Although these networks contain(ed) a significant amount of illegal activities, they have demonstrated the opportunities of this disruptive technology. Today, many existing file sharing companies are examining new ways for the legal distribution of content. Furthermore, new companies – like Kontiki, Qtrax and RawFlow – were established that exploit peer-to-peer characteristics for secure content distribution. The question remains whether these peer-to-peer systems will be a viable solution for publishers and consumers. This is why it is important that we provide a better understanding of the characteristics, threats and prospects of peer-to-peer. We argue that peer-to-peer systems possess the capability of turning the internet into a valuable multimedia channel that will provide content providers and users with a rich arsenal of content.

In this paper, we will first explore different types of peer-to-peer applications to provide an overview of the capabilities of these models and we will show that different types of applications are merging. Subsequently, different peer-to-peer architectures will be analyzed in order to provide a classification based on the degree of (de)centralization of the topology and the presence or absence of structured resource location. The appropriateness of peer-to-peer as a publishing model will be described in the final section. This includes an exploration of the question what criteria must be met for a peer-to-peer system to be a suitable application.

## 2 Characteristics and Challenges of Peer-to-Peer

It is important to understand the pros and cons of systems in order to be able to evaluate them. It is therefore necessary to provide an outline of the characteristics and threats of peer-to-peer systems so that one can comprehend the possibilities and advantages peer-to-peer offer in comparison with other models. There is still no generally acknowledged unambiguous definition of the concept peer-to-peer which causes a discussion about what can(not) be accepted as peer-to-peer. Several authors have tried to formulate their own definitions of peer-to-peer [e.g. 3-5]. In spite of these definitions, we still belief that the following definition is the most accurate and comprehensive: "*The term 'peer-to-peer' refers to a class of systems and applications that employ distributed resources to perform a function in a decentralized manner. The resources encompass computing power, data (storage and content), network bandwidth, and presence (computers, humans and other resources)*" [6]. The following principles lay the foundation of peer-to-peer: (1) resources are being shared within peer-to-peer systems, (2) the systems are partially or fully decentralized and (3) the systems are self organizing depending on the extent of (de)centralization [7]. Peer-to-peer systems make use of the unutilized resources of the peers for instance on the level of storage capacity, bandwidth, CPU cycles and content.

Peer-to-peer systems have often been described as the counterpart of client/server networks [8, 9]. In client/server systems, centralized servers manage and control the network, provide services and resources whereas the clients consume these resources. Several client/server networks can hardly meet the demand for resources because of an increasing number of users, higher bandwidth traffic and the arrival of a variety of applications. The major drawbacks of client/server systems in comparison with peer-to-peer is that the client/server models suffer from inefficient allocation of resources and limited scalability which can result in bottlenecks and eventually in single points of failure. Furthermore, additional users stand for additional costs as they consume more bandwidth of the system.

Nodes in peer-to-peer networks do not only act as clients, but they exhibit server functions as well. This is why nodes or peers have been described as servents (SERVer + cliENTS). As said, client/server networks are not scalable and are susceptible to bottlenecks and single points of failure whereas peer-to-peer networks are characterized by: scalability, decentralization, transient connectivity, cost efficiency, fault tolerance, self organization, sharing of resources and autonomy [10, 11]. Other components that often proof to be important in peer-to-peer networks are security, anonymity, resilience and efficiency of resource location [9]. In theory, peer-to-peer systems exhibit positive network externalities in a way that additional users add value to peer-to-peer networks by introducing extra resources in the system. In this way, users preserve the system and influence the functioning, performance and control of the network by making their resources available. Therefore, it is critical for a peer-to-peer system to be able to cope with the transient presence of nodes, network/computer failures and that the network is able to self organize itself in the absence of centralized coordinating components.

An important challenge for peer-to-peer networks is security [10]. Peer-to-peer systems add risks to the network by distributing control to unknown nodes or peers and it therefore requires new security treatments. Further, the unstable and transient connectivity of nodes has consequences for the availability of resources [12, 13]. The resources of nodes are no longer available to the community if they are offline. Moreover, most users are free riders in that they consume resources while not providing anything to the system [14, 15]. The fact that most nodes are mostly free riders and are only online for a limited period of time makes resource availability a critical factor in the viability of peer-to-peer systems. A reliable peer-to-peer system therefore must be able to detect and recover from failures, guarantee content availability and avoid single points of failure. Subsequently, scalability stands for both opportunities and threats of peer-to-peer [16, 17]. Scalability reveals itself in the load in terms of bandwidth and storage capacity, the number of nodes that can be reached, the number of hops to reach nodes, the amount of resources that can be consumed, etc. without interfering the system's performance. Several file sharing systems contain millions of nodes that send terabytes of data across the network. Although peer-to-peer systems are theoretically inherently scalable, it often turns out to be a major challenge in real terms as we will further explore in the section on architectures.

# 3 Applications

In this section, we will provide an outline of different existing and new peer-to-peer applications. In this way, we gain insight in the fact that peer-to-peer is more than just file sharing. During the last few years, applications are becoming more integrated so that it becomes harder to draw a line between different types of applications [18].

## 3.1 Communication: Instant Messaging and Telephony

The first category of applications encloses communication systems such as Instant Messaging (IM) and telephony. These applications furnish the infrastructure mainly for real time or synchronous communication among users [5, 17]. Communication systems try to avoid as much central control as possible in order to reduce costs and to improve fault tolerance. These systems often merge into integrated applications that provide collaborative tools and file sharing on top of communication.

IM is a type of application that can utilize peer-to-peer aspects for their services, which of course does not mean that all IM systems exploit peer-to-peer characteristics. Some IM applications function within a client/server model, whereas other systems – e.g. ICQ, AOL instant messaging and Yahoo Messenger – make use of centralized peer-to-peer systems. The topology of these systems consists usually of a centralized directory server that contains information such as which nodes are online and who might communicate with whom. The communication then takes place directly between peers without intervention of the server.

Voice over IP (VoIP) is another application domain of peer-to-peer systems and has become more widely known particularly because of Skype [19, 20]. Skype is a peer-to-peer VoIP application that provides telephony, IM and audio conferencing via a peer-to-peer system.

## 3.2 Grid Computing

It has been widely debated whether grid computing can be accepted as peer-to-peer. In either way, grid computing and peer-to-peer networks are both distributed systems that are build to share resources [5, 21]. Grid computing is the coordinated use of resources – computers, processor capacity, sensors, software, storage capacity and data – which is being shared within a dynamic and continuously changing group of individuals, institutions and resources [17, 22]. In contrast to peer-to-peer systems, grids stress the standardized, secure and coordinated sharing of resources with a better guarantee of Quality of Service (QoS). The philosophy behind grids, which is largely the same for peer-to-peer systems, is that we can generate an enormous capacity by coupling several computers and their peripheral equipment in a network. The comparison has often been made with electric power: it should be as easy to get resources from the internet as it is simple to draw electricity from a wall socket [17]. Peer-to-peer and grids might evolve into a convergence in which the benefits of grid computing (interoperability, security, QoS, and standardized infrastructures) and of peer-to-peer (fault tolerance, scalability and self organization) will be combined. SETI@home (Search for ExtraTerrestrial Intelligence) is probably the most referred project in this area [23]. The processing of the radio signals has been distributed to the personal computers of users to save costs. In SETI@home, a central server receives the data, split it into small units, distributes these units to the clients and coordinates further transactions. The clients process the data using their unused capacity and send the results back to the server. Other similar projects are e.g. FightAIDS@home, Distributed.net, Entropia, Genome@home and Folding@home.

## 3.3 Collaborative Tools

A third application domain, in which peer-to-peer has already proven its usefulness and value, consists of tools for users to collaborate on certain tasks within groups [18, 24]. This type of software pursues the collaboration of users, even if some of these users find themselves to be outside the corporate LAN. An example of a peer-to-peer groupware tool is Groove Virtual Office, which has attributes such as file sharing, document management, chat, agenda and discussion groups. Security issues such as integrity, authentication and authorization are more critical in a confidential business environment so as to malicious users don't have the possibility to access, read or change the information [25]. It is obvious as well for peer-to-peer groupware applications to have opportunities for e-learning purposes. Peer-to-peer groupware integrates several elements like IM and file sharing.

## 3.4 File Sharing and Content Distribution

Peer-to-peer content distribution is the most well-known application area of peer-to-peer systems and it contains file sharing systems (e.g. Napster, Gnutella, eDonkey), distributed storage applications (e.g. Freenet) and content delivery networks (e.g. Kontiki). These applications offer companies and users the possibility to publish, store and exchange files and other content [5, 6]. The hype of peer-to-peer file sharing started in 1999 with the arrival of Napster [3]. Napster demonstrated the opportunities of peer-to-peer file sharing which resulted in the development of new systems such as Gnutella, Freenet, KaZaA and BitTorrent. Androutsellis-Theotokis and Spinellis [5] make a distinction between file exchange systems and content publishing/storage systems. On the one hand, file exchange systems are little sophisticated file sharing applications such as the former Napster that only contains some basic functionality and mostly doesn't address issues such as resource availability and security. It is mostly this type of applications that appears in the news because of copyright infringements. On the other hand content publishing and storage applications are more elaborated systems to publish, distribute and store content. These applications focus more on aspects such as security, availability and authorization.

Peer-to-peer streaming is a specific type of content distribution and it probably represents the most successful 'legal' peer-to-peer application. The traditional streaming technologies, such as unicasting and multicasting, are characterized by the fact that additional consumers of the streaming imply more costs. High costs, bottlenecks, single points of failure, lack of scalability and poor quality of most streaming technologies causes e.g. internet television to be still in its infancy. Peer-to-peer streaming, that has some similarities with multicasting, might offer some solutions for these problems by providing cost efficiency, scalability and quality of content [26, 27]. In peer-to-peer streaming applications, clients act as servers as they send units of the stream to other clients in the network. Most commercial peer-to-peer streaming applications integrate centralized components in their architecture to control and secure the content distribution. Examples of peer-to-peer video and or audio streaming are: Rawflow, Octoshape, Coopnet, Splitstream, Peerstreaming and Abacast.

## 3.5 Wireless and Ubiquitous Peer-to-Peer

Peer-to-peer systems offer some opportunities for wireless systems, such as MANETs (Mobile Ad hoc NETworks), and varies from cellular networks to wireless LANs [28]. Wireless communication networks can be considered to be peer-to-peer if the signals are being transferred directly between the appliances. In comparison with personal computers, the capacity of wireless equipment – for instance storage capacity, content and bandwidth – are increasing as well which offers new opportunities for peer-to-peer systems to be applied on mobile phones, Smartphones, PDA and laptops. Wireless peer-to-peer applications evoke other issues than 'traditional' peer-to-peer systems such as the power of batteries and the location of apparatuses when users are moving. The mobility of users combined with a transient connectivity of nodes make that self organization is an even bigger challenge for wireless peer-to-peer systems. Other challenges are emanating from the following characteristics: (1) wireless resources such as storage capacity, bandwidth and processing power are still limited, (2) the performance and capacities are fluctuating and (3) the availability of resources is barely guaranteed without a centralized component. It is therefore necessary for mobile peer-to-peer systems to develop applications with an efficient search and location infrastructure and routing model as to avoid zigzag movements.

Furthermore peer-to-peer systems exhibit characteristics – e.g. self organization, sharing of resources, collaborating apparatuses – that seem to be similar to some aspects of ubiquitous computing [29]. Similar to peer-to-peer systems, ubiquitous computing architectures must cope with autonomous communicating systems that are marked by transient connectivity. The parallel features of peer-to-peer and ubiquitous computing make that it doesn't seem illogical to integrate these systems.

## 4 Classifying Peer-to-Peer Architectures

Given that peer-to-peer systems have several different features, we endorse the fact that there might be different ways to classify peer-to-peer architectures. We argue that most peer-to-peer architectures distinguish themselves from each other based on the extent of (de)centralization and on the presence of structure in object location and routing. Based on this we distinguish the following combinations: centralized unstructured, pure unstructured, hybrid unstructured and pure structured systems.

### 4.1 Degree of Decentralization

Systems might be considered as peer-to-peer when at least some elements are decentralized. The degree to which centralized en decentralized components are applied in the network can vary between systems. In other words, in contrast with what has been suggested in some definitions of peer-to-peer, not all peer-to-peer networks are completely decentralized. We make a distinction between centralized, pure decentralized and hybrid peer-to-peer topologies (Figure 1) [5, 30, 31].

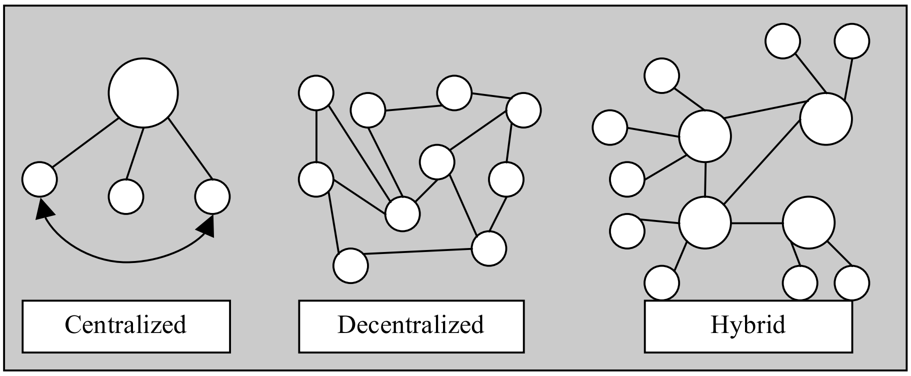


**Figure 1: Degree of Decentralization**

*Centralized* peer-to-peer architectures, such as the former Napster, contain a central server that executes vital functions for the system. This central server is mostly used as a directory server that stores an overview of the available nodes and resources in the network. In this way, the central directory server makes it possible for peers or nodes to find, locate and share resources with other peers. Peers eventually exchange data directly between each other without the intermediation of the server which makes it a simple but quite efficient architecture. This type of architecture exhibits the following drawbacks. The whole system stops functioning if the central servers cannot be reached for whatever reason. In other words, the major disadvantages of these systems are the risks of bottlenecks and single points of failure which imposes limited scalability. The advantage of using central directory servers is that if the sought data is available, the search algorithm can mostly guarantee the retrieval of the data.

*Pure decentralized* architectures consist of nodes that perform functions without the intervention of centralized components. These types of architectures have theoretically unbounded scalability and a high level of fault tolerance. In addition, these systems are autonomous and self organizing in a sense that the peers are responsible for the functioning and viability of the network. In practice, a great deal of these systems has limited scalability because self organization causes a lot of traffic to keep the network running. Another characteristic is that several of these systems have low levels of QoS in the domain of resource location because sometimes only a limited proportion of the network can be reached. Examples of pure decentralized peer-to-peer networks are Gnutella 0.4, Freenet and Chord.

*Hybrid* systems are often hierarchical networks that adopt elements of both centralized and pure decentralized architectures in which they combine the advantages (e.g. efficient resource location, scalability) while avoiding the drawbacks (e.g. bottlenecks, limited QoS) of these systems. In hybrid peer-to-peer systems, some peers have more capacities than others and therefore these peers are granted with more responsibilities. These nodes, that perform more functions in the network, are named super nodes or ultranodes. These super nodes are in fact regular nodes that also serve as a kind of directory server, with the difference that these super nodes are dynamic and can suddenly disappear from the network. In this way, nodes with better capacities have more responsibilities in the organization and functioning of the system whereas nodes with fewer capacities are less loaded. This reduces e.g. the possibility of delaying resource location because of links with dial-up connections. This leads to a better performance of the system because of less traffic and better search functionalities. The risks of bottlenecks and single points of failure are limited because of the use of several super nodes in stead of one central directory server. KaZaA and Gnutella 0.6 are some examples of suchlike architectures. Further, we argue that BitTorrent can also be regarded as a hybrid system.

## 4.2 Degree of Structure

Whether a system is structured depends on how nodes and data are positioned in the network [5, 8, 31].

*Unstructured.* A system is unstructured when nodes and data are positioned without certain rules and in an ad hoc manner in the network. The location of data is not connected with the topology of the network which results in cumbersome and little efficient search methodologies – such as the 'query flooding model' (cf. Gnutella) – that hamper scalability. On the one hand, most unstructured systems are characterized by much consumption of bandwidth in the matter of traffic of messages. Unstructured networks cannot guarantee that data, if available, can be found because the system is often not capable of reaching the whole network. On the other hand, these systems are mostly quite resilient. Another advantage is that these systems – e.g. Napster, Gnutella, KaZaA – mostly support keyword-based search.

*Structured.* In this type of networks, nodes and data are being placed in a structured way in the network as to be able to efficiently locate data which increases the possible scalability. The nodes, data or other resources are connected to specific locations. Distributed routing tables make it possible to efficiently, i.e. in a smaller number of hops, acquire search results. Structured systems are, in comparison with unstructured systems, more scalable, more reliable and fault tolerant. On the other hand, these systems have the disadvantage that it is more difficult to support keyword-based search because one needs to know the key to be able to locate the associated data. Another shortcoming is that these systems laboriously handle the transient connectivity of nodes whereby the system needs to reconfigure the structure constantly. Examples of structured systems are Chord, CAN, and Tapestry. Freenet is often called a 'loosely structured' network because it is not rigidly structured in that the location of the data is not totally specified.

## 4.3 Centralized Unstructured Systems

These peer-to-peer networks (e.g. Napster and Publius) have a centralized topology and display several client/server characteristics [16]. This type of peer-to-peer networks contains a central server that functions as a directory server [5, 32]. But, this directory server has fewer tasks than servers in client/server networks. In this way, a server in peer-to-peer networks is less loaded than servers in client/server networks. How does this system function? When peers log in to the system, they announce their presence and give some information (e.g. IP address, bandwidth of the connection, number and metadata of files that are being shared, etc.) to the directory server. In this way, there is one server that keeps an index of all available resources in the network. If a peer is searching for information, it sends a query to the server asking for available peers who share the requested information. The server subsequently searches his database and returns the result to the peer who initiated the query. Based on these results, the peer can decide to make a direct connection with a peer from the search results to download the requested data. In a nutshell, the search process is centralized, via the directory server, and the eventual exchange of data or other resources takes place in a peer-to-peer manner. The data is not stored on the server but on the hard disks of the peers.

The most important advantages of this type of architectures are that it is easy to implement and that the server is less loaded in terms of bandwidth and storage capacity. This is because the directory server doesn't vouch for the sending, distributing or storage of the data. Although this type of architectures use an unstructured search

infrastructure, they do provide good performance to be able to find the data if it is available [9, 11]. Another advantage of a centralized system is that it gives more opportunities to manage and control the network for security. Finally, these systems support keyword-based search which is important to users.
The disadvantages of the system mainly stem from the possible bottleneck at the server which is also hazardous for a single point of failure [11]. This has implications for the scalability of such systems because each additional user induces extra load in terms of traffic and storage capacity.

## 4.4 Pure Decentralized Unstructured Systems

The most striking feature of pure decentralized unstructured systems – e.g. Gnutella 0.4 – is that there is no centralized component which means that all nodes are directly connected to each other. Nodes function as clients, servers, routers and cache [5, 11]. They act as servers, not only for the storage and transfer of data, but also to search for data. Nodes can be involved as routers to help send messages through the network. Peers have an index of their own data and not of others' data. Therefore, to find the demanded data, it is important to be able to reach as many peers in the network as possible.

The advantage of this category of peer-to-peer architectures is that there is no single point of failure and that it is fault tolerant. The failure of one or even several of the nodes has little impact on the performance of the network. It is essential for these systems to be autonomous and self organizing in order to be able to cope with the transient connectivity of nodes. In the absence of a central infrastructure, the major challenge is to elaborate an efficient search method that is capable of achieving satisfying search results in the presence of transient nodes [9, 11]. Even if sought data is available in the network, unstructured peer-to-peer systems cannot offer guarantees that it would be able to find it. We will explain this with the example of the query flooding model. In the query flooding model, a node broadcasts a query to all his neighbors, his neighbors in their turn broadcast the query further to their own neighbors and so one. This process runs until a limited number of hops is reached according to the TTL (Time-To-Live). This TTL is essential to prevent messages from saturating the network by endlessly flooding it. But, this causes as well that the whole network is often impossible to reach which means that scalability is limited. Scarce content in a large file sharing network for example might be difficult to find because it is too many hops away. Of course several researchers have developed or adapted search methods to address these flaws such as: Random Walkers, Adaptive Probabilistic Search, Breadth First Search, etc. [for a more profound overview see: 33].

## 4.5 Hybrid Unstructured Systems

Hybrid unstructured peer-to-peer networks – such as KaZaA [34], eDonkey and Gnutella 0.6 [32] – have been developed with the objective of combining the advantages (e.g. better search results and fault tolerance) and circumventing the drawbacks (e.g. scalability and bottlenecks) of centralized and pure decentralized peer-to-peer systems [5, 16, 17, 35]. On the one hand, Napster had a centralized topology and they had to pull the plug on the server which ended the functioning of the whole network. This demonstrated the danger for bottlenecks and single points of failure in centralized topologies. On the other hand, Gnutella 0.4 as a pure decentralized system had to contend with an overload of messages because of the query flooding model. Hybrid peer-to-peer systems try to cope with these problems by introducing hierarchy in the system via the use of super nodes. Super nodes are peers with more capacity, such as bandwidth or storage capacity, than the average peer and therefore they are chosen to perform more functionality in the system. Super nodes mostly have the following tasks:

- Keep record of a directory list with information of a part of the peers and their data;
- Keep record of a directory list with information of some other super nodes;
- Search through the directory list in case a peer sends him a query;
- Redirect queries to other super nodes to be able to have better search results.

In other words, the hybrid architecture includes a combination of a centralized and a decentralized topology and therefore can be regarded as the convergence of these systems. There is not one central server, but there are different servers (super nodes) that all have responsibilities for a part of the node population. The super nodes are interrelated in a decentralized manner, whereas the normal nodes are related to their super node in a centralized way. The super nodes function as directory servers for a part of the peer-to-peer population.

The use of hierarchy in the system increases the chances for better, more efficient and faster search results because these systems utilize the available resources more intelligently [11]. The division of labor is more balanced in hybrid systems because nodes with more capacities get more responsibilities so as to nodes with less capacity don't get overloaded. Slow connections are avoided in this way, which results in an overall better performance of the system. Data and nodes are inserted in the network in an unstructured way, so that resource location also occurs in an unstructured way.

### 4.6 Pure Decentralized Structured Systems

Pure structured peer-to-peer systems – e.g. Chord, CAN, Freenet (loosely structured), Kademlia, Pastry – are self organizing networks without centralized components to store and retrieve data. If the content or other resources are available, unstructured networks offer guarantees that it will be able to find it within a limited number of hops [9, 11, 36]. These systems are structured because the resources and nodes are mapped into an address space in the network so as to be able to efficiently retrieve them. The indexing of this address space is distributed among the nodes in the system which makes every node responsible for a part of the indexing. These systems utilize Distributed Hash Tables (DHT) to structure the network: "*Distributed Hash Tables provide a global view of data distributed among many nodes, independent of the actual location*" [36]. DHTs manage the data in the system and it contains a routing system. Data can be efficiently retrieved because the DHT provides the system with a routing scheme to easily find the node that hosts the sought data. A unique key is created and assigned to every data to serve as an ID. The keys are mostly generated using hash functions such as SHA-1 and MD5. Hash functions are operations executed on data with a unique key as a result that has a fixed length regardless of the size of the data. The peers or nodes are responsible for a part of these keys in the address space. Peers are assigned with keys that most closely approach their own ID. It is important to know the unique key of the data to be able to retrieve it. During a search, the query is continuously redirected from node to node and it is getting closer to its destination or key in every hop. Every node in the system has a routing table of several other nodes in the network. A node that receives a query but doesn't have the sought information locally, routes the query to a node that, according to his routing table, is numerically closer to the destination.

There are two options to store and find data in this type of structured systems [36, 37]. In the first option, the nodes store only the unique keys of the data which serve as pointers to the actual data that is being stored somewhere else. The node that inserted the data is responsible for making the data available. In the second option, the nodes do not only store the keys, but the data as well. The second option implies that nodes store content that other nodes initiated. In this way, even if the node that initiated the data goes offline, the data remains available to other users because another node stores it.

It is a challenge for pure structured systems to maintain and update the routing tables in the transient presence of nodes. The updating process of the nodes' routing tables causes a load on the network. The resilience and structure of the network might be harmed if a large amount of the population would suddenly (dis)appear with a decreasing performance as a result. Another disadvantage is that these systems don't support keyword-based search because the search method requires the exact key [9, 11]. Further, a problem of load balancing might occur because nodes might be responsible for: (1) a big part of the address space, (2) a data rich address space and (3) very popular content [36]. Structured peer-to-peer systems exhibit the advantage that they have high levels of scalability and that they have an efficient search method that offers high guarantees for search results.

## 5 Appropriateness of Peer-to-Peer for e-Publishing and Openness

### 5.1 Peer-to-Peer Criteria

In this section, we try to elaborate some criteria to decide whether a peer-to-peer solution might be appropriate. We would first like to remark that these criteria are not meant to be formulas for success. One of the most important questions that has been posed is: when and for what types of content are peer-to-peer applications appropriate? It is difficult to formulate an unambiguous answer to the first part of the question. The second part of the question, which is for what types of content peer-to-peer is appropriate, is simpler to answer. In my opinion, peer-to-peer is content independent as it is distribution model and not a content model. Roussopoulos, Baker and Rosenthal [38] tried to answer the question when peer-to-peer might be appropriate. They formulated several criteria to determine whether the implementation of peer-to-peer aspects is the appropriate method for the distribution of a certain kind of content:

- Cost savings: peer-to-peer solutions make it possible for companies and other organizations with limited budget to distribute their content to the masses;
- Relevance of resources: the content must be important to the consumers so that they are more willing to participate in helping to distribute the content;

- Trust: the indispensable need to have peers cooperating is hard to achieve when peers distrust each other;
- Rate of change and criticality: A peer-to-peer application probably will not succeed if there is a high rate of change (peers entering and leaving the system) in an untrustworthy environment. Peer-to-peer has more chances for success when the rate of change is low and the criticality of the information is low.

These researchers formulated the following conclusion: "(…) *the characteristics that motivate a P2P solution are limited budget, high relevance of the resource, high trust between nodes, a low rate of system change, and a low criticality of the solution*" [38]. Although we do agree with the conclusions of these authors, we argue that the practice is more complex and that the appropriateness of peer-to-peer depends on technical (e.g. architecture) and social aspects. The applicability of peer-to-peer systems is independent of content type. We will first discuss the technical aspects to subsequently expand the social aspects.

From a technical perspective, peer-to-peer systems provide solutions for mass content distribution in that they are characterized by cost reduction, scalability and performance. Publishers and content providers might implement a peer-to-peer solution for cost saving objectives. Every additional consumer in a client/server model produces extra costs, whereas this effect is more limited in peer-to-peer networks because additional users mean extra resources for the system. Peer-to-peer systems are especially important for scalability solutions. The connection of the server in a client/server model becomes silted up little by little in an environment of an increasing user population which causes a bottleneck. Peer-to-peer systems can prevent the occurrence of bottlenecks by utilizing the available resources in the peer community. Excellent developed peer-to-peer applications give evidence of great performance in the presence of mass populations. From this, we can conclude that the three main reasons for choosing a peer-to-peer solution are cost of ownership, scalability and performance. Furthermore the architecture of peer-to-peer systems has certain aspects that make some topologies more suitable for specific applications. We explain this with some examples without having the aim to be exhaustive. The larger the consumer mass, the more a decentralized architecture is appropriate in order to avoid bottlenecks. Commercial content providers mostly want to preserve control over the distribution of content so that they are ensured of payment and they are capable of avoiding copyright infringement. Therefore, most commercial and legal peer-to-peer systems have a centralized topology. A centralized topology exposes itself to limited risks of bottlenecks and it is therefore suitable for a limited user population.

From a social perspective, a characteristic feature of peer-to-peer networks is that the performance of the system is not only dependent on technical functionality, but also on user behavior. A critical mass of active participating online peers and content availability are critical to the viability and success of peer-to-peer networks. The number of online peers and the number and quality of content these peers share, determine the value other peers can derive from the network. These last two sentences contain several features that determine whether a peer-to-peer solution might be appropriate:

- Critical mass: it is not necessary for all peers to provide the system with there resources, but it is necessary to have a sufficient amount of peers that contribute to the system so that the content remains available;
- Online: peers have to stay online after they have downloaded content so that others can download this and other content from them;
- Quantitative availability: it is important to have a sufficient amount of content available;
- Qualitative availability: it is not sufficient to have large amounts of content available, but it is critical as well to have resources available other peers are interested in.

From this analysis, we argue that content is suitable to be distributed via peer-to-peer networks if: (1) the distribution of the content is very resource consuming, (2) it is being consumed by a mass, (3) the consuming mass is online enough, and (4) there are enough users willing to cooperate in distributing the content. Publishers, consumers and media companies might consider using peer-to-peer networks if their content meets these criteria.

Critical readers may now wonder whether peer-to-peer is only appropriate for mass content distribution. The answer to that question is no, but there are some important conditions for peer-to-peer to be successful in the presence of small user communities. Peer-to-peer solutions might succeed if the population of peers is often online simultaneously. It is favorable to have a community of peers with strong ties and with similar interests in the sense that they consume the same kind of content. The importance to stimulate users to cooperate via incentives increases in small populations of peers in order to ensure content availability. A centralized peer-to-peer architecture might be an appropriate system for small user communities because the risks for bottlenecks at the server are more limited. A few examples of content that is mostly consumed by small communities are user generated content and content that is consumed within organizations (e.g. corporate communication).

## 5.2 Openness

Openness is a new buzz word that supports philosophies of open access [39, 40], open content and open source in which information, knowledge and content is universally available as a public good and is often for free. Whereas this content is mostly free to users, it is often expensive to the organizations hosting this content; e.g. the Budapest Open Access Initiative recognizes this assumption: "*While the peer-reviewed journal literature should be accessible online without cost to readers, it is not costless to produce*" [39]. Comtella (http://bistrica.usask.ca/madmuc/comtella.htm), LionShare (http://lionshare.its.psu.edu/) and Edutella (http://edutella.jxta.org/) are three peer-to-peer systems that support open sharing of educational content such as papers, articles and other educational and research tools (videos, images, presentations, demos, etc.). All these projects have been initiated with the objective of making free and open educational and research material more accessible to all interested persons. Why might peer-to-peer be appropriate for open initiatives? In our opinion, peer-to-peer might be applicable to 'open' environments if it meets several of the above mentioned criteria. The storage and distribution of content in open initiatives are often resource and consequently money consuming. There is a large simultaneous online population that possesses many unused resources. These two fundamental aspects, that are at least necessary for the implementation of a peer-to-peer system, are characteristic for several open initiatives. Whether there will be a critical mass of cooperating peers is hard to predict because it depends on the complex interaction of several factors. The main raison for using a peer-to-peer system is cost reduction on the level of storage capacity and bandwidth. If peer-to-peer systems turn out to be successful applications in open environments, it will result in decreasing storage and bandwidth expenses. Decreasing resource expenses reduce the barriers for open initiatives which in his turn can lead to more accessible content.

The internet originally displayed peer-to-peer characteristics in that every computer was mutually connected to other computers in the network and most computers acted as clients as well as servers [17, 41]. In those days, the internet was mainly used for research and military purposes. If the actual open movements will succeed in using peer-to-peer systems for research and educational purposes, then it might be regarded as the *renaissance* of the internet.

# 6 Discussion and Conclusions

Our analysis suggests that peer-to-peer systems in some cases might provide solutions for the flaws of client/server systems. Client/server models suffer from limited scalability, bottlenecks, cost inefficiency and single points of failure. These characteristics of client/server models set limits for the amount and largeness of available content on the internet. In this paper, we have demonstrated that in some cases, peer-to-peer systems might provide solutions for the drawbacks of client/server systems in that they have already proven their abilities in terms of scalability and cost efficiency. Further, we have shown that peer-to-peer comprises more than file sharing, such as communication, collaboration, and grid computing. The importance of characteristics and (dis)advantages of peer-to-peer systems varies from architecture to architecture depending on the degree of (de)centralization and whether it is structured or not. This can be represented as a pendulum between: (1) risks of bottlenecks, possible single point of failure, more control (centralized) and (2) scalability, fault tolerance, self organization (decentralized). Whether a peer-to-peer system is structured or not determines the efficiency of node and resource location, at which structured systems are more efficient.

One of the major questions of several content providers is when a peer-to-peer solution might be appropriate. Therefore, we tried to elaborate criteria to decide whether a peer-to-peer solution might be suitable. We have to remark that these criteria are not meant to be rules for success as it does not imply that users will adopt and use the system. These criteria imply that peer-to-peer is not always a good solution and that client/server systems will sustain. Besides more technical criteria such as scalability, we paid attention to some social criteria as well. Peer-to-peer systems are not only dependent on technical criteria, but also on social aspects for it are the users that make their resources available and cooperate or free ride in distributing content. More research on social aspects is needed because there is little information on user behavior in peer-to-peer systems. Social research is necessary because users have never had such a powerful impact on a system as the end users influence the performance of peer-to-peer networks by (not) providing their resources to the community. Whether peer-to-peer solutions might be appropriate for open initiatives depends on whether the system meets the aforementioned criteria. It seems likely that peer-to-peer is suitable in some open access and open content systems because scalability, cost efficiency and a large simultaneously online user population are all criteria that are often met in these applications.

The results of this broad analysis provide a better understanding of the capabilities and application domains of peer-to-peer systems. The internet today is being marked by an increasing amount of content and an increasing size of this content which causes more loads on the distribution, storage and consequently costs. That is why publishers and other media companies are trying to find solutions for scalability and cost problems and therefore need to explore innovative platforms such as peer-to-peer systems.

What will the future bring for peer-to-peer? There are several essential issues that need to be remedied in order for peer-to-peer to be able to succeed. There is still a lot of work to be done to address problems of standardization, security, Digital Rights Management and asymmetric connections with unbalanced upload/download ratios. The years 2006 and 2007 might become the turning point for peer-to-peer networks because this is the period that new 'legal' peer-to-peer services have entered the market [42]. Currently, mainly the opportunities of peer-to-peer for video, film and television are being explored by different companies (e.g. Joost, RawFlow, Kontiki, BitTorrent, In2Movies, etc.) and in different workshops [e.g. 43, 44]. Peer-to-peer television is one of the examples that meet the formulated criteria. But it is not all about video. Peer-to-peer is a distribution system in that it is content independent. It is remarkable that almost every 'legal' commercial peer-to-peer system implements centralized components in their architecture. This is probably to ensure control, security and QoS. In this way, these commercial peer-to-peer platforms combine the strengths of peer-to-peer systems with those of client/server models. A lot of non-technical questions remain unanswered, e.g. are users willing to cooperate in a network when they have to pay the consumed content and for what kind of content are they willing to pay. To learn more about the possibilities of peer-to-peer networks, it is essential that the research community explores new applications, environments, content to experiment with peer-to-peer. Peer-to-peer networks have the capacities for a scalable, accessible and cost efficient model for the distribution of content. If peer-to-peer systems turn out to be a success for content publishing, it may lead to new business models that change the way content is distributed. It is hard to predict whether peer-to-peer will become a success for legal purposes. Only the future will tell how peer-to-peer will evolve.

## Acknowledgements

FLEET (an interdisciplinary research project on FLEmish E-publishing Trends) is an IWT SBO project, with research partners IBBT SMIT, Cemeso, LSTS, MOFI, ICRI, CUO, MICT, ECDC and TNO.

## References

[1] OECD. Peer to Peer Networks in OECD Countries. 2004. http://www.oecd.org/dataoecd/55/57/32927686.pdf.

[2] KOLWEY, M.; LECHNER, U. Towards P2P Information Systems. *The Fifth International Workshop on Innovative Internet Community Systems: IICS 2005, Paris, France*, 2005.

[3] SHIRKY, C. Listening to Napster. In A. ORAM (Eds.), *Peer-to-Peer: Harnessing the Power of Disruptive Technologies*, Sebastopol, CA: O'Reilly & Associates, Inc, 2001, pp. 21-37.

[4] SCHOLLMEIER, R. A Definition of Peer-to-Peer Networking for the Classification of Peer-to-Peer Architectures and Applications [poster]. *The First International Conference on Peer-to-Peer Computing, Linköping, Sweden*, 2001.

[5] ANDROUTSELLIS-THEOTOKIS, S.; SPINELLIS, D. A Survey of Peer-to-Peer Content Distribution Technologies. *ACM Computing Surveys*, 2004, vol. 36, no. 4, pp. 335-371.

[6] MILOJICIC, D. S.; KALOGERAKI, V.; LUKOSE, R.; NAGARAJA, K.; PRUYNE, J.; RICHARD, B.; ROLLINS, S.; XU, Z. Peer-to-Peer Computing [Technical Report]. 2002. http://www.hpl.hp.com/techreports/2002/HPL-2002-57R1.pdf.

[7] ABERER, K.; HAUSWIRTH, M. An Overview on Peer-to-Peer Information Systems. *Workshop on Distributed Data and Structures, Paris, France.* 2002.

[8] STEINMETZ, R.; WEHRLE, K. What Is This "Peer-to-Peer" About? In R. STEINMETZ & K. WEHRLE (Eds.). *Peer-to-Peer Systems and Applications*, Berlin Heidelberg: Springer-Verlag, 2005, pp. 9-16.

[9] WANG, C.; LI, B. Peer-to-Peer Overlay Networks: A Survey. 2003. http://comp.uark.edu/~cgwang/Papers/TR-P2P.pdf.

[10] SCHODER, D.; FISCHBACH, K.; SCHMITT, C. Core Concepts in Peer-to-Peer Networking. In R. SUBRAMANIAN; GOODMAN, B.D. (Eds.), *Peer-to-Peer Computing: the Evolution of a Disruptive Technology*, London: Idea Group Publishing, 2005, pp. 1-27.

[11] SCHMIDT, C.; PARASHAR, M. Peer-to-Peer Information Storage and Discovery Systems. In R. SUBRAMANIAN; GOODMAN, B.D. (Eds.), *Peer-to-Peer Computing: the Evolution of a Disruptive Technology*, London: Idea Group Publishing, 2005, pp. 79-112.

[12] BHAGWAN, R.; SAVAGE, S.; VOELKER, G.M. Understanding Availability. *The Second International Workshop on Peer-to-Peer Systems, Berkeley, CA, USA*, 2003.

[13] CHU, J.; LABONTE, K.; LEVINE, B.N. Availability and Locality Measurements of Peer-to-Peer File Systems. *Conference on Scalability and Traffic Control in IP Networks, Boston, USA*, 2002.

[14] ADAR, E.; HUBERMAN, B.A. Free Riding on Gnutella. *First Monday*, 2000, vol. 5, no. 10, http://www.firstmonday.dk/issues/issue5_10/adar/.

[15] HANDURUKANDE, S.B.; KERMARREC, A.-M.; LE FESSANT, F.; MASSOULIÉ, L.; PATARIN, S. Peer Sharing Behaviour in the eDonkey Network, and Implications for the Design of Server-less File Sharing Systems. *EuroSys 2006, Leuven, Belgium*, 2006.

[16] DING, C.H.; NUTANONG, S.; BUYYA, R. Peer-to-Peer Networks for Content Sharing. In R. SUBRAMANIAN & B.D. GOODMAN (Eds.), *Peer-to-Peer Computing: the Evolution of a Disruptive Technology*, London: Idea Group Publishing, 2005, pp. 28-65.

[17] TAYLOR, I. J. *From P2P to Web Services and Grids: Peers in a Client/Server World*. London: Springer, 2004.

[18] SCHODER, D., FISCHBACH, K.; SCHMITT, C. Application Areas. In R. STEINMETZ & K. WEHRLE (Eds.). *Peer-to-Peer Systems and Applications*, Berlin Heidelberg: Springer-Verlag, 2005, pp. 25-32.

[19] BASET, S.A.; SCHULZRINNE, H. An Analysis of the Skype Peer-to-Peer Internet Telephony Protocol – Technical Report. 2004. http://www.rootsecure.net/content/downloads/pdf/skype_protocol.pdf.

[20] JENNINGS, C.; BRYAN, D.A. P2P For Communications: Beyond File Sharing. *Business Communications Review*, 2006, vol. 36, no. 2, pp. 36-40.

[21] TALIA, D.; TRUNFIO, P. Toward a Synergy between P2P and Grids. *IEEE Internet Computing*, 2003, vol. 7, no. 4, pp. 96, 94-95.

[22] FOSTER, I.; IAMNITCHI, A. On Death, Taxes, and the Convergence of Peer-to-Peer and Grid Computing. *The Second International Workshop on Peer-to-Peer Systems, Berkeley, CA, USA*, 2003.

[23] ANDERSON, D. SETI@home. In A. ORAM (Eds.), *Peer-to-Peer: Harnessing the Power of Disruptive Technologies*, Sebastopol, CA: O'Reilly & Associates, Inc, 2001, pp. 67-76.

[24] MAUTHE, A.; HUTCHISON, D. Peer-to-Peer Computing: Systems, Concepts and Characteristics. *Praxis in der Informationsverarbeitung und Kommunikation*, 2003, vol. 26, no. 2.

[25] ASVATHANARAYANAN, S. Potential Security Issues in a Peer-to-Peer Network from a Database Perspective. In R. SUBRAMANIAN & B.D. GOODMAN (Eds.), *Peer-to-Peer Computing: the Evolution of a Disruptive Technology*, London: Idea Group Publishing, 2005, pp. 131-144.

[26] LIU, Z.; YU, H.; KUNDUR, D.; MERABTI, M. On Peer-to-Peer Multimedia Content Access and Distribution. The *International Conference on Multimedia and Expo*, *Toronto, Canada*, 2006.

[27] STOLARZ, D. Peer-to-Peer Streaming Media Delivery. *The First International Conference on Peer-to-Peer Computing, Linköping, Sweden*, 2001.

[28] KELLERER, W.; SCHOLLMEIER, R.; WEHRLE, K. Peer-to-Peer in Mobile Environments. In R. STEINMETZ & K. WEHRLE (Eds.). *Peer-to-Peer Systems and Applications*, Berlin Heidelberg: Springer-Verlag, 2005, pp. 401-417.

[29] KANGASHARJU, J. Peer-to-Peer and Ubiquitous Computing. In R. STEINMETZ & K. WEHRLE (Eds.). *Peer-to-Peer Systems and Applications*, Berlin Heidelberg: Springer-Verlag, 2005, pp. 457-469.

[30] BACKX, P.; WAUTERS, T.; DHOEDT, B.; DEMEESTER, P. A comparison of peer-to-peer architectures. *Eurescom Summit 2002, Heidelberg, Germany*, 2002.

[31] POUREBRAHIMI, B.; BERTELS, K.; VASSILIADIS, S. A Survey of Peer-to-Peer Networks. *The 16th Annual Workshop on Circuits, Systems and Signal Processing, Veldhoven, the Netherlands*, 2005.

[32] EBERSPÄCHER, J.; SCHOLLMEIER, R. First and Second Generation of Peer-to-Peer Systems. In R. STEINMETZ; WEHRLE, K. (Eds.). *Peer-to-Peer Systems and Applications*, Berlin Heidelberg: Springer-Verlag, 2005, pp. 35-56.

[33] TSOUMAKOS, D.; ROUSSOPOULOS, N. A Comparison of Peer-to-Peer Search Methods. *The International Workshop on the Web and Databases, San Diego, Florida, USA*, 2003.

[34] LIANG, J.; KUMAR, R.; ROSS, K.W. Understanding KaZaA. 2004. http://cis.poly.edu/~ross/papers/UnderstandingKaZaA.pdf

[35] LEIBOWITZ, N.; RIPEANU, M.; WIERZBICKI, A. Deconstructing the Kazaa Network. *The Third IEEE Workshop on Internet Applications, San José, CA, USA*, 2003.

[36] WEHRLE, K.; GÖTZ, S.; RIECHE, S. Distributed Hash Tables. In R. STEINMETZ & K. WEHRLE (Eds.). *Peer-to-Peer Systems and Applications*, Berlin Heidelberg: Springer-Verlag, 2005, pp. 79-93.

[37] BALAKRISHNAN, H.; KAASHOEK, M.F.; KARGER, D.; MORRIS, R.; STOICA, I. Looking up Data in P2P Systems. *Communications of the ACM*, 2003, vol. 46, no. 2, pp. 43-48.

[38] ROUSSOPOULOS, M.; BAKER, M.; ROSENTHAL, D.S.H. 2 P2P or Not 2 P2P. *The Third International Workshop on Peer-to-Peer Systems, San Diego, USA*, 2004.

[39] BOAI. Budapest Open Access Initiative. 2002. http://www.soros.org/openaccess/read.shtml.

[40] JOHNSON, R.K. Open Access: Unlocking the Value of Scientific Research. *Journal of Library Administration*, 2004, vol. 42, no. 2, pp. 107-124.

[41] MINAR, N.; HEDLUND, M. A Network of Peers: Peer-to-Peer Models Through the History of the Internet. In A. ORAM (Eds.), *Peer-to-Peer: Harnessing the Power of Disruptive Technologies*, Sebastopol, CA: O'Reilly & Associates, Inc, 2001, pp. 3-21.

[42] WINSBURY, R. 2006 – The year that P2P comes in from the cold? Mass video broadcasting over the Internet. *Intermedia*, 2006, vol. 34, no. 2, pp. 12-17.

[43] ARNOLDUS, M. Creative Commons Nederland: Workshop on technical and legal aspects of peer-to-peer television. 2006. http://creativecommons.nl/nieuws/wp-content/uploads/2006/04/Verslag%20P2P-TV%20Workshop.pdf.

[44] KOZAMERNIK, F. EBU Seminar Report: From P2P to Broadcasting. 2006. http://www.ebu.ch/en/technical/trev/trev_306-p2p.pdf.

# A Lifeboat Doesn't Do You any Good if it's not There when You Need it: Open Access and its Place in the New Electronic Publishing Paradigm

*Ian M. Johnson*

Aberdeen Business School, The Robert Gordon University, Garthdee Road
Aberdeen AB10 7QE, Great Britain
e-mail: i.m.johnson@rgu.ac.uk

## Abstract

This paper draws on the results of recent research into digital publishing in Latin America sponsored by the European Commission's ALFA programme. It outlines the growth in publishing in the region. It aims to stimulate reflection on the impact of a system in which most of the publishing is supported by institutions rather than commercial companies, and considers authors' aspirations for their work to achieve recognition, attitudes towards peer review and other aspects of journal quality, the indexing and availability of full text journals, and the sustainability of institutionally supported publishing. Examples are drawn from publishing in the field of librarianship and information sciences on which the original research project was focused.

**Keywords:** Latin America; electronic journals; quality control; findability; sustainability

## 1 Introduction

> "*Our actions must embody these new 'realities' because even when people realize that they are on the Titanic and the iceberg is right ahead, we still need to see the lifeboat in order to jump ship*" [1]

Research into the diffusion of innovation notes the significance of communication channels in transferring awareness and understanding of innovations.[2] A key part of that process has been the scholarly peer-reviewed journal. We are currently in a state of transition in global scientific communication as the new Information and Communication Technologies are becoming more generally available and more powerful. There is no denying the ability of the Internet to disseminate information rapidly, and it is rapidly being accepted that online access to the full text of scholarly papers should become the norm. However, access to 'free' information on the Web has stimulated a perception that the traditional scholarly journals might be replaced by new services compiled, edited, produced, marketed and distributed without the intermediary services of a publisher, a concept that has been promoted by the emergence of pre-print repositories and of electronic journals produced by individuals. The potential of these new models has proved particularly attractive to researchers and their funders, who had become concerned about the so-called 'scholarly communication crisis', and to librarians who have become concerned about the 'serials crisis.'

In the transition between paper and electronic publishing, discontent about the way in which information is handled is rising, and new experimental models are bound to capture attention. Like most phenomena where one order has to be replaced by a new order, a certain amount of disorder or chaos is inevitable in the transition phase. The 'noise' from the chaos has inevitably reached the ears of governments, and the debate has moved into the political domain. In these circumstances, it becomes a debate in which the awkward questions must be asked and answered clearly, or the solution that emerges may be one that will have to be revisited in more critical circumstances. It also requires us to take such parallels as exist and to examine them to assess what might be learned that is relevant.

Latin America provides an interesting paradigm through which to examine Open Access publishing, because the majority of journals published within the region are published by universities or with financial support from national research councils, other public institutions or professional associations. Commercial journal publishing has been inhibited not only by the relatively weak economies in the region, by the poor infrastructure of the book trade[3],[4], and by the lack of formal training in publishing. Nonetheless, developments in electronic publishing are taking place in Latin America, and it provides some notable examples of Open Access provision. It may thus offer some realities as a contribution to a consideration of issues in the more industrialised, wealthier countries that could otherwise easily be dismissed as false hypothesising. The paper begins with a review of the growth of scholarly and electronic publishing in Latin America, and then focuses on the key issues in the debate about Open Access: quality, visibility, findability, and sustainability.

## 2 Methodology

This paper draws on the results of recent research into journal publishing in Latin America, undertaken with support from the European Commission's ALFA programme ('**A***merica* **L***atina* - **F***ormacion* **A***cademica*'). The aim of the ALFA Programme is to support collaboration between European and Latin American Universities. In common with most of the research and development programmes that the Commission sponsors, it has to be based on a multinational partnership. In the case of ALFA, the requirement is that there should be at least 3 Universities from Latin America and 3 from the member states of the European Union. In this project, the Robert Gordon University's partners were Queen Margaret University College, Edinburgh; Universidad Nacional Autonoma de Mexico (CUIB); Universidad Nacional del Sur, Bahía Blanca, Argentina; Universidad Federal do Parana, Curitiba, Brazil; Hogskolan i Boras, Boras, Sweden; and Universidad Carlos III, Madrid, Spain. The aim of the project was to identify professional journals published in the region with a view to ensuring their wider availability through digitisation and thus contribute to professional education and development. Many of the projects supported by the European Commission's Research and Development programmes have short names that are intended to capture the underlying idea. This project was no exception. REVISTAS – 'journals' in Spanish - became an acronym for **RE***d* **VI***rtual* **S***obre* **T***odas las* **A***merica***S,** which translated into English as something meaningful**:** **'**a virtual network across the Americas.'

REVISTAS, focused on the feasibility of digitising journals as an aid to professional development in the field of Librarianship and Information Sciences, but journal publishing in this field is probably representative of many, if not most, disciplines in the region. The paper aims to discuss how traditional patterns of scholarly communication in that region are being or may be impacted by the shift to electronic media and the emergence of alternative approaches to publishing in a way that draws parallels between the Latin American experience and that in other countries where electronic publishing has become more widespread.

As well as reviewing much of the literature on the topic, the project team compiled a list of serial titles based on a number of indexes, journal articles and selected library catalogues. A comprehensive search would need to cover both the print and electronic catalogues of every institution that has taught librarianship and information studies, as well as every National Library in the region, and it must be acknowledged that more titles probably remain to be discovered by individuals more familiar with LIS publishing in their own countries. This is almost implicit in the wide disparity between the numbers of journals reported for each country. A final web search was carried out in early March 2007 using the metasearch engine 'Dogpile' to check for online versions of the list of titles that had been gathered to date[5].

## 3 The Growth in Scholarly Publishing in Latin America

There is no reliable evidence for the number of scholarly serials published in Latin America, but there is clear evidence of growth in the number of publications appearing in the languages spoken in the IberoAmerican communities. Whilst data from the ISSN International Centre[6] shows growth in the number of records for English language serials was c.19% between 2001 and 2006, it also demonstrates much faster growth in records for serials published in Portuguese, Catalan, and Spanish.

| Language | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | Increase |
|---|---|---|---|---|---|---|---|
| Portuguese | 13,244 | 13,277 | 13,294 | 13,310 | 21,324 | 21,361 | 61% |
| Spanish | 37,064 | 39,782 | 41,859 | 43,850 | 48,222 | 51,112 | 38% |
| Catalan | 1,034 | 1,163 | 1,248 | 1,340 | 1,479 | 1,555 | 50% |

**Table 1: Number of ISSN records per language (Source: ISSN International Centre).**

Only 11 of the countries in which these languages are used have national ISSN Centres. The data from them suggests that there are at least 24,816 records for serials published in Latin America.

| National Centre | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | Increase |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| Argentina | 7108 | 7391 | 7722 | 7954 | 10,040 | 11,006 | 55% |
| Brazil | 10001 | 10001 | 10001 | 10000 | 18572 | 18573 | 86% |
| Chile | 1510 | 1559 | 1758 | 1813 | 2065 | 2244 | 47% |
| Colombia | 1754 | 1754 | 1742 | 1743 | 1743 | 1798 | 3% |
| Costa Rica | 146 | 146 | 146 | 146 | 146 | 146 | 0 |
| Ecuador | 159 | 159 | 159 | 159 | 159 | 159 | 0 |
| Mexico | 3432 | 3432 | 3431 | 3431 | 3431 | 3431 | - |
| Portugal | 3924 | 3924 | 3924 | 3924 | 3924 | 3924 | 0 |
| Spain | 18876 | 21309 | 22576 | 24382 | 26303 | 27851 | 48% |
| Uruguay | 1526 | 1771 | 2019 | 2091 | 2225 | 2315 | 52% |
| Venezuela | 1704 | 1704 | 1704 | 1704 | 1704 | 1704 | 0 |

**Table 2: Number of records in national ISSN Centres (Source: ISSN International Centre)**

The apparent absence of growth in the number of recorded serials in some countries, rapid increases in other countries, and variations in the number of ISSN records for countries with similar populations suggests that there is probably significant under-recording in the ISSN system. Moreover, Latin American journals have not always registered an ISSN[7],[8], and it may be speculated that part of the growth in records may be attributed to belated registration. For example, the growth in records for serials emanating from Brazil (8572) is more than those in Portuguese (8117). There may also be some discrepancy between the data held at the national and international ISSN centres because of the difference between these linguistic and geographic analyses. Moreover, some of the serials published in Latin America are published in English. For example, 14% of the 239 Open Access journals indexed by ISI are English language journals that originate in Latin America.[9] Moreover, the number of titles apparently published in Portuguese and not recorded by the national centres in Portugal and Brazil is 1,136, which may or may not represent the output of the other countries from where publications in Portuguese may emanate (notably Angola, Goa, Mozambique, and Macau).

Not all the serials recorded by the ISSN Centre could be considered to be scholarly journals. More relevant data may be drawn from Latindex, the main directory of journals that is compiled within the region and principally intended as an aid for university libraries, which lists 15,578 titles, including 2,468 online journals.[10] However, there may also be some under-recording in Latindex. The main index to library science articles about Latin America that is compiled within the region[11] has recently been demonstrated to have indexed fewer than half the serials in the field that are now known to have been published.[12]

## 4 The Transition to Electronic Publishing

According to the ISSN International Centre, some 50,000 serials are available internationally in computerised formats, compared with 1.2 million in print. Although some under-recording may again be suspected, this reflects significant change since the first experiments with electronic publishing commenced in 1992.

As a contribution to resolving the problems of scientific communication in the region, the participants in a Conference convened by the International Council of Scientific Unions in Guadalajara in 1997 argued that the mechanisms for the promotion and distribution of scientific publications must be improved and suggested "...the establishment of a Latin American scientific electronic periodicals collection." [13]

Their thinking may have been influenced in part by the proximity to the establishment of *SciELO* (*Scientific Electronic Library Online*)[14] in Brazil in 1997. Its Open Access service has since spread to several other countries in the region. In addition, many other journals have established an online presence. For example, of the 220 journals in the field of Library and Information Sciences that are known to have been or are currently being published in the region, the full texts of 48 have now been made available online, but only 2 have met the criteria for inclusion in SciELO. Others are moving in the direction of online publication: 8 more journals publish an Electronic Table of Contents and Abstracts online, and 13 publish their Table of Contents.[15]

To provide access to these open access journals, a number of aggregator services have been established. The Brazilian Nuclear Information Centre maintains *LivRe*, a portal to more than 2,500 journals, not all of them in Spanish or Portuguese.[16] A more selective service is provided by *RedALyC, Red de Revistas Cientificas de*

*America Latina, el Caribe, Espana y Portugal*, which is hosted by the Universidad Autónoma del Estado de México, and provides access to some 300 peer-reviewed journals in Spanish, Portuguese and English.[17]

Several commercial database publishers also make a selection of journals in Spanish and Portuguese available. Grupo Océano, a Spanish company, has developed 6 databases covering different fields of knowledge.[18] EBSCOHost has developed 3 databases[19], whilst Thomson Learning promotes *Informe*.[20] Dialnet also includes some Spanish language content.[21] The most recent entrant to the field is ProQuest, which has developed a new collection of full-text scholarly journals *Publicaciones y Revistas Sociales y Humanísticas (Prisma).[22]* There appears to be some overlap between these services, and some even include titles that are freely available through SciELO.

## 5 Visibility

The growth in the number of serials may be explained by the growth of the local economies and consequently in national Higher Education and research systems. Latin American scholars are no different from those anywhere else in the world in the desire for their work to achieve recognition and make some impact in their field. In common with Higher Education institutions all over the world, Latin American Universities are experiencing the need to manage their educational, research and associated assets more effectively and transparently than in the past. They recognise that making their research and scholarly outputs more readily available will contribute to growth in the recognition of both scholars and institutions, and support the development of new and more fertile relationships between academic staff and departments both nationally and internationally, as well as stimulating economic and social development. Making them available could also facilitate much needed changes in teaching and learning, facilitating the development of a pedagogical environment that is information-rich and fosters the student-centred approaches to learning which are the key to success in the Twenty-First Century ‘Information Society.’

There is generally an expectation that - unless research is related to state security or defence, or is commercially confidential - the results will be published, i.e. that a report will be written for the sponsor, and that it will be summarised in whole or in parts in papers in scholarly journals, and perhaps in magazines intended for practitioners or the lay reader. Part of the assumption that these papers will be published is made possible by an understanding that “scientists in the public sector are largely motivated by intellectual curiosity, peer recognition and the promotion of the public interest rather than by private economic gain.” [23]

The expectation that the results of research will be published as a journal paper is reinforced by the reward system in the academic world – a reward system that is supported by governments. Research Councils in some Latin American countries have given career incentives and financial rewards to academics who publish in journals of high recognition and visibility. Paradoxically, it is often countries that provide support for the publication of indigenous journals that also focus their reward system for researchers on the publication of their work in international journals.[24],[25] Researchers in Latin America naturally want their papers to be published in international journals to improve access to their work and increase its global impact. These tendencies are enhanced in countries in which national research assessment and funding practices favour submission to international journals over submission to national journals. Elite Latin American researchers in all disciplines have therefore sought to maximise the potential impact of their research by submitting their manuscripts to well-established European and North American journals. For example, a study of the productivity of Mexican PhD holders trained abroad found that the majority had selected international journals indexed by ISI as their publishing outlets.[26]

Since the evaluation of research work can be influenced to some extent by the visibility and reputation of the journal in which the work is published, the choice of highly visible, prestigious journals as publication outlets has become crucial, especially for scientists.[27] If journals published in Latin America are to raise their attraction power for researchers in the region to select them as outlets for their research papers, they will clearly have to demonstrate that they are of comparable quality. This implies that quality control procedures will be in place, that other researchers will easily be able to find and access those journals, and that they will become sufficiently well established to become well known.

# 6 Quality control

The publisher of any journal is responsible for decisions about the level of quality control that is exercised by determining whether papers should be submitted to independent peer review. This will normally involve selecting (and often remunerating) an editor whose standing at least matches the perceived or expected standing of the journal, as well perhaps as some degree of oversight over the selection and activities of the members of any editorial advisory board. The editor makes a significant contribution to the standing of a journal by selecting experts from the editorial advisory board and/or others who can confirm that papers meet an acceptable standard in terms of their academic content. Paradoxically, although themselves largely drawn from the academic community, Latin American scholarly publishers and editors have not consistently addressed the crucial issues of quality control that affect the impact of the contributing authors' research. In the absence of any imperative to improve sales and distribution, peer review mechanisms in Latin American journals have been lax.[28],[29] One consequence has been that the journals often duplicate coverage of subjects, or reprint papers from elsewhere, whilst possibly leaving significant gaps in the coverage of sub-disciplines.[30]

In general, most scholarly publishing in Latin America has been handled by academics.[31] One commentator observed that scholarly publishing in the region seemed to be operated by highly committed and altruistic academics trained to do research, but not necessarily to run publishing houses.[32] These academics develop publishing skills on the job or in some countries through targeted professional development schemes. Latin American scholarly journals, supported by or through public institutions, depend on the annual budgetary allocation to enable them to sustain regular publication. They have been affected by regular financial crises in the region[33], and have not always succeeded in maintaining a regular publication schedule. These institutional journals are frequently not sold through subscription mechanisms but exchanged for journals from other universities or associations. They rarely reach a wide international audience. Library collections often contain incomplete files of a journal.

To overcome the problems that are endemic in journal production in the region, inclusion in SciELO requires adherence to rigorous guidelines requiring peer evaluation and regular production "thus establishing challenges for the enhancement of the scientific output in the participating countries".[34] However, the consequence of this policy of selectivity is that SciELO currently makes only c.248 journals available online in full text, a small proportion of the total published in Latin America.

Quality control also includes the technical preparation of the journal. Journal publishers incur significant costs in getting the product to the reader. Editorial offices have to be maintained for logging new papers, tracking their progress, and generally communicating with the authors, referees, printers, etc. A small but increasing number of publishers in Latin America are now using Open Journal Systems[35], for managing journal production, particularly in Brazil where it has been translated into Portuguese (SEER - Sistema Eletrônico de Editoração de Revistas[36]). Despite the fact that most papers are now submitted in electronic format, there is almost always a certain amount of effort necessary to check citations for accuracy, to create links to CrossRef, and to complete the proof reading and copy-editing. Some publishers and editors of open access journals appear to be attempting to transfer some or all of this responsibility to their authors. Whether that will be acceptable to authors and practicable remains to be seen.

The adoption of online format for some journals had not overcome the problem of irregular publication and consistent access. Some journals had not been published for several years; others had been only short-life experimental projects. Some other, single issues of papers or journals appear to have been converted into Permanent Document Format (pdf) solely at the initiative of their author or editor and made available through a repository or an aggregator such as RedALyC. In some cases, the URL had changed without a link to the new URL being created, or the links to the text of articles were broken. There is already evidence that some online journals are not attracting sufficient papers to maintain a regular publication cycle.

# 7 Findability

Making the journals or papers available online is of little value unless there are good indexing and abstracting services to guide the potential users to papers that are relevant to their interests. The international visibility of scholarly periodical publishing in Latin America has been the object of a number of studies.[37],[38],[39] Whilst the regions major news magazines and newspapers are indexed in several subscription-based online sources[40],[41],[42], only a small proportion of scholarly periodicals from developing countries is indexed and abstracted by the major scientific secondary databases.[43],[44] A central archive of indexing data and a cross-site searching facility for SciELO is based in the original office at BIREME – *Biblioteca Regional de Medicina* in São Paolo. Recently, the SciELO index to its Brazilian journals has recently been uploaded into OCLC's WorldCat.[45] SciELO Chile will be uploaded shortly, and other SciELO partners are expected to follow. This provides an alternative access point for potential users of the journals included in SciELO, and will arguably raise their visibility and use, at least amongst OCLC's member libraries. OCLC has also recently added to its database the indexes (*Clase* and *Periódica*) that have been compiled by the Dirección General de Bibliotecas at the Universidad Nacional Autónoma de México (UNAM-DGB) for the last 28 years, covering 400 of the region's journals in the arts, humanities, social and pure sciences. [46] A similar number of journals, possibly the same collection, are now included in SCOPUS. However, the full texts of few of these journals are available online.

Research papers made available internationally through electronic publishing appear to have a higher national and global impact than achieved through publication in an indigenous printed journal. However, it is also important that the indexing service is widely known, and this is by no means the case. An interesting example is provided by the most substantial index to journal articles on Librarianship and Information Sciences from or about Latin America that is compiled within the region, itself the sole survivor of attempts made to establish such a service in several countries. INFOBILA was initiated in 1986 by the Universidad Nacional Autónoma de México Centro Universitario de Investigaciones Bibliotecológicas.[47]It is based on collaboration with a network that covers 13 countries in addition to Mexico: Argentine, Brazil, Chile, Columbia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Panama, Peru, Uruguay and Venezuela. It has been freely available online since 1997 and has recently been redeveloped with the capacity to include or link to the full-texts of journals. However, it indexes only a handful of the online journals produced in the region. It is also disappointing that it seems possible that INFOBILA may be almost completely unheard of in the countries in which the indexed journals originate. An impromptu survey of the c.350 participants of a conference on digital libraries in Argentina in October 2005 suggested that only about 10 people there were familiar with INFOBILA.
There is some evidence to suggest that, as a result of their availability through SciELO, a number of English language journals originating in Latin America and indexed by ISI are attracting more attention and more citations by other researchers than previously. A study of the 5 journals published in English in Brazil that have been indexed by ISI for at least 5 years, and available in full-text on SciELO[48] for at least 2 years revealed that their impact factor had more than doubled since their inclusion in SciELO.[49] Interestingly, Thomson ISI has recently agreed to begin including journals in Spanish in its Citation indexes from January 2006, possibly under pressure from its considerable Spanish customer base (as well, perhaps, as incipient competition from new indexing services such as Google Scholar[50] and SCOPUS[51]). The impact of this on author preferences for publishing outlets for their research remains to be seen.

Having good and well-known indexes goes only part of the way towards making the full text available. The difficulty in tracing the printed journals has been exacerbated by a relatively large production of new titles with small readerships and short life cycles.[52] Commercial publishers seeking to digitise some of the region's journals have already experienced difficulty in finding complete sets to digitise, and the searches conducted for the REVISTAS project confirmed the haphazard distribution of copies of many of the printed journals. However, few of the online journals appear to have taken the steps necessary to publicise their existence, or to ensure that their contents are discovered by registering with a variety of aggregators and search engines. In many cases there was no evidence of registration of ISSNs. Coverage by the IberoAmerican e-journal aggregators was poor. *Livre* included only 29 of the 90 librarianship and information sciences journals published in Spanish and Portuguese (including those published in Europe), whilst *RedALyC* included only 8. The principal European aggregator of journals in Spanish and Portuguese (and other open access journals) is *REI, Recursos Electronicos de Informacion*, a service maintained by the Universidad de la Rioja in Spain.[53] REI is maintained on behalf of REBIUN, the Spanish University libraries consortium and is not limited to peer reviewed journals, but still included only 15 of the 90 titles. Moreover, the aggregators and indexes are not necessarily well known. *RedALyC* was not known to the REVISTAS partners from the region, nor to senior LIS professionals based in the same city as its host institution. Bypassing the aggregators may overcome their deficiencies. SciELO is now beginning to use CrossRef[54] to create links to and from the full text of papers in the journals that it hosts, but there are few signs that this practice has yet been more widely adopted.

# 8 Sustainability

Much of the fabric of online publishing in Latin America remains supported by institutions. Most of the journals are published by universities. SciELO is supported by FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo in collaboration with BIREME - Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde, and CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico. Anecdotal comments suggest that its principal funders may be questioning why a local agency should be funding what has become an international service. RedALyC is alleged to depend to some extent on international aid from the Spanish government.

The technology lends itself to creating electronic collections that offer a range of features that add value, and that will be increasingly expected as standard features of e-journal services. Features such as cross-file searching, browsing, saved search histories, Table of Contents alerts by email, and citation linking must be introduced into the region’s electronic journals, presenting challenges in terms of the availability of both skills and finance. That must raise concerns about the future sustainability of these journals. There is already evidence that some online journals had not been published for several years; others had been only short-life experimental projects.

The insecure financial base must raise concerns about long-term preservation of those electronic journals that do appear, particularly as many of the region’s National Libraries do not have a preservation policy that extends to electronic media produced in their country – or the resources to implement one.

# 9 Discussion and Conclusions

In Latin America, the rewards and recognition for researchers and other academics are closely linked to the perceived quality of their published outputs. The evidence indicates quite clearly that the absence of any commercial imperative to raise quality to improve sales and distribution has had a damaging impact on quality controls. The publishers and editors of the vast majority of the electronic journals that have emerged to date seem largely to be continuing their previous neglect of quality control.

Visibility is clearly an issue for Latin American researchers. They want their publications to be highly visible. The evidence tends to indicate past failure in efforts made in Latin America to raise awareness of the contents of journals and to ensure adequate distribution of copies to meet potential demand. Making the full text of journals freely available online alone has not yet proved any more effective.

“Findability precedes usability”.[55] The evidence is that the region’s printed journals have not been well served by international or indigenous indexing services. Although some efforts are now being made to improve the arrangements for indexing, these only serve to highlight the limited availability of full-text sources.

The final issue to emerge from this study was concerns about the sustainability of publications and related services that depend on institutional support. The evidence tends to indicate that, to date, personal or institutional circumstances have contributed to the short life of many Latin American journals and newsletters. Simply switching to electronic media has not yet entirely resolved these issues.

The problems that have been discussed may be peculiar to Latin America. The examples drawn from this review of the region’s literature of Librarianship and Information Sciences may not be exactly paralleled in every discipline, and further research to test the findings from this study on a wider scale is needed, and needed soon. However, the realities of scholarly communication in Latin America should prompt a pause for reflection by anyone interested in securing the future of scholarly communication at a time when the existing system is undergoing significant changes.

The aim of the paper was to use these realities to provide a fresh perspective on some of the global implications of the shift to electronic publishing, particularly to inform the debate about open access publishing, and to point to issues that still appear to need further consideration before significant changes in the system of academic communication are put in place. The situation that now exists may, in some respects, be resembled to the position of the ‘Titanic’ approaching the iceberg. The scale of the problem that confronts us is enormous. Just like the bulk of the iceberg, the vast majority of research papers are out of sight, not hidden below the surface but because they have not yet been written. The arguments about the most appropriate course to steer are complex. Faced with what is perceived as a major threat to scholarly communication, there are members of the research community and the library community who seem to want to abandon ship immediately without any clear idea of whether that is the safe course of action. The experience of institutionally supported publishing in Latin America and the faltering emergence of electronic journals there suggests that open access publishing could prove to be as much use as a safeguard for scholarly communication as a trap door on a lifeboat.

## Acknowledgement

The author gratefully acknowledges the contribution made by the partners in REVISTAS and particularly Dr. Virginia Cano to the preparation of this paper.

## Notes and References

[1] REINSBOROUGH, P. De-colonizing the revolutionary imagination: values crisis, the politics of reality and why there's going to be a common sense revolution in this generation. *The Journal of Aesthetics and Protest*, 1 (2), August 2003. [online]: http://www.journalofaestheticsandprotest.org/1/de_colonizing/8.html [Accessed 8 April, 2005]

[2] ROGERS, E.M. *Diffusion of innovation*. 4th ed. 1995. New York, U.S.A.: Free Press.

[3] JOHNSON, P.T. A brief overview of the book trade in Spanish speaking Latin America *in Seminar on the Acquisition of Latin American Library Materials (19, 1974, Austin, Texas). Final report and working papers.* 1976. Amherst, Mass.: SALALM Secretariat pp. 55-59

[4] BABINI, D.; SMART, P. Using digital libraries to provide online access to social science journals in Latin America. *Learned Publishing*, 19 (2), 2006, 107-113

[5] 'Dogpile' [online]: http://www.dogpile.com/ [Accessed 1 March 2007]

[6] ISSN International Centre [online] http://www.issn.org/ [Accessed 15 April 2007]

[7] CANO, V. Bibliographic control and international visibility of Latin American periodical publications. *in*: *Indicators for developing countries, edited by* R. Arvanitis and J. Gaillard. 1992. Paris: ORSTOM. pp. 511-526.

[8] CANO, V. Characteristics of the publishing infrastructure of peripheral countries: A comparison of periodical publications from Latin America with periodicals from the US and the UK. *Scientometrics,* 34 (1), 1995, 121-138.

[9] MCVEIGH, M.E. *Open Access journals in ISI databases: analysis of impact* factors and citation patterns. 2004. [online]: http://www.thomsonscientific.com/media/presentrep/essayspdf/openaccesscitations2.pdf [Accessed 15 January 2006]

[10] Latindex -Sistema Regional de Información en Línea para Revistas Científicas de América Latina, el Caribe, España y Portugal. [online]: http://www.latindex.unam.mx [Accessed 15 April 2007]

[11] INFOBILA [online]: http://cuib.laborales.unam.mx [Accessed 15 April 2007]

[12] JOHNSON, I.M.; CANO, V. Electronic publishing in Librarianship and Information Sciences in Latin America – a step towards development? Forthcoming.

[13] CETTO, A.M.; ALONSO, O., *editors. Revistas cientificas en America Latina – Scientific Journals in Latin America.* 1999. Paris: International Council of Scientific Unions; Mexico: UNAM, CONACYT, and Fondo de Cultura Economica. pp. 461-466

[14] SciELO [online] - http://www.scielo.org/ [Accessed 7 August 2004]

**[15]** JOHNSON, I.M. REVISTAS – online LIS journals in Latin America. Focus on international library and information work, 2007. (forthcoming)

[16] LivRe [online]: http://livre.cnen.gov.br/ [Accessed 1 March 2007]

[17] RedALyC [online]: http://www.redalyc.com/mx [Accessed 3 January 2006]

[18] Grupo Océano [online]: http://www.oceano.com/oceano/oceano.html [Accessed 5 December 2005]

[19] EBSCOHost [online]: http://www.epnet.com/ [Accessed 15 December 2005]

[20] Informe [online]: http://www.gale.com/pdf/facts/inform.pdf [Accessed 5 December 2005]

[21] Dialnet [online]: http://www.dialnet.com.mx [Accessed 12 April 2005]

[22] Prisma [online]: http://www.il.proquest.com/division/pr/05/20050408.shtml [Accessed 5 December 2005]

[23] UHLIR, P.F. Re-intermediation of the Republic of Sciences: moving from intellectual property to intellectual commons. *Information Service and Use*, 23 (2/3), 2003, 63-66

[24] VESSURI, H. Estrategia de valoracion de las revistas cientificas Latinoamericanas [Strategy for evaluation of Latin American scientific journals]. in: *Publicaciones cientificas en America Latina; edited by* A. Cetto and K. Hillerud. 1995. Mexico: Fondo de Cultura Economica. pp. 200-210

[25] BONILLA , M., *and* PEREZ ARAGON, M. Revistas Mexicanas de Investigación Científica y Tecnológica. *Ínterciencia* 24(2), 1999, 102-106.

[26] LICEA de ARENAS, J.; SANTILLÁN-RIVERO, E.; ARENAS, M.; VALLES, J. Desempeño de becarios Mexicanos en la producción de conocimiento cientifico ¿de la bibliometria a la politica cientifica? *Information Research*, 8(2), 2003. paper no. 147 [online]: http://InformationR.net/ir/8-2/paper147.html [Accessed 15 January, 2006]

[27] RAVETZ, J.R. *Scientific Knowledge and its Social Problems*. 1971. Oxford: Clarendon Press

[28] CANO, 1992, ibid.

[29] MENEGHINI, R. Brazilian production in Biochemistry: the question of international vs domestic. *Scientometrics*, 23 (1), 1992, 21-30.

[30] DIAZ, I.G.; AGUILAR, G.S. Las revistas científicas: su problemática en América Latina y El Caribe. [The problems of scientific journals in Latin America and the Caribbean.] in: *Revistas científicas en América Latina - Scientific Journals in Latin America; edited by* A.M. Cetto and O. Alonso. 1999. Paris: International Council of Scientific Unions; Mexico: UNAM, CONACYT, and Fondo de Cultura Economica, p. 231

[31] CANO, V. International visibility of periodicals from Ireland, India and Latin America. *Knowledge and Policy,* 6 (3-4), 1992-1993, 55-78

[32] GOMEZ, Y.J. A proposito de un Ejercicio de Evaluacion de Seriadas Cientificas. [A proposal for an evaluation of scientific journals.] Paper presented at the *Second International Workshop on Scientific Publishing in Latin America,* Guadalajara, Mexico November 27-30, 1997. [Unpublished]

[33] BABINI; SMART, 2006, ibid.

[34] GREENRIDGE, E. An overview of the PAHO Virtual Health Library. in: *Models of Cooperation in U.S., Latin American and Caribbean Libraries: the first IFLA/SEFLIN international summit on library cooperation in the Americas; edited by* B.E. Massis. 2003. Munich, Germany: K. G. Saur. pp.45-51

[35] Open Journal System [online]: http://pkp.sfu.ca/ojs/ [Accessed 15 April 2007]

[36] SEER - Sistema Eletrônico de Editoração de Revistas [online]: http://www.ibict.br/secao.php?cat=SEER [Accessed 15 April 2007]

[37] CANO, 1992, ibid.

[38] KRASUKOPF, M.; VERA, M.I. Las revistas cientificas de America Latina acreditadas en el ISI [Latin American journals indexed by ISI]. in: *Publicaciones cientificas en America Latina, edited by* A. Cetto and K. Hillerud. 1995. Mexico: Fondo de Cultura Economica. pp. 168-175.

[39] CETTO, A.M.; HILLERUD, K., *editors*. *Scientific publications in Latin America.* 1995. Mexico: ICSU, UNAM, and Fondo de Cultura Economica. p. 305

[40] Info-Latinoamerica [online] - http://www.nisc.com/factsheets/qila.asp [Accessed November 2003]

[41] Latin American Newsletters [online] - http://www.latinnews.com [Accessed November 2003]

[42] Prensa Latina [online] - http://www.prensa-latina.cu/English/ [Accessed November 2003]

[43] CANO, 1992-3, ibid

[44] GONCALVES DA SILVA, L.; SILVA FERNANDES, R. La cobertura de las revistas Latinoamericanas por los Servicios de Indizacion: el caso de las revistas brasilenas. Paper presented at the *Second International Workshop on Scientific Publishing in Latin America*, Guadalajara, Mexico, November 27-30, 1997. [Unpublished]

[45] OCLC WorldCat [online]: http://www.oclc.org/worldcat/default.htm [Accessed 3 January 2006]

[46] *Clase* and *Periódica* [online]: http://www.dgbiblio.unam.mx/ [Accessed 7 May 2006]

[47] INFOBILA is available, free of charge, directly though the UNAM-CUIB web site [online]: - http://cuib.laborales.unam.mx [Accessed 7 August, 2004]

[49] ALONSO, W.J., *and* FERNANDEZ-JURICIC, E. Regional network raises profile of local journals. *Nature*, 415, 2002, 471-472

[50] Google Scholar [online] – http://scholar.google.com/ [Accessed 7 May 2006]

[51] SCOPUS [online] – http://www.scopus.com/scopus/home.url [Accessed 7 May 2006]

[52] GUIMARAES, J.P. Opportunities and common goals for research in the Americas. in: *Science and technology in the Americas, perspectives on Pan American collaboration, edited by* J. Stann. 1993. Washington, D.C., U.S.A.: American Association for the Advancement of Science. pp. 65-72.

[53] REI, *Recursos Electrónicos de Información* [online]: http://aps.unirioja.es/biblio/recursos?sub=1 [Accessed 3 June 2006]

[54] CrossRef [online]: http://www.crossref.org/ [Accessed 3 January 2006]

[55] Findabilty.org [online]: http://www.findability.org/archives/cat_findability.php [Accessed 12 January 2007]

# Expectation and Reality in Digital Publishing: Some Australian Perspectives

*Bill Martin, Hepu Deng; Xuemei Tian*

School of Business Information Technology, RMIT University, Melbourne, Victoria, 3000, Australia
e-mail: {bill.martin; hepu.deng; xuemei.tian}@rmit.edu.au

## Abstract

This paper presents a brief summary of the findings of a Web-based survey of the views of Australian publishers, on the potential impact of digital technologies, followed by three case studies conducted between January and April 2007. The survey results indicate that the most influential technologies currently in use in publishing are the Internet and the World Wide Web, with little or any interest being shown in for example, semantic technologies. There is however, widespread realization of the importance of providing enhanced customer value through digital content and delivery channels, with consequent implications for changes to value chains and the emergence of new and transitional business models, which however, are likely to complement rather than replace existing business models. The case studies drawn from a set of eight selected to include a range of value propositions and business models suggest that in Australia publishers are optimistic about the prospects of digitisation but are nonetheless cautious in its uptake and application.

**Keywords:** digital publishing; business model; value chain; case study; Australia

## 1 Introduction

Traditionally the publishing industry has played a key role in the dissemination of knowledge and for centuries was a forerunner of what today would be described as a *knowledge-based business* [1]. Until the closing decades of the last century, publishing and associated printing activities were based upon old technologies, with clear implications for business processes and relationships among the main stakeholders in what was basically a linear progression from the creator of content to its publication in print form [1, 6, 7].

The advent of digital technology has potentially limitless implications for publishing both in hard copy and electronic formats [2, 6, 7]. Combined with advances in electronic commerce [8] it offers the prospect of new value propositions and business models for those who are able to take advantage of developments in digital technology. Digital publishing incorporates several characteristics including an infrastructure that gives multiple options with digital content available in various formats and viewing modes according to customer requirements and basic editing processing and updating of information on the server, leading to reductions in processing time and the fast, efficient transmission of content, with subsequent economic benefits [5]. This said, even the latest digital tools and applications are at best enabling mechanisms whose adoption must relate to the overall business strategy and purpose [4, 5].

This paper presents the initial findings from an Australian government-funded research project looking at the implications of digital technologies for the publishing industry in Australia, with particular emphasis upon current and emerging stakeholders, competition, and value propositions and business models, current and potential. The findings (which are still to some extent interim in nature) have emerged from a variety of research activities including literature review, focus groups, a national online survey of publishers and the conduct of case studies. The project adheres to the generally accepted view of publishing as a set of content industries comprising sectors for book, journal, newspaper, directory, magazine, music, maps and multi-media publishing [1, 2]. However, its major focus is on book publishing in Australia. This paper concentrates largely on three case studies conducted during the research.

## 2 Methodology

Following the conduct of an extensive literature review (including analysis of secondary documentation such as Annual Reports) and of three focus groups, the major research methods employed were those of survey and case study. After several unsuccessful attempts to obtain access to relevant membership listings, the researchers made use of a commercial listing service. They provided a list of 65 publishing companies throughout Australia. In the event this turned out to be an exercise of somewhat limited value in that the great majority of addresses obtained were those of newspaper and magazine publishers (particularly publishers of rural newspapers), most of whom had no interest in participating in the project. However all those responding were in fact book publishers and their responses, limited in number as they were, tended to support the major assumptions underlying the survey. The case study was operated on the basis of a set of protocols designed in order both to facilitate consistency in the handling of responses to key issues raised in the survey and to ensure the presence of a certain amount of rigour in the conduct of the case study exercises.

## 3 Analysis of Survey Results

On a more positive note, the conduct of a survey had always been regarded as being part of a triangulation process involving the literature review and focus groups and the conduct of case studies. The data analysis resulted in identification of the general extent of progress made towards planning and implementing digital initiatives, and more specifically, those factors that influenced this process and issues with regard to industry trends, stakeholders and competition, propositions and business models. Of the 65 surveys mailed, and subsequently re-mailed to publishers, only 14 were returned completed. Although a response rate of almost 22 percent from a Web-based industry survey would appear to compare well with the reported norm for such exercises of 4% to 6% [3], the researchers make no claims for significance. The findings will be presented in a forthcoming paper and are summarised here as follows:

- 70% of respondents reported increased growth in revenues from existing products/services and nearly 60% from new products/services.
- The main benefits anticipated from digital technologies are in the areas of new niche markets, repackaging and repurposing of existing content, consumer-generated content and the enhancement of value chains.
- The most profound effects expected from digital publishing are in the areas of specialist business/professional/academic publishing, government and web-based publishing.
- The critical success factors for digital business models were identified as technical robustness, consumer acceptance and financial logic.
- Subscription-based and content creation business models were the most highly regarded, frequently in the context of niche markets.
- Key organisational changes anticipated included:
    - ❖ Introduction of digital media divisions.
    - ❖ Introduction of an integrated platform for all editorial operations, print and digital.
    - ❖ Changes in human resource practices to suit a digital environment.
    - ❖ Organisation-wide promotion of cultural change to suit a digital environment.
    - ❖ Introduction of new strategies for the digital market.

## 4 Background to the Case Study Element

About half of the candidates for case research emerged from the online survey exercise and the rest were obtained later by direct approach. The three cases reported here are drawn from a group of eight that will be completed as part of the research project. These are all exploratory and descriptive in nature, rendering them suitable for the kind of interpretive research undertaken in the project. The case studies gave researchers the opportunity to meet face-to-face with senior members of publishing companies and discuss the results of the survey analysis with them. The case study instrument was designed to enable respondents to take ownership of the process and respond within the boundaries of meta-level questions [8].

Interviewees were asked a combination of open and closed questions and were free to add anything else they thought important. The case studies operated on the basis of a standard set of protocols relating to research design, operating procedures and data analysis techniques [3]. This was to guard against bias and ambiguity and to ensure as far as possible that a logical chain of evidence could be seen to operate from the initial research questions to the ultimate conclusions [9-11]. This, for example, led to the use of *How* and *Why* type questions for exploring operational links over time and *What* type questions for exploring new phenomena such as digital developments. The protocol specified detailed procedures in relation to data collection during the interview process. Every interview was recorded and transcribed, with the transcriptions being read and independently analysed by two of the three-person research team, one of whom had not participated in the particular interview session under analysis. In addition, the teams of two interviewers both kept separate field notes, which again were later subject to mutual and then third party scrutiny. Finally with regard to data analysis, the strategy was designed to link findings and interpretations not to generalisable outcomes but to contexts beyond the immediate, to extrapolation to other situations and environments [11].

## 5 The Case Studies

In the three cases reported here, the firms are identified only by the use of numerals. They comprise respectively a university press (Company A), an educational publisher (Company B) and an electronic publisher with close connections to a conference operator (Company C). The major focus will be on their business models, which for present purposes are perceived as a description of the roles and relationships among a firm's consumers, customers, allies and suppliers that identifies the major flows of product, information and money and the major benefits to participants [12]

The business models of the three firms were identified following cross comparison of each company across a range of constructs identified as important to successful business models. These are:

- Customer base
- Value proposition
- Value chains
- Core competencies
- Products and services
- Partners
- Use of and Attitudes to Technology
- Risks and opportunities
- Business mdoels

### 5.1 Customer Base

There is a considerable similarity in the makeup of the customer bases of these three firms, serving as they do a largely academic or educational market. However, one area of difference is that Company A as well as Company B is engaged in the provision of publishing services to conference organizers. Specifically, the customer base of these three companies can be described as follows:

- For Company A, the customer base has remained remarkably stable for the last 16 years, with the main difference being with regard to expansion into overseas markets. Most of their customers are libraries (notably academic, state and corporate) and small publishers, with a minority of direct sales to end users over the Web.

- For Company B, the customer base is comprised of teachers and pupils in the primary and secondary sectors.

- For Company C, the customers are mainly academics either seeking to publish their own papers or access those of others, on either a subscription or a per item basis. There is also a small but growing segment of custom in the library market and substantial revenue from the provision of publishing services to the associated conference business.

## 5.2 Value Proposition

All three firms offer customers a range of value propositions including:

- Companies A and C offer the benefits of a full electronic publishing service including provision of software, metadata, file conversion, content management and quality.

- Company B, while offering a digital dimension in the form of PDF formats and Website *question and answer* facilities, has as its major value proposition the ability to delivery content in the form of hard copy textbooks.

- Companies A and C offer the benefits of online access to and delivery of aggregated and indexed content based on a common technology platform and sophisticated search technologies.

- Company A offers provision of an additional marketing, sales and promotion channel to its customers.

- Company A provides archival services.

## 5.3 Value Chains

The value chains of the three firms are all familiar in scope although that for Company B is much the more traditional: author to publisher to printer to distributor/bookseller to reader [1]. While in essence the same, the value chains for companies A and C are much more geared to a digital environment with the major stages entailing:

- Stage 1: Acquiring content from authors or owners (via licensing or payment).

- Stage 2: Obtaining and converting digital files involving PDF and XML formats, creation of metadata and databases, editing and quality assurance.

- Stage 3: Printing (frequently outsourced) with content held in digital repositories.

- Stage 4: Sales, marketing, promotion through representatives, print media and virtual and physical book shops.

- Stage 5: Archiving content

All three companies were confident of maintaining their place in what they expect to be changing value chains in the near future. They were not concerned about possible disintermediation as a result of technology, but all agreed that booksellers have reason to be concerned.

## 5.4 Core Competencies

All three companies identified as core competencies the provision of high quality content (in either print or digital formats), the ability to organize content including editorial competencies and the ability to negotiate licensing and royalty arrangements, and the provision of networks of business partners and services including production, distribution, marketing and selling. Those competencies emerging as specific to individual companies included:

- Meta data creation, file conversion and content management (Companies A and C).

- Competencies in current and emerging classroom content delivery methods (Company B).

- Competencies in curriculum development and assessment (Company B).

- Technology competencies (Companies A and C).
- Conference management competencies (Company C).

## 5.5 Products and Services

With respect to the products and services offered, Company B is clearly different from Companies A and C. This is because the main source of revenue for Company B is through the sale of hard copy textbooks, with a modest trade in e-Books in PDF format and the delivery of classroom content via digital whiteboards.

In contrast Company A earns 96% of its revenue from digital products and services including:

- Bibliographic databases which also form the basis of the search infrastructure.
- Online databases giving access to fully indexed full text journal articles by using a single search interface.
- E-Press: a cover-to-cover aggregation of journals, monographs, conference papers, reports, occasional series and other *grey* literature published in Australia and hitherto not widely available online.

Company A has a minor trade in hard copy books (some 4% of output) and offers a full e-publishing service to a growing client list.

Company C also sells consultancy services (both publishing and technological) , as well as hosting conferences, the revenues from which subsidizes publishing activities including:

- Access to digital content in the form of monographs, single papers and electronic journals.
- Access to journal contents via an archive of titles and abstracts.

## 5.6 Partners

All three firms have common partnership arrangements in the form of links to authors, printers, marketers, distributors and booksellers. Company B has a particular relationship with schools and Company C with its associated conference business. Of the three, Company A has the most diverse range of partners which in addition to those mentioned in the foregoing include the National Library of Australia, the Copyright Agency, a range of government departments, such as the Attorney General's Department, various research centres in fields as diverse as family studies, criminology, agriculture and languages and a range of small publishing operations seeking to go digital. All these partners in a variety of ways add value to the products and services of the case companies.

## 5.7 Use of and Attitudes to Technology

There were clear differences in the attitudes of the three case firms towards the take-up and development of technology. Both Company A and Company C had from the outset relied upon the use of technology to gain market share and a competitive edge. They had sought to market a technology-intensive value proposition. Company B was much more pragmatic, linking developments in technology infrastructure and applications much more closely to market demand. Although a multinational company with ample financial and other resources, Company B did not maintain an active research and development program as such, preferring to monitor general developments and if necessary respond appropriately.

Companies A and C were easily the most enthusiastic of the six case companies interviewed prior to the writing of this paper. They both strongly endorsed the potential for *many-to-many* forms of communication including contributions from end users and distributed content and cognition. They were particularly interested in the potential of the Semantic Web and Web 2, not least given their respective histories of involvement with and expertise in metadata creation, file conversion and content management. They are heavy users of XML for the management of often relatively small print runs and the transition from source to print and web outputs using open standards and a high degree of automation. To this extent they see themselves as already beginning to engage with the notion of the semantic web, but realize that there is a long way to go before this comes to fruition.

This commitment to technological development at both Companies A and C is simply a reflection of their continued appreciation of the value of technology to the sustainability of their businesses. Hence while both Companies A and B outsource aspects of meta data creation and file conversion (in the case of Company C to Mumbai) this has been done more for technical and quality reasons than simply to cut costs. For Company C the Mumbai operation is critical to its global data harvesting activities, which in turn are central to the marketing of conferences and the recruitment of authors and journal editors.

Both Companies A and C have invested heavily in proprietary content management and workflow systems. Key files and databases at Company A are based on Terratext Foundation software developed within the company's parent institution and for which Company A has a permanent licence. Company C, following extensive work with almost 20 standards, has developed a core publishing and workflow management system (CG Publisher) which it claims is the first fully online publishing environment in the world. It manages publishing proposals, version control for drafts and editions, and contracts and automatically places completed texts (print and electronic) into an easily managed self-publishing site, as well as into personal sites for each of the authors. Company C believes that it is largely owing to the existence of its core management systems that it has been able to grow its business ten-fold during the last three to five years

Although nowhere as engaged in the development and application of technology as the other two firms, Company B is by no means oblivious to its importance. In addition to a small-scale involvement in the production of e-Books, Company C delivers content under licence to classrooms using its own range of electronic whiteboards. To date the uptake of this technology has been constrained both by a shortage of relevant content and by school budgets.

## 5.8 Risks and Opportunities

Company A sees very little on the horizon as regards potential risks, and in particular nothing in the way of threats from new entrants or from developments in technology. In terms of good governance they are focusing on keeping costs down, for instance in relation to royalty and licensing fees and looking at opportunities for improving their delivery infrastructure in order to reduce the unit costs of production. There is little sign of any potential problems from for instance channel or supplier conflict. The company is very comfortable with ongoing developments in Open Access publishing, which it regards as being highly domain and content-specific and where the future may lie in the publication of material that is not saleable on a commercial basis. Company A is currently participating in a local repository experiment, for which they are providing input on software and content management. However they see this more as a goodwill gesture than as a commercial venture. So far as technology is concerned, they have been early adopters of digital opportunities and they would see further opportunities in the digital publishing space owing to their strengths in metadata creation and management and in indexing and searching. They are also intending to pursue new markets comprised of library consortia and large libraries in Asia, the United Kingdom and North America, to repackage and reformat existing materials for corporate and enterprise markets, and to develop new products both with regard to aggregated services and content.

Despite its relatively low key presence in digital markets, the future whether in terms of technological or related change holds few fears for Company B. Hence, although much has been made of potential disintermediation in the value chain for publishing consequent upon the empowerment of authors or on competition from new players in the market, Company B is confident that whereas booksellers may be adversely affected, changes in publisher-author relationships are just as likely to be in its favour. There could be a risk of channel conflict were they to move to any substantial form of direct-to-customer sales, or indeed to any wholesale attempt to deliver content through their Website (hence conflicting with the traditional book selling model). On the other hand, threats from the wholesale digitization of texts, say by Google or Microsoft, are seen as more a matter for old material than for new. Their customers on the other hand, want new and dynamic content. Curricula are constantly changing and publishers have possibly unique expertise, not only in updating content but also in scoping and sequencing it in relation to course changes and more generally in the organization of content. Their view is that if the Internet has taught us anything it is that *more is not necessarily* more when it comes to, timely, relevant and high quality content of credible provenance. The major threats posed to educational publishers in Australia for the foreseeable future are not those of digitization, but rather of government policies, not just as regards the funding of education but also in relation to support for the creation of content. In Australia, governments at both state and Federal level, appear to see the future of digital content production as lying outside the mainstream publishing industry. For example, a national effort to produce such content through a body called *the Learning Federation*, has largely succeeded only in producing sets of learning objects (animated content intended to illustrate the use of say mathematical or scientific concepts for use by teachers) which according to surveys conducted by the Copyright Agency are used by very few schools. Company B is of the opinion that they could profit from the opportunity to develop and supply customized digital content. Indeed, a diversification in content creation, along with realistic funding for school hardware and software, could result in a transformation of the firm's value proposition to the point where additional revenue streams would accrue not just from digital content but also from the provision of hardware and software.

Ironically one potential area of risk for Company C stems from its inherent technological strengths. The fact that their technology is so sophisticated means that it is costly both to implement and to amend for different purposes. Moreover, there is a very high sunk cost in a relatively small pool of technical staff, with the accompanying risk of knowledge loss and damage to the business through the departure of key people. So far as competition is concerned, Company C is extremely comfortable, given that it owns the conference business which underpins the supply of content to its publishing arm. They perceive potential opportunities through the development of semantic technologies and given their existing expertise in connecting and processing digital documents they believe that they have every reason to be positive about the future.

## 5.9 Business Models

Authors such as Timmers [13] and Weill and Vitale [12] have categorized business models by type, arguing that for any organization, the business model can be constructed from any two or three atomic models drawn from this categorization. An analysis of the business processes of the three case study companies revealed that all of them contained at least two of the following atomic e-business models:

- Direct-To-Customer: This involves a small but growing B2C model operating as pay-per-view with customers paying either by monthly account or by credit card.
- Content provider: providing content (information, digital products and services) via intermediaries.
- Intermediary (Aggregator): bringing together buyers and sellers by concentrating information.
- Shared infrastructure bringing together a range of players (some of them competitors).

The models shown below will employ a schematic developed by Weill and Vitale [12] wherein the following components and relationships will be depicted as follows:

- Participants: Represented as:
  - ❖ Squares (firms of interest).
  - ❖ Left- and –right-facing pentagons (customers and suppliers).
  - ❖ Split squares (partners – organizations whose products or services help to enhance demand for those of the firm of interest).
- Relationships: Where solid lines between participants indicting a primary relationship and dotted lines an electronic relationship between the parties.
- Flows: Where arrows represent the major flows between participants and can either be money ($), a product or service, digital or physical (0) or information (i).

**Business model for Company A**

The company sees itself as having a hybrid business model that involves publishing and aggregating largely on a business-to-business basis. It began basically in 1989 as a cost-recovery model, but since 1997 it has operated as a commercially sustainable (but not-for-profit) publisher and aggregator. It contains elements of all the four atomic models listed above. Figure 1 shows an overview of its business model.

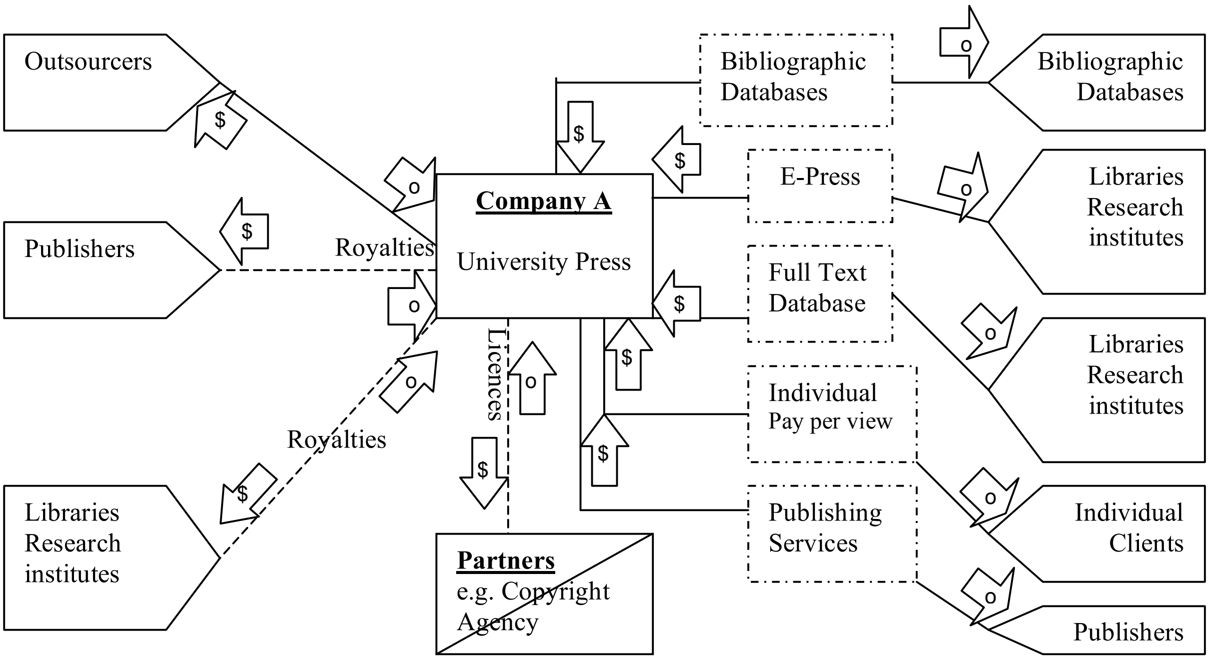


**Figure 1: The business model for Company A**

**Business model for Company B**

The business model for Company B is based largely on the traditional market for textbook sales, but again it contains elements of at least three of the atomic models listed above. Figure 2 shows the business model for Company B.

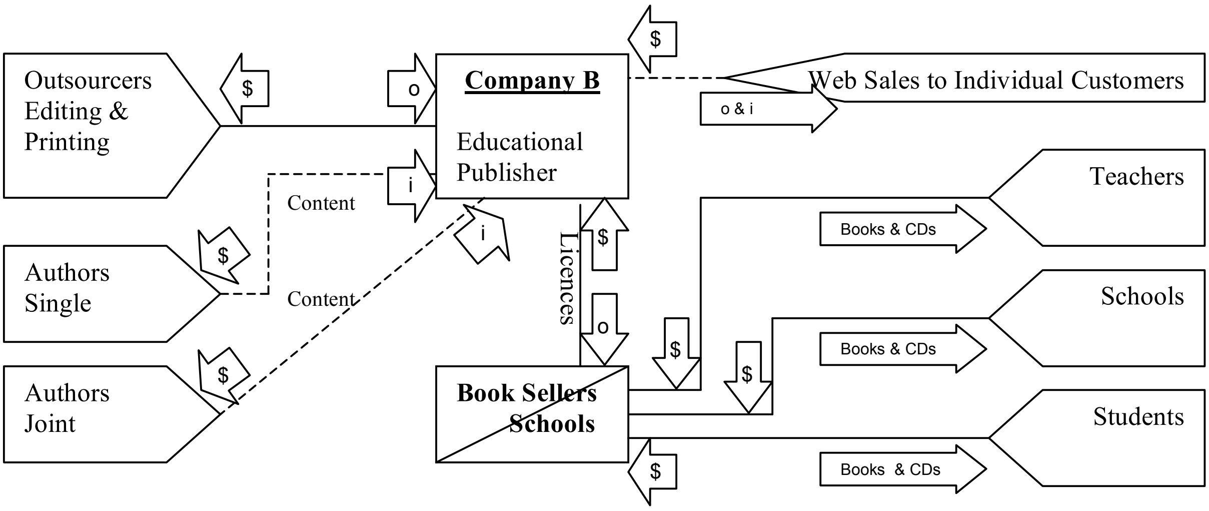


**Figure 2: The business model for Company B**

**Business model for Company C**

The business model for Company C is largely that of a full service provider, with elements of direct to customer and content provider models included. Essentially Company C sells publishing services to conference attendees including peer reviewed publication of single papers and sales from an online book store. Figure 3 presents the current business model for Company.

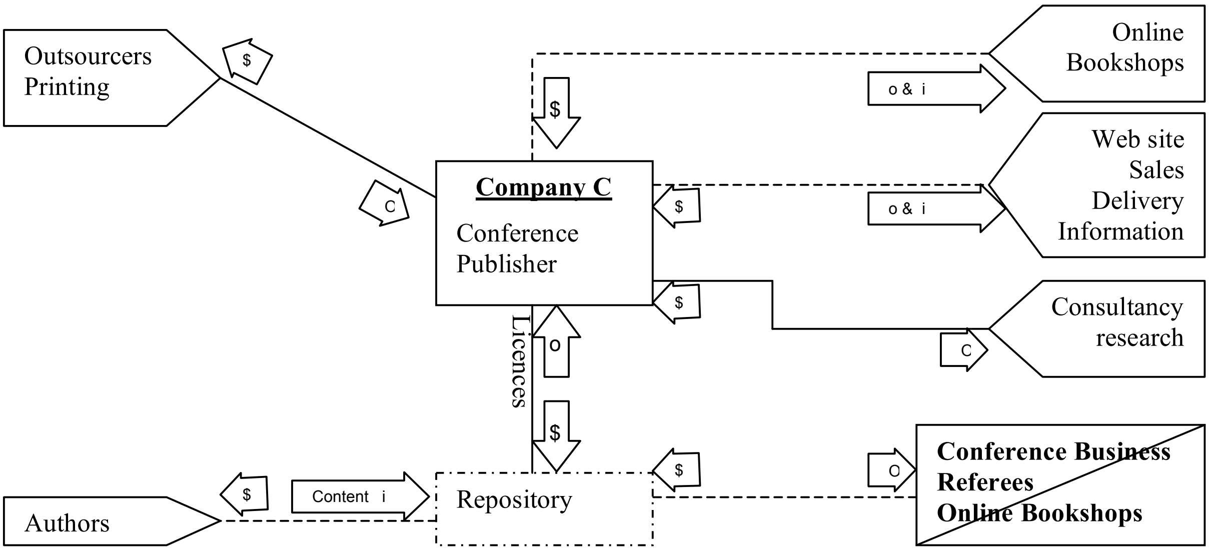


**Figure 3: The business model for Company C**

# 6 Conclusions

What has been reported here are findings from three of what will ultimately be eight case studies seeking to identify current and future business models for book publishing in Australia. The case study protocols, the structure of the interviews and the nature of the questions posed were all determined by feedback from focus groups and a national online survey. On the basis of what has been learned from the six cases conducted until now, the researchers perhaps over-estimated the likely impact of technology on the thoughts and deeds of publishers, while underestimating the continued popularity of the printed book. To some extent this is not so apparent in the context of the three cases reported in this paper. Companies A and C are major users of leading edge technologies and see the future very much in terms of the exploitation of technology for business sustainability. Company B a highly successful and profitable multinational publisher of educational texts, remains much more focused on traditional perceptions of value and on channels for its delivery, while maintaining a careful watch on market developments. For Company C this already entails the ability to respond to what for it is a minority demand for digital content, and evidence from the other three cases not covered in this paper suggests that publishers are *hedging their bets* to the extent that many of them have a growing presence in markets for digital products and services. This is certainly the case for example, with the industry partner for our research project (CCH Australia), which while regarding itself as a traditional publisher operating in niche professional markets, nonetheless generates up to one-third of its revenue from digital sources. The overall conclusion, therefore, is that publishers are *making haste slowly* in response to the potential inherent in digital technologies, whose potentially disruptive presence is more than balanced by a range of organizational, commercial and market factors.

# References

[1] COPE, B; KALANTZIS, M (2002), Managing Knowledge: Communication, Learning and Organization Change. In Cope and Freeman eds, *Developing Knowledge Workers in the Printing and Publishing Industries*, Common Ground Publishing, Melbourne.

[2] COVEY, DT (2003), Copyright Permission: Turning to Dust or Digital. *International Journal of the Book*, Volume 1, Common Ground Publishing Melbourne.

[3] DUBE, L; PARE, G (2003), Rigor in Information Systems Positivist Case Research: Current Practices, Trends and Recommendations. *MIS Quarterly*, 27, 4, 597-635.

[4] DAVIS, M; WALTER, M (2003), Next-Wave Publishing Technology: Revolution in Process and Content. *Seybold Publications*, 3, 15, 3-15.

[5] KLEPER, M (2001), *The State of Digital Publishing. Prentice Hall PTR.*

[6] JANSEN, B (2003), The Future of the Book: Format and Technology. *International Journal of the Book*, 1, Common Ground Publishing Melbourne.

[7] MASON D; COPE, B (2001), Australian Book Production in Transition. In Cope and Mason eds, *Creator to Consumer in a Digital Age: Australian Book Production in Transition*, Common Ground Publishing, Melbourne.

[8] MOLLA, S; HEEKS, R; BALCELLS, I. (2006), Adding Clicks to Bricks: A Case Study of e-Commerce Adoption by a Small Catalan Retailer. *European Journal of Information Systems*, 15, 424-38.

[9] OLIVER, S; KANDADI, K (2006), How to Develop Knowledge Culture in Organizations: A Multiple Case Study of Large Distributed Organizations. *Journal of Knowledge Management*, 10, 4, 6-24.

[10] ROWLANDS, I; NICOLAS, D (2005), New Journal Publishing Models: An International Survey of Senior Researchers. *A CIBER report for the Publishers Association and the International Association of STM Publishers*, London, CIBERhttp:www.ucl.ac.uk/ciber-pa-report.pdf

[11] WALSHAM, G (1995), Interpretive Case Studies in IS Research: Nature and Method. *European Journal of Information Systems*, 4, 74-81.

[12] WEILL, P; VITALE, M (2001), *Place to Space: Migrating to E-Business Models*, Boston, Harvard Business School Press.

[13] TIMMERS, P (1998), *Business Models for Electronic Markets*. European Commission, Brussels, http: //webarchive.org/web/20030612192921/http://lists.commerce.net/archives/eco-wg/199901/msg00010.html.

# Libraries as Publishers of Open Access Digital Documents: Polish Experiences

*Marek Nahotko*

Institute of Information and Library Science, Jagiellonian University
ul. Gronostajowa 7, 30-387 Kraków, Poland
e-mail: nahotko@inib.uj.edu.pl

## Abstract

This article presents the experience of Polish libraries in the field of electronic publishing. There have been described some solutions applied for creating digital libraries and institutional repositories. Nowadays, Polish libraries seem to be passing from the stage of digitalization of their own collections (usually of historic value) to publishing new digital-born documents in their own institutional or multi-institutional repositories. This activity should be (and is) developed in co-operation with university press companies.

**Keywords:** digital library; open access; electronic publishing; Poland

## 1 Introduction

In Poland, a large number of societies dealing with research communication (authors, librarians and users – readers of scientific publications) are satisfied neither with the operation of today's research communication system nor information exchange. Besides, negative effects can be noticed not only in research communication, but also in other information processes and communities, what is visualized, for example, in a constant decrease in the level of books reading at public libraries. The reasons for the dissatisfaction are, among other things, a sharp increase in the prices of publications, copyright-related issues, problems pertinent to intellectual property as well as a still longer and longer time interval between arriving at research results and their publishing. Polish libraries are active participants in the discussions held about the items said, and they seek to extenuate the problems, which come up and to submit some proposals of practical solutions aimed both at reforming the system of publishing and relative processes. Their actions are chiefly focused upon the items connected with access to the resources. The first goal is heading off the 'crisis of journals'; the problem consists in fighting against prohibitive prices as they make it considerably harder for one to find the desired publication. The other is the limitation of effects brought about by 'access crisis', which means not only limitations in permanent access to documents already published, e.g. by preventing one from gaining access to scientific electronic journals when the pertinent license has expired, but also impediments to having access to older publications, still esteemed by users.

## 2 Methodology

In this article have been utilized the data collected during research into Internet websites, dedicated to projects related to electronic publishing, whose initiators are mainly libraries operating on *dLibra* software. Then, there has been also presented a case study of a digital library which, due to a relatively long period of operation, its experience and achievements, is a good example for illustrating the trends, which nowadays dominate in Poland in the field of e-publishing performed at libraries. While analyzing the problems under presentation, there has also been made use of interviews held with librarians – authors of new forms of publishing; the Internet bulletin board was also used for this purpose. The analysis has covered as well the relation of publishers of Polish scientific journals towards the idea of electronic availability of their publications. The journals placed on the list of the so-called score journals by the Polish Ministry of Research and Sciences have been marked out for the analysis said. An author publishing in such a journal is awarded proper score, highly appreciated when his research achievements are subject to evaluation. According to the Ministry, those are top level journals in their fields; therefore, their publishers should take care of spreading the contents under publication also electronically. Hence, it might be supposed that the position of journals not included on this list must be yet worse.

# 3 Results

Recently, some initiatives related to electronic publishing, and first of all, to making digital libraries, have appeared in Poland. This phenomenon is typical for the library sector at the beginning of the 21st century. Not only do libraries collect traditionally their resources and render them available to the public, but they also take over some new tasks, for instance, electronic publishing of documents. On the turn of the 20th and 21st centuries, in many forums (for example, at numerous conferences), Polish librarians debate the issue of involving libraries in electronic publishing. Nowadays, another stage which consists in the implementation of practical solutions, has commenced. Initially, attention was focused on the digitalization of libraries' own collections, mainly for their protection and archiving. Later on, there appeared also some projects aimed at electronic publishing of newly born documents, usually born-digital. Such documents are published in digital repositories through the mediation of librarians who administer them.

**Position of libraries**

In Poland, libraries started their actions related to building digital libraries with the digitalization of their own collections at the beginning of the 21st century. The survey carried out in 2003 showed that the digitalization was performed in 25 libraries, and in 14 of them were special purpose-built laboratories [1]. According to the latest information collected (end of 2006), 115 libraries deal with digitalization – 51 of them are at university level schools, including almost all university libraries (16) and technical university libraries (11); instead, 48 of them are public libraries.

Initially, most initiatives were short-term tasks, whose purpose was mainly taking immediate actions and performing services as ordered by users. In consequence, they turned out to operate with certain irregularities and were affected by various, often subjective factors (equipment, staff, finances). Some libraries established special divisions within their organizational structures; others preferred services performed by external companies.

As digitalization was developing, it was necessary to made certain decisions on the selection and choice of materials for digitalization. Those problems were often solved by library managers and specialists, employed mainly at the divisions dealing with special collections, the collection acquisition and circulation as well as with it protection and preservation. In most libraries, digitalization programs cover first of all old prints, manuscripts, incunabula and 19th century journals and magazines. The main purpose of the actions mentioned was not only protection of valuable resources, but also meeting the new needs of library users.

Reasons for digitalization of library collections:

- The necessity of protecting the collections possessed against destruction and making them accessible to a large number of readers; those collections are of high value from the viewpoint of cultural heritage;
- The need of rendering university press books, textbooks and other learning materials more and more accessible, as well as of making them adapted to use in distant learning (e-learning);
- Readiness for involvement in the promotion of university level schools through popularization and spreading of research and culture potential as well as intellectual production of university staff (series, journals);
- The necessity of becoming active participants in the national strategies eEurope, ePoland and in the strategy UE i2010, as well as the need of participating in the initiative of digitalization of collections of the top Polish libraries.

Lastly, the digitalization process of the resources held in Polish libraries has been quickening its pace. The libraries plan to maintain this pace in the nearest future. Among other things, this process is favored by a reduction in prices of IT hardware and services.

Anyway, the quality of digitalizing operations still needs improving. A part of electronic versions have been compiled from poor quality materials, e.g. old microfilms, which will cause them to be useless soon and the digitalization process will have to be repeated. Another unsettling information is the lack not only of any uniform standards for recording and archiving digital documents, but also of certificates to determine the durability of digital records media (some collections are recorded on CDs). Therefore, neither the future nor the accessibility of such resources is certain.

In order to coordinate the works and to secure a more close co-operation among those libraries which render digital publications accessible, there has been established a consortium called Digital Library Alians [2]. Its aim is to develop and to intensify the actions related to acquisition, presenting and popularization of digital resources connected with both the cultural heritage in different Polish regions and research resources produced at respective Polish university level schools. In consequence, the co-operation and the funds raised due to joint efforts should lead to a development of regional networks of digital libraries which are supposed to constitute a stable structure, and to an unification of standards and an optimization of the solutions adopted.

Polish libraries – both academic and public – participate in electronic publishing in two ways, by:

- Making digital libraries - which contain documents digitalized (scanned) from originals, collected and stored in traditional (printed) library resources. Originals are often of value, and access to them is hindered. Those are usually historic documents to which the copyright is no longer applicable;

- Creating repositories of digital documents (articles, PhD, MA theses, reports etc.), whose authors are researchers employed at institutions provided with a repository. Those are documents for which the copyright provisions are applicable.

In both cases said, libraries become publishers and editors of electronic documents; in most cases those documents are digitalized copies of traditional publications. So far, in neither case we can say about any traditional roles played in both fields by libraries. Therefore, because of new roles involved, librarians face many new problems to be solved in the matter of new technology and organization application.

There can be distinguished two organizational models of libraries and/or digital repositories in Poland:

- institutional (academic libraries, Polish National Library), including 7 libraries working on dLibra software;

- regional models, focused mainly on major university level schools, sometimes also on regional public libraries; they consist of 2-23 institutions, chiefly libraries, but also of museums and archives.

**dLibra Software**

Most initiatives presented so far are based on Polish software called dLibra, compiled at the Poznan Supercomputing and Networking Center (http://dlibra.psnc.pl/). Nowadays, this software is applied in a few dozen libraries (Fig. 1). This software serves for professional making of collections of digital objects. It allows to collect and to render digital objects available in Internet in various formats (eg. txt, html, pdf, djvu). Each object may consist of any number of files and is described with metadata (MARC, Dublin Core etc.). Each implementation of dLibra software includes the three main elements working in the client-server configuration (Cf Fig. 2):

- The server of the digital library/repository, responsible for the performance of all library functions, usually operating on the dedicated hardware, not accessible directly to end users;

- An application of the editor and administrator (client) which allows them to enter digital objects, their descriptions and execution of other similar functions:

- End user’s application (client), based upon Web interface, and allowing one to have customized access to the objects within a collection.

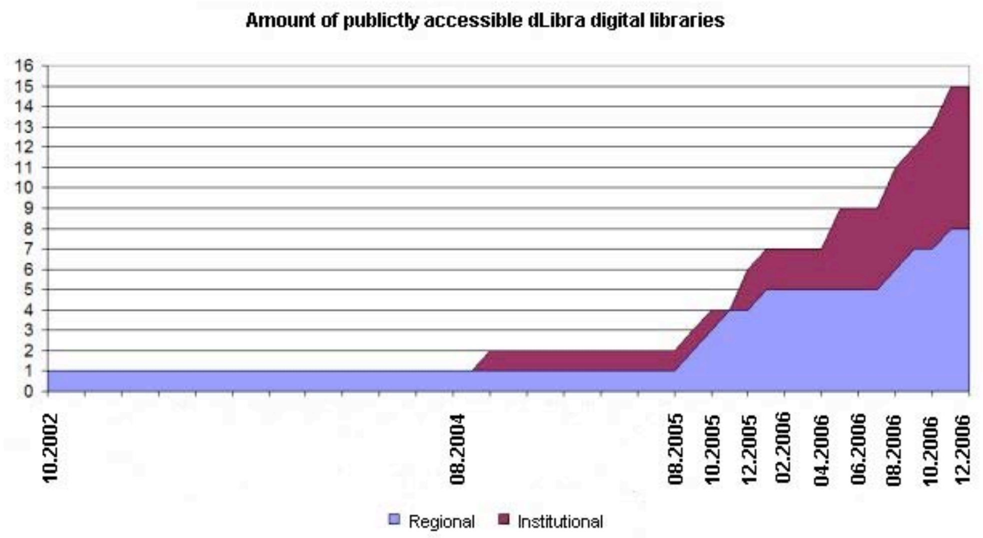


**Figure 1: Progress of dLibra digital libraries in Poland**

dLibra allows us to implement the majority of international standards, presently under application worldwide, for example RSS, XML, RDF, MARC, Dublin Core or OAI-PMH. It can be upgraded by independent programmers under condition of free access to the newly prepared software.

Publications are placed to the system by their authors directly or with librarian's intermediation. Any author of a publication can modify texts previously compiled, which leads to their new editions. Those editions consist of files which also may have various versions. Editions, in turn, can be published or not; they can be also made accessible for a certain time until the fixed date.

It is also possible to make group publications serving, in turn, for combining single publications which have some common features e.g. successive journal issues or series. Within a group publication may operate other groups, too. Each group is provided with its own description. Publications may be grouped into collections. Each publication may belong to more than one collection. In the case of assigning a group publication to a collection, all publications within a group are automatically assigned to this collection. Collections may be divided into sub-collections, which leads to a tree structure. Collections are provided with their own descriptions, also copied to a sub-collection with a possible modification.

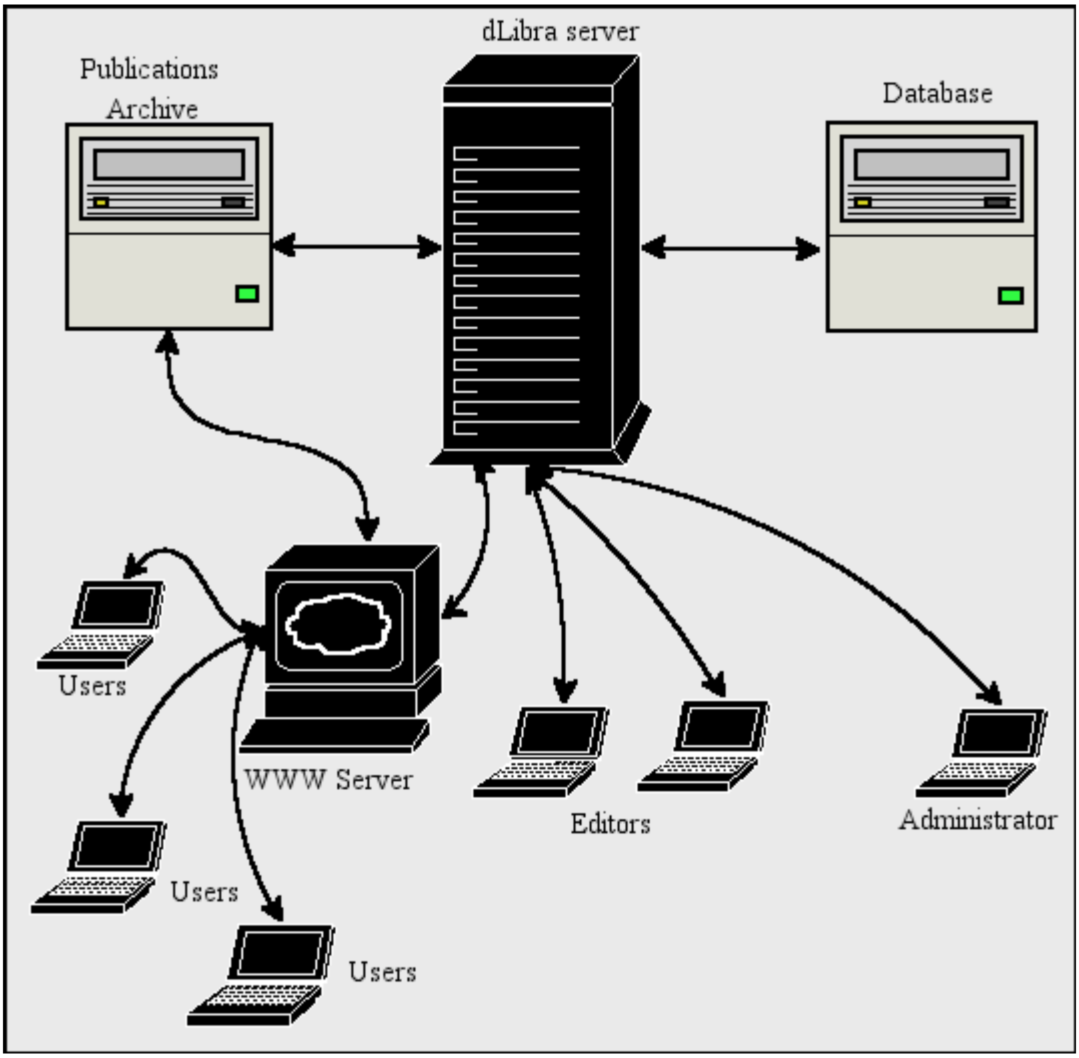


**Figure 2: dLibra client-server architecture**

Publications collected by dLibra software are indexed by popular search tools, like Google; hence, those are not resources of any hidden (invisible) Web. The architecture designers paid much attention to indexing of descriptions of digital objects by search engines, which has led to quite a good effectiveness. Another functionality, very important for the user, is a possibility of searching the contents of all dLibra resources from the level of each system implementation (Cf Fig. 3, option: 'Search remote libraries'). In consequence, irrespective of the library the user has chosen at the beginning of his/her search, he/she can search all digital libraries consisting one network, with one search tool [3].

**Case Study – *Kujawsko-Pomorska* Digital Library**

One of the oldest and largest digital libraries in Poland is the Kujawsko-Pomorska Digital Library (KPDL). In order to set it up, in 2003 there was established a consortium of libraries led by the Nicolaus Copernicus University in Torun (north-western part of Poland) and its library, with participation of other two regional high schools. It is also planned to co-operate with local public libraries. Each cooperating institution places their own digital resources on a joint platform, and administers them in the scope of compilation, updating and access rules. KPDL is a part of the Project of building open information society ePoland, which in turn constitutes a part of eEurope. EU's financial share was 75%.

The resources of KPDL digital objects consist of three collections:

- Research and teaching collection aimed at improving the quality of teaching by securing access to digital copies of textbooks, monographies and research articles;

- Cultural heritage collection which is to include the most valuable rare books, manuscripts, books published in the 19th and 20th century, to archive records, music notes, emigration, cartography and iconography collections;
- Regional records which will include publications, articles and occasional materials on the history of the Region of Kujawy and Pomerania [4].

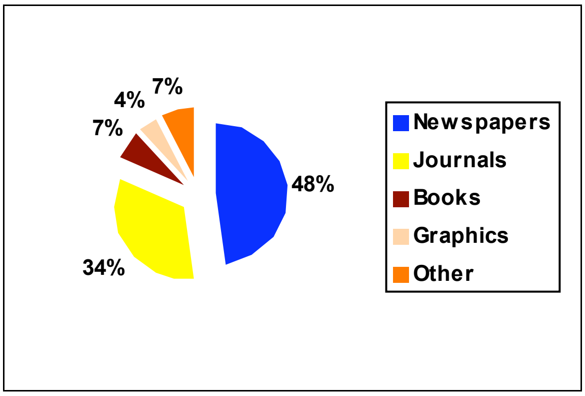


**Figure 3: Kinds of objects in KPDL**

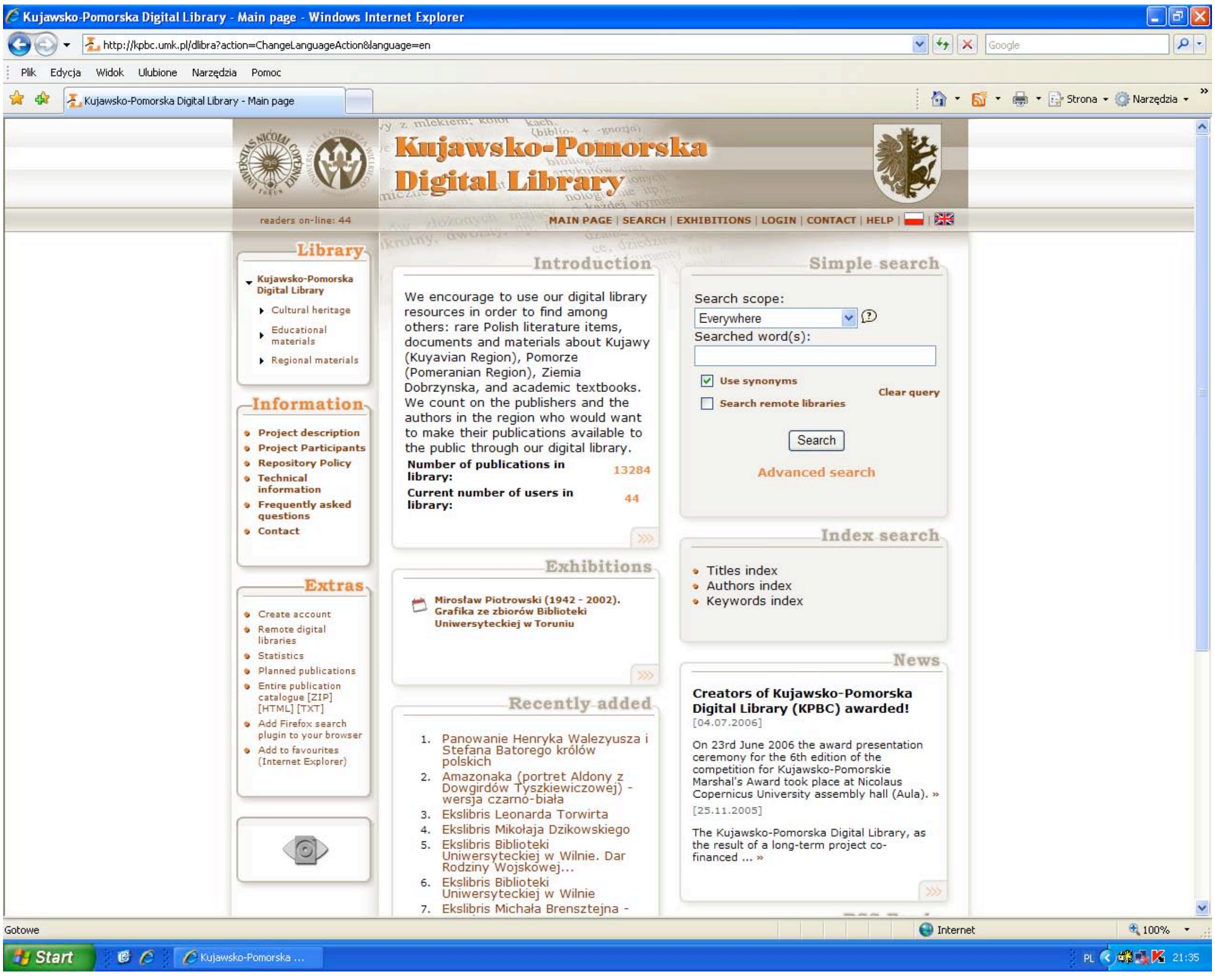
Collections are divided into smaller groups as needed. Materials are assigned for digitalization by three libraries, which bear joint responsibility for the KPDL resources, namely two academic and one medical (Cf Fig. 3). Access is first of all given to teaching materials in the field of medicine (university notebooks, journals and monographies published before 1945, and a self-published journal 'Biological and Medical Sciences'). Other branches of science are represented by regional historic journals, subject to digitalization in co-operation with two chief regional public libraries. Instead, the Nicolaus Copernicus University Library offers access to the sources on the history of the region, emigration collections, engravings and Vilnius records [5], so much essential for the history and tradition of this University. The Library also digitalizes teaching materials (e.g. set books for philologists), not subject to copyright restrictions. So as to avoid repetitions, the lists with materials to be digitalized are agreed upon both electronically and at monthly meetings of the editing staff.

The co-operating university level schools publish as well their own, contemporary materials and research papers. The authors of such works sign proper licence agreements in which they may reserve the range of access to their work: no limitations in the whole Internet, at their own university only or access to the users of consortium libraries only. On the same basis, they give their consent to the KPDL for electronic publishing of their texts. The authors hand over their works free of charge.

It was not difficult to select the software for the digital library under project, since libraries operating on dLibra software had been already in existence (Fig. 4). It was acknowledged that such a platform is provided with the fundamental functions, indispensable for any digital library: cataloguing and giving access to text and graphic files, searching of documents through any words taken from the description or contents of the document, collections management, navigation within a publication or limitation of access to a selected group of users. A significant feature of dLibra was the compatibility allowing one to work with the library system Horizon, used in the libraries of the region.

Before a publication can appear in the KPDL, it must go through certain stages (in the brackets are those who are responsible for their performance):

- Section and assignment of documents for digitalization (selecting librarians) based on rules as agreed upon;
- Compilation of objects ready for digitalization as the so-called list (selectors);
- Queuing of the documents assigned to digitalization and queue control (editor);
- Technical works on a document and its handover to the digitalization lab (selectors);
- TIFF format scanning and archiving (technicians);
- Processing of OCR scanned files into DjVu format (CT staff, technicians);
- Compilation of a bibliographic description for a local catalogue in Horizon system (MARC 21 format), conversion into dLibra (Dublin Core format) (catalogers);
- Publication in the digital platform (editor);
- Control of metadata in dLibra and Horizon systems, amelioration of the resources and possible corrections (main cataloger) [6].


Kujawsko-Pomorska Digital Library - Main page - Windows Internet Explorer
http://kpbc.umk.pl/dlibra?action=ChangeLanguageAction&language=en
Plik Edycja Widok Ulubione Narzędzia Pomoc
Kujawsko-Pomorska Digital Library - Main page
Strona Narzędzia
MAIN PAGE | SEARCH | EXHIBITIONS | LOGIN | CONTACT | HELP
Gotowe
Internet
100%
Start
Kujawsko-Pomorska ...
PL
21:35

Kujawsko-Pomorska Digital Library

readers on-line: 44

Library

- Kujawsko-Pomorska Digital Library
  - Cultural heritage
  - Educational materials
  - Regional materials

Information

- Project description
- Project Participants
- Repository Policy
- Technical information
- Frequently asked questions
- Contact

Extras

- Create account
- Remote digital libraries
- Statistics
- Planned publications
- Entire publication catalogue [ZIP] [HTML] [TXT]
- Add Firefox search plugin to your browser
- Add to favourites (Internet Explorer)

Introduction

We encourage to use our digital library resources in order to find among others: rare Polish literature items, documents and materials about Kujawy (Kuyavian Region), Pomorze (Pomeranian Region), Ziemia Dobrzynska, and academic textbooks. We count on the publishers and the authors in the region who would want to make their publications available to the public through our digital library.

| | |
|---|---|
| Number of publications in library: | 13284 |
| Current number of users in library: | 44 |

Exhibitions

Mirosław Piotrowski (1942 - 2002). Grafika ze zbiorów Biblioteki Uniwersyteckiej w Toruniu

Recently added

1. Panowanie Henryka Walezyusza i Stefana Batorego królów polskich
2. Amazonaka (portret Aldony z Dowgirdów Tyszkiewiczowej) - wersja czarno-biała
3. Ekslibris Leonarda Torwirta
4. Ekslibris Mikołaja Dzikowskiego
5. Ekslibris Biblioteki Uniwersyteckiej w Wilnie. Dar Rodziny Wojskowej...
6. Ekslibris Biblioteki Uniwersyteckiej w Wilnie
7. Ekslibris Michała Brensztejna -

Simple search

Search scope: Everywhere
Searched word(s):
☑ Use synonyms
☐ Search remote libraries
Clear query
Search
Advanced search

Index search

- Titles index
- Authors index
- Keywords index

News

Creators of Kujawsko-Pomorska Digital Library (KPBC) awarded!
[04.07.2006]
On 23rd June 2006 the award presentation ceremony for the 6th edition of the competition for Kujawsko-Pomorskie Marshal's Award took place at Nicolaus Copernicus University assembly hall (Aula). »
[25.11.2005]
The Kujawsko-Pomorska Digital Library, as the result of a long-term project co-financed ... »

**Figure 4: Main page of KPDL**

Procedures compiled for respective stages provide the following actions:

1. Assignation of documents:
a. determination what should be digitalized and how to do it,
b. compilation of lists with no more than 15 items, with detailed data for processing (scanning quality, color, and others),
c. uploading of lists on a joint disk accessible in Intranet.
Responsible: librarians responsible for respective collections.

2. Queuing of documents (lists):
a. queuing of lists, priority assignation, setting the sequence of scanning,
b. keeping the lab informed about any queue and that the performance of a task is possible,
c. hand-over of lists to the digitalization lab,
d. constant control of the digitalization process and of the compliance with the procedures.
Responsible: KPDL Editor

3. Scanning and archiving:
a. a staff member orders documents for scanning by contacting the person who is signed under the list,
b. the staff member signs the list for a librarian who supplies the materials, and the former considers the list as a lending form,
c. scanning (in the lab) of documents as queued on the list; any remarks about the scanning result are addressed at the editor,
d. processing of the scanning result with the programs enclosed to the scanner (picture correction, framing etc.),

e. entering the file under a standard name (shelf number is a file name or folder name for many files); assignation of catalogue numbers to the box with the carrier of the archival digital version of the document; physical description of the archival digital version (file format, carrier type and recording date, resolution, color detail level etc.),
f. transmission of final files for further processing,
g. delivery of materials to the division of formal working against receipt, like in b/.
Responsible: Digitalization lab staff.

4. Processing of the files created in the scanning process:
a. obtaining of file formats (DjVu, HTML, PDF and others) as planned in the process of assignment and preparation; OCR for some objects,
b. handover of files to the editor for further operations.
Responsible: Digitalization KLab Staff, CT specialists.

5. Publication on the digital platform:
a. combining an object with a description,
b. uploading of files either to one or more dLibry collections as indicated,
c. supplementing of a description in dLibra with the archival version metadata, if any.
Responsible: KPDL Editor.

6. Compilation of a bibliographic description:
a. cataloguing in Horizon,
b. placing final descriptions in the target dLibra collection,
c. constant amelioration and quality control of descriptions in dLibra,
d. making corrections to dLibra, review of indices,
e. handover of books to the bookstacks or to the reading room as specified on the list.
Responsible: librarians responsible for respective special collections.

7. Control of metadata and transmission of a description to NUKAT [7]:
a. entering of the new data related to a digital object to the existing records in NUKAT database,
b. compilation of new records with a re-routing to a KPDL object,
c. combining of existing records in NUKAT database.
Responsible: chief cataloger.

According to the list said, the respective tasks are carried out by a team consisting of various members who have different levels of skills and qualifications:

1. Project coordinator – administration, finances, cooperation with partners, promotion, content and quality supervision, negotiations on copyright with authors.

2. Coordinator's deputy – supervision over the CT part of the project, hardware, software, contracts with suppliers, tenders, standards.

3. Administrator – project files, finances, reporting, personal matters, correspondence and others.

4. Editor – edition of digital library objects, idea of resources and its administration, coordination of works of the team which compiles and enters documents.

5. CT specialist – software, supervision over dLibra software, engineering solutions, statistics.

6. Chief selection specialist – selection of documents for digitalization, selection of materials from special collections, work coordination.

7. Chief cataloguer – bibliographic description, metadata, amelioration of entire resources, standards.

8. Technicians – digitalization, supervision over the lab, scanning standards, objects archiving.

**Situation of the university press publishers**

Nowadays, in the process of research paper publishing, the role of the author of a publication, viz. the compilation of a text becomes the easiest one. But problems will start soon after. First, one must get some funds (grant) for publication. When finally apportioned, this money turns out to be halved. That is why the publisher usually refuses any royalty for the author, considers the entire project as an unprofitable task, which has no positive effect upon the development of any enterprise. All advertising and marketing activities spell only more expenses; that is why from the publisher's viewpoint the best solution would be withdrawing the item already published from distribution at all. Other activities carried out by researchers in the course of the publishing procedure, like preparation of reviews are also performed free of charge, which yet worsens the unfavorable situation mentioned.

In addition, access to the information is hindered by improperly arranged book trade. For being commercial entities, bookshops do not deal with academic books as usual [8]. Their activities are targeted at mass consumers who, for example, purchase such items, like Harry Potter; instead, the sale of single copies of scientific texts goes beyond the boundaries of commercial risk.

A large number of university press publishers try to send their items by themselves, often through Internet bookshops; however, not all of them resort to such a solution. Then, there will also appear other problems, for instance, mail-order sale of a low circulation book whose publishing has been refunded, and in consequence, the publisher has already got their profit; such a situation may be seen as a contingency, not profit. University press publishers divide their items into those they have got to publish at cost due to their role played in academia, and those on which they can gain some profits. Of course, professional marketing refers to the latter.

In turn, this means that in the process of publishing of academic items it is necessary to find a new solution in which libraries, especially academic ones, may and should actively participate. Academic libraries begin to take over the functions pertinent to university press publishers. Practically, those publishers are by definition non-profitable entities, always in deficit. This will lead to establishing a kind of electronic library publishing houses, a part of which will become electronic repositories of publications supposed to be used free of charge within a reciprocally advantageous cooperation held with other institutional and regional repositories. Due to such a cooperation, it will be possible to economize on publishing and by giving access to a large number of non-commercial (yet valuable) low circulation publications. Eventually, a large number of university press publishing houses will become superfluous. Their today's number arises from the fact that each university level school/ college, however small it may be, has an ambition to have their own publishing house. In consequence, there are many microscopic publishing companies, and in many cases their professionalism and potential are exiguous.

Moreover, the process of publishing academic texts will be accelerated; as of today, it takes years to bring them out as publications. Such a situation results in part from the top-down order coming from the government agencies in the matter of having the quality of research and academic publications evaluated. The aftermath of such an evaluation is the list of scientific journals as published by the Polish Ministry of Sciences; any author who has published in them is assigned a score enhancing the evaluation of his research achievements. Therefore, the editors of the magazines from this list have their hands full for a few year time or more.

# 4 Discussion

As a result of the actions taken by all persons and institutions involved in contemporary research communication, one can notice a change in the roles assigned in the process of making electronic publications. After a short time, indispensable for preparing such changes, we can expect serious modifications to the operation of publishing companies and assignments performed by their staff; nevertheless, such modifications will also refer to librarians and book dealers. In a more or less conscious way, representatives of those professions get ready for the changes and modifications to come soon. Such changes will also refer to scientific community. Authors become editors and publishers; instead, publishers deal with the aggregation of contents and contribute with their own value added. Librarians turn into digital librarians, which is related to their participation in electronic publishing.

According to the actions enumerated and reported in the case study, the establishment of a repository administered by the traditional library or consortium of libraries entails changes to the organization because, those take over new functions. A part of tasks performed in the digital library demands only that librarians should change their way of working and their habits (e.g. transit from MARC 21 cataloguing to metadata, like Dublin Core). Other actions entail completely different skills, so far typical rather for publishing companies than libraries, e.g. compilation and edition of digitalized contents.

Differences between traditional and electronic publishing can be described by dividing each research communication process into four stages at which concrete functions are performed:

- Description of the idea and conceptions arising from conducted research;
- Certification of values of the described ideas and research results;
- Distribution of ideas and results by making them accessible to prospective readers interested in them;
- Archiving of results so that they might be utilized successively.

It is evident that such functions are performed by each system of research communication, either traditional or contemporary, based upon new digital technologies.

| Function | Process | Performed by: | Financed by |
|---|---|---|---|
| Registration | Delivery of a digitalized text | Librarians | Repository |
| Certification | Review | Researcher – reviewer | Publisher of the printed original |
| Circulation | Open repositories | Librarians | University level schools, local government |
| Archivization | Permanent access | Librarians | University level schools, local government |

**Table 1: Model of Polish repositories**

In compliance with the model said, one may state that Polish initiatives are endowed with a certain, separate set of features if compared with similar initiatives developed in other countries. In the Polish model, authors do not provide repositories with their latest publications; instead, old items (often 100 and more years old, to which the copyright provisions are inapplicable), are supplied by the librarians who have scanned them. Since Polish repositories (mainly regional) are often established by public libraries, they are financed from local government budgets.

Eventually, Polish repositories are dominated by archival resources which consolidate the role of the library in the field of archiving and museum functions, but such repositories do not play the main role usually attributed to them, since they make no contribution to the acceleration of research communication. Such a situation may result from the absence of agreements in the field of author's rights and copyright. It is still unclear which solutions will be adopted. Especially, two of them are under consideration: solutions as applied in Wikipedia, and Creative Commons. As of today, respective institutions make their own agreement/ contract models to be used while receiving texts from their authors to be published in the repository.

A solution of today's difficulties related to research communication in Poland seems to be a cooperation between libraries and university publishers so as to make an Open Access publishing system. The base for such a cooperation should be a modification to today's Polish model of academic/research publishing. Scientific institutions publish their own items (mainly journals and series) to be exchanged for those from other institutions (also from abroad). Many scientific centres convert their published items into the electronic form. Those journals are usually of non-commercial nature – they are financed from different sources – directly by government agencies and research institutions.

The transition from the traditional library via digital library to the digital publisher is a part of processes leading to a development of digital research and science. It should facilitate innovativeness, by making new ways of production and popularization of research results. Due to new IT methods and technologies, it is possible to make research communication more streamlined at all its stages – from making texts, via evaluation of their quality, administration, circulation and archivization.

For assessment of Polish academic publishers, of interest could be also some data regarding scientific journals published by small university press publishers and research societies. Parallelly be stressed that practically all such journals are brought out by small publishing companies, since there is no large, commercial publishing house which deals with this business. Among the journals placed on the polish government list (available on website WWW: http://www.nauka.gov.pl/mein/_gAllery/13/66/13662.pdf), one may find top score items; if you publish in them, it will contribute greatly to the evaluation of your academic achievements. About 75% of the journals have their own websites. As far as journals with websites are concerned, in 48% of cases, one may gain access to full texts (Cf Fig. 5). Instead, 52% of them place on such websites contents and/or abstracts (sometimes very general data). It means that in Poland the process of making the contents of scientific journals available to the public in electronic mode has already started, but in this field we are still behind the level of 80-90% of journals available online, as in Western Europe and USA.

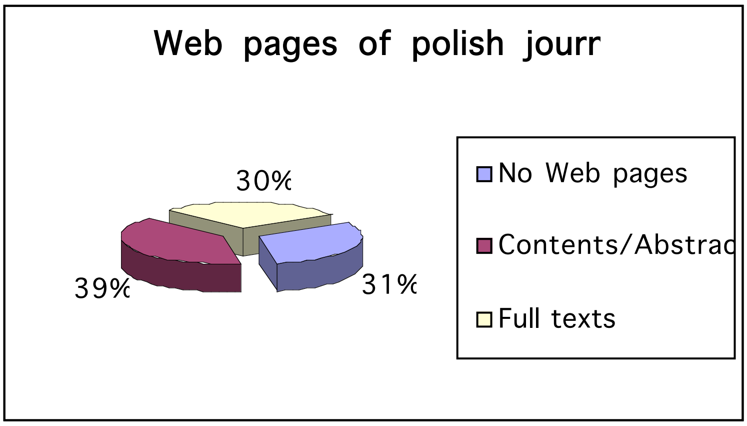


**Figure 5: Accessibility of Polish journals on the Web**

A large part of those publications is published in English. As far as the magazines from the government list, available in Internet, are concerned, 75% of them are available in English. They are made accessible in wide networks, which can thus secure a possibility of their popularization (wide distribution), also due to the absence of a language barrier. In particular, this refers to journals dealing with medicine, sciences and mathematics.

Quality control often consists in selecting materials by the editors; therefore, there is often no typical reviewing. However, this is not any problem in small circles of specialists in narrow fields of science, since those specialists know each other. In such circles, it is easier to perform quality control according to the reputation of respective researchers and research institutions (faculties, institutes).

## 5 Conclusions

In Poland, there have appeared some new initiatives related to electronic publications. One of them is the initiative of Interdisciplinary Center for Mathematical and Computational Modelling (ICM) [9] in Warsaw, and the Library of the Warsaw University, which proposes the establishment of a national repository of research texts named DIR. The repository could include as well the collections from the digital libraries already existed. The core of the project is to be the Virtual Library of Science, already in operation (http://vls.icm.edu.pl). The cooperating institutions might provide DIR with their own electronic documents (scans or versions made as electronic) with the metadata added. The model of cooperation assumes the storage digital objects by various ways:

i. In DIR, only;
ii. In local repository (digital library), only;
iii. In both places at the same time.

In ICM-DIR, materials would be given a final retouch, and their presentation in Internet could be secured. In consequence, there would be created a central, but scattered collection of Polish science accessible via one searching interface.

A model of open access repository of digital objects in conjunction with new publishing processes as performed in libraries proves to be more and more successful in Poland. The process of delivering the value added as a result of publication is not contradictory to the values of the open access idea. Experience shows they may complement and support each other, particularly when researchers try to expand and develop their research due to the application of new forms and possibilities of electronic publishing. The creation of digital libraries and repositories in Poland is an interesting example of the process of integration of digital libraries, repositories and publishing activities as is carried out by librarians in their traditional libraries. Those processes are under way now; therefore, not all their effects are known yet.

In such a way there appears a change in the publishing paradigm, like in the period following the invention of printing, when texts previously available only in manuscripts went to print on a large scale. Materials previously available in print are nowadays digitalized so as to be included in the worldwide resources of digital objects. At the same time, new born-digital objects are being collected, and their availability will be higher because the original (author) version of all publications is digital and has been ready to use for a long time.

## Notes and References

[1] KOWALSKA, M. (2006). Digitalizacja zbiorów w bibliotekach polskich – próba oceny doświadczeń krajowych. *Biuletyn EBIB*, 11(81), http://www.ebib.info/2006/81/a.php?kowalska

[2] ROŻNIAKOWSKA, M.; MARGAS, M. (2006). „eBiPol" – Biblioteka Cyfrowa Politechniki Łódzkiej na tle innych inicjatyw bibliotek cyfrowych w kraju od strony technicznej, formalnej i projektowej. *Biuletyn EBIB*, 4(74), http://www.ebib.info/2006/74/rozniakowska_margas.php

[3] KALOTA, T. (2006). Marzenie o polskim systemie rozproszonych bibliotek cyfrowych. *Biuletyn EBIB*, 4(74), http://www.ebib.info/2006/74/kalota.php

[4] CZYŻAK, D. (2005). Kujawsko-Pomorska Biblioteka Cyfrowa – stan zaawansowania realizacji projektu ZPORR. *Biuletyn EBIB*, 9(70), http://ebib.oss.wroc.pl/2005/70/czyzak.php

[5] Nicolaus Copernicus University in Torun continues traditions of polish university in Vilnius, existed since 1579. After 1945 it was removed to todays polish territory, as well as a lot of polish citizens.

[6] BEDNAREK-MICHALSKA, B. (2006). Kujawsko-Pomorska Biblioteka Cyfrowa – pragmatyka tworzenia biblioteki cyfrowej. *Biuletyn EBIB*, 7(77), http://www.ebib.info/2006/77/michalska.php

[7] NUKAT – National Universal Central Catalogue of scientific libraries in Poland (http://www.nukat.edu.pl).

[8] WOJCIECHOWSKI, J. (2006). Dostęp półotwarty. *Forum Akademickie* no. 11, pp. 24-26.

[9] ICM is well known in Poland becouse of central subscription of abroad scientific journals for polish libraries.

# Use of Open Access Electronic Journals by Chinese Scholars, and an Initiative to Facilitate Access to Chinese Journals

*Ruoxi Li[1];Fytton Rowland[2]; Zichuan Xiong[2]; Junping Zhao[4]*

[1]Chongqing Normal University, Shapingba District, Chongqing Municipality, 400047, Republic of China
e-mail: cc86@163.com
[2,3]Department of Information Science, Loughborough University, Leicestershire LE11 3TA, United Kingdom
e-mail: [2]J.F.Rowland@lboro.ac.uk; [3]shawnzec@gmail.com
[4]Tsinghua University, Beijing, People's Republic of China
zhaojunping@tsinghua.edu.cn

## Abstract

Surveys were carried out with two groups of Chinese scholars – one group working in China, and a second group working in the UK. The objective was to investigate usage of Chinese-language scholarly journals and the potential for them to use an Open Access business model. The results were compared with those published by the CIBER group at university College London, whose sample of scholars was international in scope. The overseas Chinese group made very little use of journals published in China, and one of the reasons for this was the difficulty of accessing the electronic versions of these journals from the West. We therefore proposed the construction of an English-language website to provide access in the first instance to the full texts of journals published by the members of the Society of China University Journals in the Natural Sciences (SCUJNS), and we created a pilot version of this website.

**Keywords:** open access; Chinese journals; overseas Chinese scholars

## 1 Introduction

This paper reports work undertaken while R.L. and J.Z. were Visiting Scholars at Loughborough University in 2006. In their home universities in China they are the editors of scholarly journals published by these universities, and as such they are active members of the Society of China University Journals in the Natural Sciences (SCUJNS), a collective organization representing such university-published journals in China. Many Chinese journals are published directly by universities in ths way, but a lot of them are little-known in the West. Not all of them have English-language abstracts or metadata.

Recently there has been an upsurge of interest in the West in the academic journals published in China, in recognition of the amount of research being conducted in that country and reported only in its own literature. As a result a number of papers have appeared in western journals describing the scholarly publishing scene in China [1-3]. Chinese publishers in their turn have shown interest in making their journals accessible to westerners through the English language, with one of the first such initiatives coming from J.Z.'s home university, Tsinghua [4]. Most Chinese journals are not Open Access at present but the subscription fees, tailored as they are to both Chinese cost levels and Chinese affordability, seem very moderate to Westerners.

As part of a programme of research investigating possible futures for Chinese-language scholarly journals published by universities in China, surveys were carried out of two groups of Chinese academics, one group based in China and the other group made up of expatriate Chinese scholars now working in the United Kingdom. The purpose of the surveys was to ascertain the knowledge of, and attitude to, Open Access (OA) journals among these groups. Differences between the home-based and overseas Chinese scholars, and between these groups and the general international group of scholars studied by the CIBER group at University College London [5] were also of interest.

Conversations with the overseas Chinese scholars showed that they made very little use of the Chinese-language literature published in China, either as authors or as readers, even if they had used it when previously resident in China. It appeared that this lack of use was in part a consequence of the Chinese journals being difficult to access from the West, even though many of them are available in electronic versions. We therefore carried out the pilot phase of a proposed operational website that would provide straightforward access from the West in the English language to Chinese electronic journals.

## 2 Surveys of Chinese Scholars

**Method**

Chinese scholars who had published papers in seven university journals published in Beijing, Xi'an and Chongqing were sent a questionnaire. About 3000 e-mail invitations to participate were sent to authors in China, and about 1000 paper questionnaires were also distributed. Over 500 responses were received, but exclusion of incomplete questionnaires from the survey reduced the fianl number analysed to 376, a response rate of 9.4%. In the UK, members of the academic or research staff at Loughborough, Nottingham and Sheffield Universities who had Chinese family names and personal names were approached individually and asked to take part. The majority of them were born in the People's Republic of China, all could read Chinese, and most were now permanently resident in the West; 50 responses from these these overseas Chinese scholars were received. The results from the China-based group and the overseas group were compared with each other, and both were also compared with those from the international group of authors surveyed by Rowlands *et al.* [5], the CIBER group, whose questionnaire we used. We are grateful to Dr Ian Rowlands for permission to use their questionnaire, and for helpful discussions.

**Results**

Of the group resident in China, computer scientists (24.5%) and engineers (28.5%) predominated, but many other disciplines were also represented. Their average age was 31.76 years. In all, 75% of them worked in Universities, and fewer than 1% in business or government. Engineers, mathematicicians and computer scientists also dominated the UK-based group.

More than one-third of both groups said they knew 'nothing at all' about Open Access (OA) , though more of the China-based group (29%) than of the UK-based group (16%) claimed to know at least 'quite a lot' about OA. Using a chi-squared test, the differences between the China-resident group and the overseas Chinese group were significant at the $p<0.05$ level, and those between the China-rsident group and CIBER's international group were significant at the $p<0.005$ level. The difference between the overseas Chinese and the CIBER respondents was not significant, however, possibly reflecting the more international orientation of the UK-based group compared with those who remain in China. The younger authors were more ignorant of OA, in contrast to the CIBER group's results, which found older scholars less knowledgeable, and this difference was significant at the $p<0.005$ level.

About three-quarters of the UK-based Chinese group associated the term 'Open Access' very strongly with 'free to access', a similar proportion to CIBER's international group, whereas only 45% of the China-based scholars thought that this was the defining characteristic of OA. Of CIBER's respondents, 47% did not associate the term 'OA' with 'author pays', but only 23% of our China group and 24% of our overseas group did not associate 'OA' with author pays. This difference was significant at the $p<0.001$ level. Fewer of the China-based authors than of CIBER's sample had ever published in an OA journal: 15.7% of our China sample claimed to have done so versus 25.7% of the international group surveyed by the CIBER team, the difference being significant at the $p<0.001$ level.

Fewer of them have self-archived their papers or put them on to an institutional repository: 17.5% of our China-based scholars and 14% of our UK-based ones had, versus 32% of CIBER's group. The difference between our two groups on this issue was not significant, but the differences between each of them and the CIBER group were significant ($p <0.001$ for the China group and $p<0.0001$ for the overseas group). The important advantages of self-archiving were seen by our respondents to be wider communication of results (38% said this was 'very important'), speed of dissemination (46% 'very important'), and increased impact (41% 'very important').

Responses to questions measuring attitudes of the respondents towards a possible OA-oriented future scholarly-communication system showed that the Chinese scholars were generally more positive in attitude towards OA journals than were CIBER's international sample, with younger Chinese respondents more optimistic about the likely effects of OA than older ones. However, our respondents differed from CIBER's on a number of points. Only 20.3% of our China group, but 78% of CIBER's international group, thought that printed scholarly journals would disappear altogether. This perhaps reflects the lesser progress towards electronic publication that has been made in China so far. Perhaps connected is another difference: 27% of our China respondents but 55% of CIBER's thought that rejection rates would fall. (High rejection rates, in some disciplines at least, can reflect unaffordable printing costs, and purely electronic journals do not suffer from cost constraints in the same way as printed ones.) It may be, though, that he apparently large differences between the international CIBER group and out respondents may in part be explained by the much lower average age of our group, given the observed lower level of knowledge of OA among the younger age groups.

When asked whether scholarly publishing in China should become wholly OA, fewer than 30% of scholars in China agreed but almost 60% of the UK-based Chinese respondents agreed, and this difference was significant at the $p<0.001$ level. The reason for this difference is not clear, given that most of our UK-based group had started their research careers in China, but it may be that those still based in China are aware of the potential financial difficulty in maintaining an OA publishing operation, while those who have moved to the West are aware of the high subscription prices of western journals and grgard the cost levels in China as sufficiently modest to make OA a feasible business model. Over 70% of both groups thought that at least a partial conversion to OA should occur in China. It was notable that those who publish frequently were less likely to favour an all-OA future (18% agreeing) than those who publish less (30% agreeing), a significant result at the $p<0.05$ level; however, there was no relationship here with the respondents' age. It may be that frequently publishing authors are more senior in the research profession, ads as such are generally more aware of the cost structures of scholalrly publication. Unsurprisingly, those who claimed to know nothing about OA were likely to make a neutral response towards a possible all-OA future, neither agreeing nor disagreeing with the proposal.

Further questions investigated financial issues. We first asked how the respondents' research had been funded. Fewer of the Chinese authors than of CIBER's international group had external grant or contract funding for their work: 40% of the international group said that the research underpinning all of their articles was funded, whereas under 30% of the China authors could say that, while only 16% of international authos said that none of their work was funded whereas over 20% of the Chinese group said this. Again, though, this may in part reflect the different age profiles of the two groups as well as their nationality.

Even so, 85% of the China-based group had paid page charges to a Chinese-language journal and 35% of the UK-based Chinese group had done so. This hints at the possibility that in some cases they had paid page charges out of their own pockets. In contrast, only 6% of the overseas group had ever paid page charges to an English-language journal, versus 38% of CIBER's international group who had paid for publication in a western-language journal at some time. These results seemed to indicate willingness on the part of Chinese authors to contemplate paying for publication in journals in their native language, whereas the major international journals published in English were perhaps perceived to be commercial successes and not in need of this financial support. Those authors who claimed to know a lot about OA were more likely to have paid page charges (92%) than those who knew nothing about OA (78%), a result significant at the $p<0.025$ level, and in agreement with the CIBER group's findings.

One important point was the amount of these payments; the median amount paid per article by Chinese authors was about 600 yuan, equivalent to about 40 pounds sterling, 60€ or US$75, a much smaller figure than is charged by western OA journals currently. The median amount paid by the overseas Chinese authors was lower, but it may simply be that their payments occurred longer ago, before they left for the West. Those who had published more frequently tended to have paid larger amounts than those who published less, perhaps explained by the moire prolific authors having more research funding. As might be expected, there was a loose relationship between the amount people had paid in the past for page charges and the amount they said they might be willing to pay to OA journals in the future; the median amount they were willing to pay was well under 500 yuan. However, there was a big difference between the China-based and the UK-based group here; only 15% of the China-based group were totally unwilling to pay anything for publication in Chinese journals, whereas 70% of the UK-based group were, a result significant at the $p<0.0001$ level. It is difficult to account for this large difference; perhaps those who have moved to the West have become used to the more commercially-based scholarly publishing system of North America and Western Europe, and do not see why they should pay for publication. Nor is it easy to square these views with the fairly positive attitude of these respondents towards OA. Willingness to pay for publication in English-language (or other non-Chinese) journals was marginally lower for the China-based group but marginally higher for the UK-based group.

When asked who should pay for the costs of publishing scholarly journals, 65% of respondents said that the scholar's department or faculty, or the research funder, should cover all or most of these costs. Those in biomedicine were the most likely to take this view, and those in engineering the least likely, though it was the view of over half the respondents in all disciplines. Over 80% said that neither authors nor readers should have to pay out of their personal pockets. Over 30% thought that all or most of the costs should be covered by central giovernment – perhaps not a surprising view in China – but around 25% felt that commercial sponsors should make a big contribution.

**Interviews**

Many of the UK-based Chinese scholars in the survey were also interviewed in the Chinese language by R.L., mostly face-to-face but in some cases by telephone. The full results of these interviews are not published here, but one important finding from them was that, having moved to the UK, they had ceased to make much use of the Chinese-language literature either as authors or as readers. They had adopted the publishing and reading habits of their Western-born colleagues, and concentrated on the major journals published in Western Europe or North America. One reason given by them for this was the fact that they regarded the mainstream Western journals as better than the Chinese ones, a result which accords with observations made in countries as varied as New Zealand and Malaysia [6, 7]. But it is also true that they said that they could not access the Chinese-published journals from their UK universities, since their university libraries did not subscribe to the Chinese journals or to the Chinese aggregation services that link to them, such as Wan Fang [8], China Academic Journals [9] and Chinese Scientific Journals Full-Text Database [10].

**Discussion**

Chinese science is known to be of high quality, with many scholars born and educated in China now working in major universities in the West. Their early work, and that of others who have stayed in China, is largely reported in the Chinese-language literature published within China. This literature is little-used by scholars in other countries, largely because of the language barrier, but evidence from this survey seems to suggest that it is difficult to gain access to these journals from the West, ebeven when they are in principle accessible in electronic form. Their subscription prices are low by Western standards, but Western universities do not in general subscribe to them so they are inaccessible to research workers in the West, whether Chinese-speaking or not.

Overall, the survey seemed to suggest that knowledge of the OA principle was incomplete among these scholars, even though almost one-third of the China-based group claimed to know quite a lot about OA.

This low level of understanding was reflected in some contradictory results: for example, the overseas Chinese group was more favourable to the idea of an all-OA future for Chinese journals than the China-based group, but less willing than the China-based group to pay publication charges! While few of them supported OA for English-language journals, the respondents were generally sympathetic to the idea that journals based in China might turn to OA, provided that individuals did not have to pay publication charges out of their own pocket. This perhaps reflected a degree of realism about the financial prospects for scholarly journals published in languages other than English. Given the relatively low costs of publishing in China, reflected in the fairly low page charges that some authors had paid in the past, it might indeed be possible for Chinese journals to be published electronically at the expense of research funders and authors' institutions. Chinese publishers – many of the universities, such as the member institutions of SCUJNS – do not publish their journals on a commercial basis or seek to make large surpluses from them, but they do have costs, modest though they may be, to cover.

## 3 Planned Website for Access to Chinese Electronic Journals from the West

**Introduction**

Interviews with the overseas Chinese scholars in the survey showed that they made very little use of the Chinese-language literature published in China, either as authors or as readers, even if they had used it when previously resident in China. It appeared that this lack of use was in part a consequence of the Chinese journals being difficult to access from the West, even though many of them are available in electronic versions. There are secondary databases based in China [8-10], but these are available on a subscription basis and few Western university libraries subscribe to them. Thus the full texts of the journals are inaccessible from outside China, even though many of them are in principle available free of charge. In addition to encouraging use of this literature by overseas Chinese scholars, it is desirable to make it accessible to others in the West who cannot read Chinese. As editors of some of these journals, R.L. and J.Z. were also aware that they are little used by non-Chinese speakers, owing to a general lack of English-language web pages to access them, even though individual papers often have short English abstracts. They would prefer that their journals were better-known, and better-used in the West and seek to provide tools to access them better.

**Proposal**

This project therefore constituted the pilot phase of a proposed operational website that would provide straightforward access in the English language to Chinese electronic journals, especially those in membership of the Society of China University Journals in the Natural Sciences (SCUJNS), the organisation that might operate the website in the longer term. R.L. and J.Z. are both active members of SCUJNS in their capacity as editors of university journals at Chongqing Normal University and Tsinghua University respectively. It was hoped that on their return to China in late 2006 they would be able to obtain funding for the development of the pilot website into an operational service. If such a service is provided, then it would be expected that visibility and impact of Chinese research work would be greatly improved outside China.

The full texts of papers in Chinese are held on the servers of their publishing organisations (mostly universities) in China. At present about one-third of the 700+ journals published by SCUJNS members are available in electronic form, but this proportion is expected to increase rapidly. The concept is that an English-language website will be created that will provide ready access to these journals, and this website in turn could be linked into sites such as the ALPSP Learned Journal Collection [11] that host many journals from not-for-profit organisations in the west.

The project was named EJUNIC (**E**lectronic **J**ournals of **Uni**versities in **C**hina). The main aim of the pilot EJUNIC website design was to create a web interface for publishers to register their journals, and to facilitate overseas readers' access to these academic resources under an Open Access mechanism. The detailed objectives include:

- To create English and Chinese language versions of the website.
- To design a registration system for publishers to mount links to their journals.
- To design a login system for publishers to keep their journal updated.
- To display all the included journals in specific pages that provide title, link, introduction, and contact information.
- To provide a browsing function with an A-Z index of the journals included
- To provide both advanced and simple search functions that allow users to search journals by title, author, keywords, and abstract.
- To establish a harvesting program that automatically collects available articles from the included journals
- To provide long (informative) English-language abstracts of the papers in the journals, and English metadata

Technical aspects of the pilot phase were implemented by Z.X., who was a postgraduate student of Electronic Publishing at the Department of Information Science at Loughborough University at the time. Functions that will be provided in the full implementation are:
*Register function* – allows publishers to mount links to their journals included in EJUNIC.
*Browser function* – allows readers to browse all the included journals in our database. An A-Z index is provided.
*Search function* – allows readers to search a particular term (title of journal or ISSN) to locate the needed material. In an operational version, we also expect to implement further function that will provide readers a powerful text level search engine to locate items within the included journals by more choices of search terms, such as title of article, author, abstract, etc. The function is expected to adopt a harvesting program based on an Open Access standard.

The registration process was fully implemented in the pilot phase, and falls into several stages:
*ISSN verification*: EJUNIC assigns the journal's ISSN as a unique username. A database was designed to hold ISSNs and the titles of their corresponding journals. Publishers will be asked to provide their ISSN to make sure they are suggesting a valid journal.

*Submitting basic information*: Once the ISSN is verified, the publisher then moves to the stage of basic information input. The required information in this stage includes a valid URL, contact person, e-mail address, and telephone number. A database (basic_info) holds this information.

*Journal verification*: EJUNIC is a website based on an Open Access (OA) protocol. It requires that all the journals included are free to access. Therefore, a verification procedure is carried out in order to confirm that the journal is OA.. If the journal is proved to be OA, a password will be sent to its publisher, which ends the whole registration process. Otherwise, we will send an email to inform the publisher of possible reasons of failure.

**Technical details**

An English-language home page was designed, and links to a number of journals, mostly those for which J.Z. is responsible at Tsinghua University, were implemented. PHP technology was used to bridge the website and a database created by MySQL. PHP is a human-readable language which is easy to write, edit, understand and expand. It is currently recognised as one of the most popular languages that used in network programming. On the other hand, MySQL is widely used in small sized databases, as it provides good flexibility in terms of database management.The designing environment was simulated by an application called APM Express 5.0

(APMEX). APMEX is a software package that associates PHP, MySQL and a database management tool, PhpSQLAdmin. It significantly eases the complex process of PHP and MySQL configuration. The coding process was completed by Macromedia Dreamweaver 8. Cascading Style Sheet technology was adapted to improve the appearance of the website.

**Future activities**

A further activity proposed for the operational phase is the provision of informative English-language abstracts of the papers in the participating journals. We recognise that this will entail negotiation with the various publishers, and locating people in China with good English-language skills to provide the abstracts.

Initially the journals incouded will be those published by SCUJNS members, but in a leter phase it is hoped that other Chinese journals published by not-for-profit organisations in China will also be brought into the ambit Of EJUNIC.

## 4 Conclusions

Earlier work, such as that of the CIBER group [5], has shown that despite the large amount of debate that takes place today about Open Access, scholars in general are still not well-informed about the OA concept. This work shows that Chinese scholars are, if anything, even less well-informed that their Western counterparts, and even scholars from China who have moved to the West permanently to work are significantly less likely than CIBER's respondents to have published in an OA journal. Despite their relative ignorance of this topic, both the Chinese groups of respondents seemed moderately favourable towards OA for Chinese-language journals, and this perhaps reflects a realism about the modest commercial prospects for these titles compared with English-language journals published by major for-profit or not-for-profit organisations in Western Europe of North America. Certainly, those resident in China seemed willing to pay author charges, and many had done so, even in some cases out of their personal pockets. The advantages of OA that they detected were similar to those mentioned by other groups – wider communication of their work and consequent higher visibility and impact for it. They also mentioned faster publication, which might be seen as an advantage of electronic publication per se, rather than OA. As the general cost level of publishing is lower in China, and many jouirnals are already published by universities directly, it may be easier to progress to an OA publishing model ('the Gold Route to OA') in China than in Western countries. In contrast, fewer of our respondents – in both the China-based group and the UK-based group – than of CIBER's group had posted copies of their articles on institutional repositories, and it may bet hat he 'Green Route to OA' ha made less progress in China than in the West.

Although much has been done to make Chinese research better known in the West [1, 2, 4, 12], and indeed major Western information providers, such as Swets with their 'Gateway to China' service [13], and NetLibrary working in partnership with a Taiwan company [14], are now providing information services, it is clear that scholars working in the West are largely not using the Chinese literature, or publishing in it, even when they themselves are originally from China and can read the Chinese language. It seems that the fact that these services are commercial and charge subscription fees, even where the original journals may be free to access electronically, leads to their being available in the West only to the largest and best-resoursed institutions. We therefore proposed that a website be provided through SCUJNS to provide direct, easy and free access to university-published Chinese journals from outside China. This would be an English-language website but would link to the full texts in Chinese held on publishers' own servers, and these would be enhanced by long, informative English-language abstracts. Both the website with its metadata, and the full texts, would be available free of charge. We produced a prototype of this website which was prerpared quite quickly using readily available open-source software, and which functioned satisfactorily. It is hoped that it might be developed into a full operational version, with other journals from Chinese publishers other than universities being added to it in a later phase, and that it might be linked to the ALPSP Learned Journals Coillection [11] to further enhance its visibility.

## References

[1] STANLEY, A.; YAN S (2007) China Opening Up: Chinese University Journals and Research – Today and Tomorrow, *Learned Publishing* **20**(1), 43-50.

[2] JIA, X. (2006) The Past, Present and Future of Scientific and Technical Journals of China, *Learned Publishing*, **19**(2), 133-141.

[3] WANG, S.; WELDON, P. R. (2006) Chinese Academic Journals: Quality, Issues and Solutions, *Learned Publishing*, **19**(2), 97-106.

[4] ZHANG, L.; YAO, Y.; ZHANG, F. ; DU, WENTAO (2006) The First Comprehensive Chinese University Journal Published in English, - the Tsing Hua Journal, *Learned Publishing* **19**(3), 204-208.

[5] ROWLANDS, I.; NICHOLAS, D.; HUNTINGDON, P. (2004) Researchers' Attitudes towards New Journal Publishing Models, *Learned Publishing*, **17**(4), 261-274.

[6] ROWLAND, F. (2005) Scholarly Publishing in New Zealand, *Learned Publishing,* **18**(4), 300-310.

[7] ZAKARIA, J.; ROWLAND, F. (2006) What are the Prospects for Online Scholalrly Publishing in Malaysia? The Cultural Constraint' In Proceedings of the ELPUB 2006 Conference, Bansko, Bulgaria, pp. 229-236

[8] Wan Fang Data, English version at http://www.wanfangdata.com (accessed 12 March 2006)

[9] China Academic Journals, English version at http://www.thtf.com.cn/www/web/en/index.asp (accessed 12 March 2006)

[10] Chinese Scientific Journals Full-Text Database, English version at http://dx3.cqvip.com/en/index.htm (accessed 12 March 2006)

[11] ALPSP Learned Journal Collection, in partnership with Swets, http://www.alpsp-collection.org/ (accessed 22 December 2006)

[12] WANG, J. (2006) Major Chinese Full-Text Electronic Information Resources for Researchers and Scholars, Serials Review, 32(3), 164-171.

[13] Swets Gateway to China, http://www.swets.com/web/show/id=84103/langid=42 (accessed 11 April 2007)

[14] NetLibrary, Chinese Language e-Resources, http://library.netlibrary.com/ChineseLanguage.aspx (accessed 11 April 2007)

# Managing Expectations for Open Access in Greece: Perceptions from the Publishers and Academic Libraries

*Banou G. Christina* [1]*; Kostagiolas A. Petros* [2]

[1] Department of Archive & Library Science, Ionian University
Plateia Eleftherias, Corfu 49100, Greece
e-mail: cbanou@ionio.gr
[2] Department of Archive & Library Science, Ionian University
Plateia Eleftherias, Corfu 49100, Greece
e-mail: pkostagiolas@ionio.gr

## Abstract

In Greece, there seems to be a growing level of awareness regarding open access among scholars, faculty staff and information professionals. Indeed, consensus regarding the necessity of open access initiatives in Greece is gradually established. The present of open access in other European settings may however be revealing the expected, though distinct, future of open access in Greece. This work focuses upon some current aspects for open access and attempts to investigate them for the Greek setting. The investigation includes five (5) important aspects of open access, i.e. a) ETDs management from the academic libraries, b) university repositories development, c) regulation of digital and/or printed scientific material quality requirements, d) cooperation and competition between libraries and academic publishers, e) understanding the role of scientific work dissemination in developing future professionals and scholars. The paper initially provides an outline for the Greek publishing industry, focusing on STM publishers and on the way they take advantage of the changes mainly in editorial and marketing terms, in a hybrid technological era. The Greek publishing industry may be representative of other national small publishing markets. Further, an empirical research is providing in order to illuminate open access from two different points of view: that of STM publishers and that of academic libraries' directors in Greece. The empirical investigation took place in February and March of 2007 and is based on seventeen experts' perceptions. The methods employed are outlined and include the development of the questionnaire for semi-structured interviews. Finally, the unexpected agreement from both publishers and academic libraries' directors regarding open access development is discussed and some specific for Greece conclusions are drawn.

**Keywords:** open access; publishing industry; academic libraries management; scholarly communication; Greece

## 1 Introduction: Setting the Scene

In Greece, there seems to be a growing level of awareness regarding open access among scholars, faculty staff and information professionals. Indeed, consensus regarding the necessity of open access initiatives in Greece is gradually established. Academic libraries, and for that matter university authorities in Greece, realize nowadays that cannot purchase access to all the scientific information their researchers expect, although some association agreements and research programs, involving publishers, the National Documentation Centre and other institutions, have assisted [1]. Publishers in Greece have been considering the development of distinct prising models for making available books, monographs and scientific articles, so they cause further pressure on institutional budgets. Overall, the current scholarly communication model, that the academia employs, seems to currently disconfirm expectations of scholars and of the Greek scholar community as a whole. A novel information and research strategy for academic libraries is required involving scholar publications which are digital, online, free of charge, and free of most copyright and licensing, compatible with printed edition.

The present of open access in other European settings may be revealing the expected, though distinct, future in Greece. Open access for Greece may constitute a greater challenge due to the language barriers, which may form two (2) distinct types of scientific publication: the ones written in Greek and the ones that are not. The expectations of the scholar community relate to the Electronic Thesis and Dissertations (ETDs) management, the development of university repositories and, finally, the regulation of the digital versus printed material quality requirements. The academic libraries in Greece ought to enhance their role within the current scholarly communication setting. On the other hand, it may safely be assumed that the scholarly community in Greece has nothing or little to gain from any publishing pricing model. Scholars mostly wish to publish their work in high impact journals (or as monographs) either open access or not, realizing gradually that openness has a positive effect on impact factors based on citations and/or other traditional and frequently used measures of research impact.

This work focuses upon some current aspects for open access and attempts to investigate them for the Greek setting. The investigation includes A. an outline of the Greek publishing industry, focusing on STM publishers and on the way they take advantage of the changing environment mainly in editorial and marketing terms, in a hybrid era. In that framework, novel publishing strategies and policies are developed. From that point of view, the Greek publishing industry may be representative of other national small publishing markets. B. the expectations of the directors of the academic libraries about open access. It is interesting and fascinating to illuminate open access from two different points of view: that of STM publishers and that of academic libraries' directors. C. an empirical investigation based on seventeen experts, that took place in February and March 2007 through a semi-structured questionnaire, in order to portray the specific aspects in Greece. The methods employed are outlined. D. the unexpected agreement from both publishers and academic libraries' directors regarding open access development is discussed and some specific for Greece conclusions are drawn.

## 1.1 The Present Scenario in the Greek Publishing Industry

The publishing industry in Greece may be characterized from the absence of conglomerates and of large foreign publishing houses. Furthermore, it is rather traditional (family owned and managed enterprises) in comparison to international markets. Specifically, the Greek Scientific-Technical-Medical (STM) publishing production [3] represents about one third (35,1%) of the annual book production. There is a steady increase, during the last five years, in the annual production of new scientific titles, that may express a turning point of the Greek STM publishing industry. In 2004, 2692 new scientific titles were published (out of 7.888 new titles of the total annual book production), while, in 1999, 2410 were the new ones [4]. It is significant that small and medium-sized STM, on the one hand, and general publishing houses, that also produce scientific publications on the other, manage to develop the profile of the Greek publishing industry; at the same time, they play a central role in scholarly communication in Greece, collaborating with the academic community.

In regard to the Greek publishing industry, focusing on the STM publishing, a number of specific features can be synopsized as follows [5]. The Greek publishing market has a rather small audience of about 14 million people, due to the Greek language, which is unique among the European languages. Hence, that market has not yet been in the focus of international conglomerates or large publishing groups; on the other hand, the Greek publishing industry is deeply influenced by them in certain terms, such as in patterns of promotion and of management practices. One of the main features of the Greek publishing industry is that almost all the Greek publishing houses are companies, owned and run by members of a family, who continue and try to innovate, respecting the tradition. Concerning the STM publishers, it is characteristic that many of the publishing houses' names consist of the surname of the founder, who, in some cases, still runs the company: Sakkoulas, Papasotiriou, Siokis, Paschalidis, Ziti, etc.

The remarkable increase, during the last fifteen years, in the total annual production of new titles (from less than 3,000 in the beginning of the 90ies to 7,888 titles in 2004) reveals the prosperity and the turning point of the Greek publishing industry [6]. More specifically, concerning the STM publications, there is a steady increase, as it was referred above. Large publishing houses in Greece produce more than 80 titles per year, medium produce 10-80 titles, while the small ones publish less than ten (10) titles annually [7]. Only seventeen are large publishers; five of them are STM. The majority of STM publishers are medium. Generally, Greek STM publishers are competitive to academic presses and to organizations with publishing activities such as scientific institutions, museums, chambers and others.

Concerning the scientific publications, the publishing houses can be categorized as follows: a. strictly STM publishers, b. general publishers, which include in their catalogues scientific texts, c. organizations and institutions that publish or order and encourage publications. The well known and specialised in scholar work publishing houses are the market leaders. They can, through their policy, influence the structure of the book market in Greece. It is significant that the majority of the new titles published annually are works of Greek academics and scientists. Out of 2692 new titles published in 2004, only 770 were translations, something that demonstrates that the Greek STM market develops an interest and a taste in the national scientific production, for which there is need to be promoted. The academic community determines to a great extent, by its special needs and expectations and through collaborations, the STM production. Furthermore, the academic community is an important knowledge producer and co-operates with the publishing houses, not only by its works, but also by editing and being responsible of series. With its high expectations and with a very good judgement, this particular audience is usually a force for change and for innovation. STM publishers are intended to a specialized and, therefore, specific, rather homogenous and steady reading audience. This target group is easily accessible, through economic ways of promotion. On the other hand, general publishing houses promote and advertise, sometimes even the scientific titles, in such a way so as to attract the majority of readers.

In the last five years, there are "new" needs in the traditional and rapidly changing Greek publishing industry. The profile of the Greek publisher was until recently formed in terms of a family company for a small market, something that gradually is changing; although family enterprises, the STM publishing houses are conscious of the competition, of the need for innovation and of the new role that they are called to play. New policies and strategies, competitive values, and new information technologies demonstrate the need of special studies and life long learning in the STM Greek publishing industry. In an era, in which information is the main product, the challenges for the STM publishers are great, and they should be innovative always bearing in mind that "publishing companies are content-acquiring and risk-taking organizations oriented towards the production of a particular kind of cultural commodity" [8].

### 1.2 Selected Research Issues for Open Access in Greece: Competition and Co-Operation

This work provides some empirical results which are based on the perceptions of the Greek STM publishers and directors of academic libraries for five (5) important aspects of open access, i.e. a) ETDs management from the academic libraries, b) university repositories development, c) regulation of digital and/or printed scientific material quality requirements, d) cooperation and competition between libraries and academic publishers, e) understanding the role of scientific work dissemination in developing future professionals and scholars. However, the Greek publishing market and academia encompass some unique features, including the language and other exclusive features. These may further establish an additional set of research objectives that are referred here as the "language" and the "digital product" issues.

The STM publishers in Greece are focusing on the conventional printed material production, and hence they do not consider openness as a "real" threat until now. Furthermore, academic libraries have recently invested a significant amount of money and effort in personnel development and information technology, have initiated additional university policies and/or mandates regarding the Electronic Thesis and Dissertations (ETDs) management, the development of university repositories and finally the regulation of the digital versus printed material quality requirements. The academic libraries in Greece are "better" organizations than Greek publishing enterprises in "gathering" scholarly work and have better access to digital distribution channels [9], as they have taken advantage of the new digital technologies [10]. The amount of accessible digital information is increasing due to the advent of information technologies, the Internet and other international networks. However, the diversity in the content of the material and in the languages text are written in, is also significantly increasing. The Greek language forms a barrier that has to be crossed. Large international publishing enterprises are intensive on digital STM publications produced mostly in other languages than the Greek language, while national publishers are resolving quality issues for Greek scientific work and thus they are developing highly prestigious conventional STM products based on the scientific work.

A realization that arose with this research is that "open access" is redefining information "transaction" for the Greek scientific and technical STM market. However, if "transactions" are costless, the most important issue is that the rights of the various parties involved should be well defined and the results of legal actions easy to forecast [11]. It would therefore seem desirable to further focus on the stakeholder's views, taking their perceptions into account when making economic decisions and/or investigating market characteristics. In that respect, "openness" is less of a substitute product within the competitive forces in the publishing industry in Greece, and it may be rather be treated as a factor that characterizes the nature of competition (and co-operation) within this particular industry. Greek publishers may co-operate and compete with the academic libraries, in regard to the language of the text and/or the nature (printed and digital) of the produced STM material.

## 2 The Empirical Research Conducted

The objective of the empirical research is to identify and then investigate the perceptions of Greek STM publishers and the perceptions of the directors of academic libraries, through five (5) important aspects of open access: a) ETDs management from the academic libraries, b) university repositories development, c) regulation of digital and/or printed scientific material quality requirements, d) cooperation and competition between libraries and academic publishers, e) understanding the role of scientific work dissemination in developing future professionals and scholars. In the following paragraphs the methods employed as well as the results of the empirical study are presented.

## 2.1 Methods Employed

The empirical research was based on semi-structured interviews that were directed to large publishers and directors of central academic libraries in Greece. Therefore, the interviewees were organization representatives (experts) selected on the following criteria: a) have an in depth experience managing STM material, b) academic libraries with central administration as well as publishers with more than 30 tittles annual production were selected. The academic publishing houses are not included in the research, mainly, due to their small annual production (less than 30 titles). On the other hand, within the group of STM experts a number of participants represent publishing organizations as the Technical Chamber of Greece.

A questionnaire was designed for the survey and a pre-test was conducted. The questionnaire includes both closed and open-ended questions. For the closed-ended questions a five-point Likert scale was used, ranging from 1=”strongly agree” up to 5=”strongly disagree”, in order to determine the extent of agreement and/or disagreement of the participants to specific statements regarding open access in Greece. The questioner included seventeen (17) questions as follows:

- Research Questions 1 to 12: aimed at investigating characteristics of the experts participated in the survey (organization title, address, telephone, email, full name, position in the organization, education, years of employment) and the organization they represent (number of employees, nature of services and/or products –conventional, digital, hybrid–, presence of electronic thesis and dissertation system, presence of repository);

- Research Questions 13 and 14: aimed at assessing the degree of agreement of the participants to the following statements, “the information provided to the scientific community in Greece is sufficient” & “clear need for the open access development in Greece”;

- Research Question 15 for “ETDs management from the academic libraries”: aimed at assessing the degree of agreement of the participants to the following statements, “support other activities”, “produce economic value”, “support scientific work”, “support co-operation” & “is a reason for competition between academic libraries and publishers”;

- Research Question 16 for “university repository development”: aimed at assessing the degree of agreement of the participants to the following statements, “support other activities”, “produce economic value”, “support scientific work”, “support co-operation” & “is a reason for competition between academic libraries and publishers”;

- Research Question 17 for “law and regulation of digital and/or printed scientific material”: aimed at assessing the degree of agreement of the participants to the following statements, “sufficient for ETDs”, “sufficient for repositories”, “sufficient for copyright issues”, “support co-operation” & “is a reason for competition between academic libraries and publishers”.

The research questions 13 through 17 were accompanied with open questions, so that the participants could state in free narrative form their additional comments. The survey was not aiming in providing results that could be generalized, although, in the lines of Behrakis [12], the information recorded and the qualitative analysis provided, produce indicative results based on expert’s opinion. Furthermore, it was found that the distance in the scale between 1=”strongly agree” and 2=”agree” was small, as well as, that the distance between 4=”disagree” and 5=”strongly disagree”. Hence, the initial form of the scale was reduced from five to three [agreement (+), rather (=), disagreement (-)]. The percentages were computed and graphs were produced, while for the open-ended questions content analysis was employed in order to determine frequency of statements of interest [13].

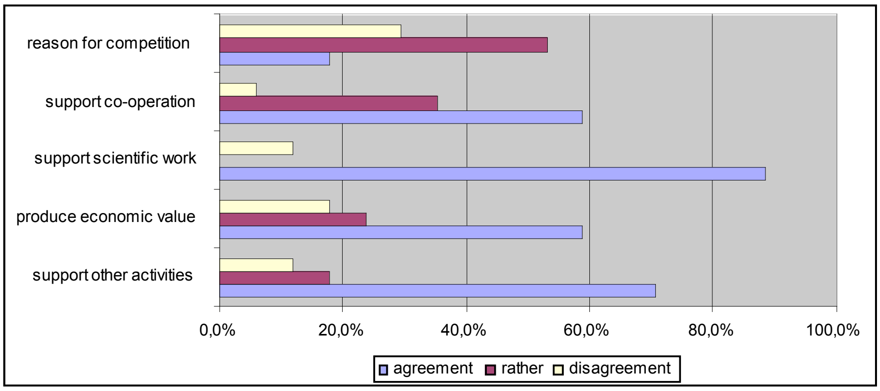


**Figure 1: Experts points of view regarding the "ETD's management" aspect of Open Access in Greece**

## 2.2 Results of the Empirical Research

The empirical research was conducted through structured interviews based on a specially designed semi-structured questionnaire, within February and March of 2007 (a pilot study took place in January 2007). Among the twenty two (22) academic libraries and the twenty (20) large Greek STM publishers conducted, representative of ten (10) academic libraries and seven (7) STM publishers agreed to participate in this survey (i.e. the 45.5% of the libraries conducted and 35.0% of the STM publishers). The overall profile of the group of experts was considered sufficient for our purposes in consideration of the following, a) the group of experts is representative including participants from the academia and the STM publishing industry, b) the group of experts consist of highly skilled and educated information professionals (the participants are all university graduates mainly from Library Science and/or Information Science departments, six of them hold a postgraduate diploma, while four of them hold a Ph.D.), c) sufficient working experience (fifteen out of the seventeen were employed in the organizations they represented for more than five years), and d) specific organizational features (six organizations employed more than twenty five professionals, while thirteen organizations provide services of both conventional and digital form). The results for each of the close–ended questions are provided bellow (Table 1 and Figures 1 to 3), while the analysis of the open–ended questions follows.

| Research Question | Issues addressed in the survey for Open Access in Greece | Agreement | Rather | Disagreement |
|---|---|---|---|---|
| 13 | "the information provided to the scientific community in Greece is sufficient" | 53.0% | 23.5% | 23.5% |
| 14 | "there is a clear need for the open access in Greece" | 88.2% | 0.0% | 11.8% |
| 15 | "ETDs management from the academic libraries" | | | |
| | "support other activities" | 70.6% | 17.6% | 11.8% |
| | "produce economic value" | 58.8% | 23.5% | 17.6% |
| | "support scientific work" | 88.2% | 0.0% | 11.8% |
| | "support co-operation" | 58.8% | 35.3% | 5.9% |
| | "is a reason for competition between academic libraries and publishers" | 17.6% | 52.9% | 29.4% |
| 16 | "university repository development" | | | |
| | "support other activities" | 64.7% | 11.8% | 23.5% |
| | "produce economic value" | 52.9% | 23.5% | 23.5% |
| | "support scientific work" | 82.4% | 0.0% | 17.6% |
| | "support co-operation" | 64.7% | 11.8% | 23.5% |
| | "is a reason for competition between academic libraries and publishers" | 17.6% | 47.1% | 35.3% |
| 17 | "law and regulation of digital and/or printed scientific material"* | | | |
| | "sufficient for ETDs" | 17.6% | 11.8% | 41.2% |
| | "sufficient for repositories" | 11.8% | 11.8% | 41.2% |
| | "sufficient for copyright issues" | 17.6% | 11.8% | 35.3% |
| | "support co-operation" | 11.8% | 17.6% | 35.3% |
| | "is a reason for competition between academic libraries and publishers" | 11.8% | 11.8% | 41.2% |

**Table 1: Results of the survey for aspect in open access in Greece (*6 of the responders could produce a reliable judgment for the statements)**

In Table 1, the overall results of the survey conducted are exhibited. In the first column of Table 1, the statements under investigation are provided; while the columns that follow provide the percentages reflecting the perceptions of the participants (STM publishers and directors of the academic libraries). More than half of the experts (53%) state that the information provided to the scientific community in Greece is adequate (23.5% "rather" adequate and another 23.5% "disagree"), whereas the majority of the participants (88.0%) agree that open access initiatives are indeed required (Table 1, research question 14). The majority of the experts agree that the ETDs management from the academic libraries may "support other activities" and "support scientific work", while they do not think that would be a reason for competition increase between publishers and academic libraries (Figure 1).

In Table 1 (research question 16 "university repository development") and in Figure 2, the results indicate a significant agreement among the participants in the survey, stating that in Greece the development of university repositories may "support co-operation" (64.7%), "support the scientific work" (82.4%) and "support other activities" (64.7%) within the university communities. Once again the experts did not indicate that this is a reason for increasing the competition between libraries and Greek publishers. In Table 1 (research question 17) and in Figure 3, the experts point of view is presented, regarding "law and regulation of digital and/or printed scientific material" aspect of open Access in Greece. For the research question 17, six of the participants did no express any opinion within the survey, stating that they need further information. However, the majority of the participants in the survey stated that "law and regulations" in the present form in Greece do not support "scientific work" and "co-operation", and they are not sufficient for the development of "university repositories", and "ETD's management" within the Greek universities.

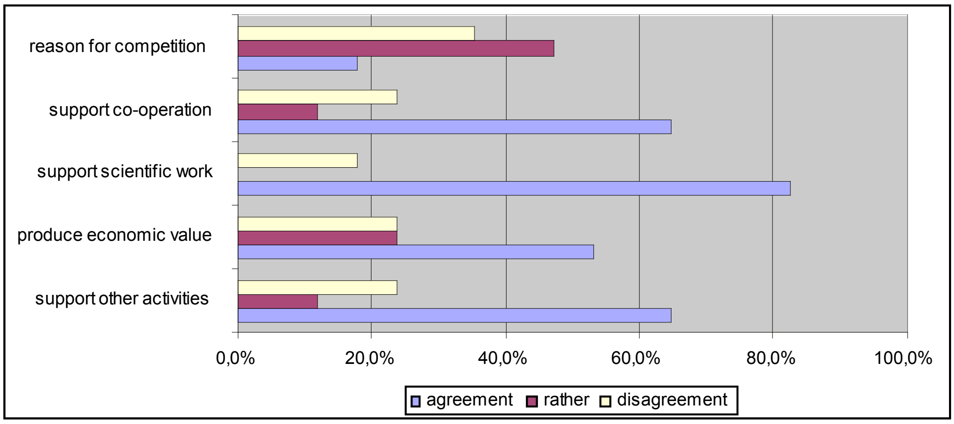


**Figure 2: Experts points of view regarding the "repository development" aspect of Open Access in Greece**

The experts expressed their views, for the distinct aspects of open access examined in the survey, in a free narrative manner through a set of five (5) open-ended questions attended each research issue (from 13 to 17). The participants frequently commented on the following: **a.** improvements have been achieved in Greece in terms of quality and quantity of the scientific information services over the last 6 to 8 years, **b.** open access in a cost – benefit analysis framework seems to prevail, reducing management cost (although significant investments ought to be made for the management of openness) and gradually reduce the "need" for costly agreements with international publishers, **c.** open access development may support improvements in Greek scholar production in both the Greek and other European languages, through better information provision and "free of subscription charges" high quality scientific communication, **d.** apprehension and support within a centrally regulated legal and investment framework for open access in Greece is required, while Greek scholars should support openness within the university communities. The participants representing academic libraries in the survey stated that open access may be used as a vehicle for further improvements (e.g. in grey bibliography management, technical reports distribution, maintenance cost reduction etc.) and of course professional development. Furthermore, the library directors stated that education and empowerment of the library staff may be a key factor for making openness a reality and that the aspects of open access studied here, may support co-operative initiatives within the academic community. It is worth mentioning that both Greek academic library directors and publishers which participated in this survey, support open access development, and although sceptical, mainly the publishers, they state that with proper regulation, e.g. embargo on the time of scholar material provision, public and private sectors can find common grounds for co-operation.

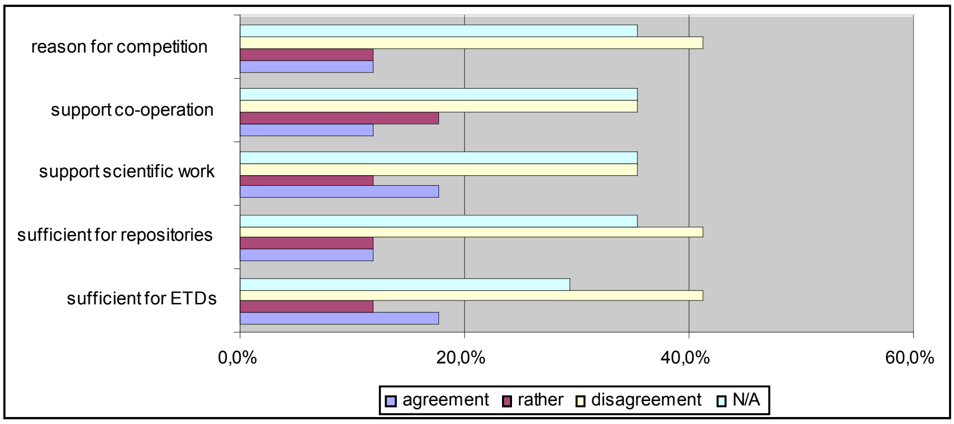


**Figure 3: Experts points of view regarding the "law and regulation" aspect of Open Access in Greece**

# 3 Discussion

The directors of the academic libraries pointed out that an important factor for ETDs management and repository development in the universities are the education and the empowerment of the library staff. Furthermore, they thought that centralized university initiatives may provide a way forward regarding the academic quality requirements for either printed or electronic scientific publications. Repositories, in particular may provide articles, pre-prints, post-prints, dissertations, PhDs, monographs or chapters from monographs, proceedings, rare material, to the scientific community, while ETD management undoubtedly enhance the quality of the scholar work production.

Some academic libraries in Greece have developed or they are developing ETD's and/or repositories: In the University of Macedonia in Thessalonica a repository has been developed and a mandate is regulating thesis and dissertations standards for digital submission. Similar efforts are undergoing in the University of Athens, for example at the Faculty of Law, where an ETDs management initiative is under development for the postgraduate courses of Civil Law. Similarly, in most Greek universities (e.g. University of Piraeus, Ionian University etc.) such initiatives are under consideration. Furthermore, a number of Greek universities have an academic press (e.g. University of Crete, University of Macedonia, etc.), that it is closely related to the academic community and the library. The academic libraries directors finally expressed their belief that the unrestricted reading, downloading, copying, sharing, storing linking and accessing scientific work, which open access incorporate, will lead to more efficient scientists and professionals in the future. However, a number of economical, political, legal and technical aspects ought to be addressed as soon as possible for satisfying user needs [14]. Universities ought to provide to all members access to the information; especially, to the information that is produced inside the university campus.

The international publishers usually resist to Open Access due to a. economic investments, b. issues of co-ompetition (cooperation and competition) with academic libraries c. legal and copyright issues. The STM publishers in Greece publish for a homogenous small market. They are engaged in printed scientific books and conference proceedings, and they collaborate with scholars and universities. Thus, the challenge for STM publishers in Greece is to realize and take advantage of open access rapid development, shaping a novel marketing strategy for the printed and/or the digital publications. It has been indicated that the author would prefer to communicate directly with the reader-user, without the intervention of the publisher. The scholarly community in Greece must become conscious that open access journals have high impact factors and thus, to gain their trust [15]. On the other hand, it has been pointed out, regarding publisher – author relation, that diachronically the publisher has been the one that takes the risk [16]. Innovation has been proved to be highly valued within the publishing industry [17].

# 4 Conclusions

A significant reason for the frustration surrounding openness in Greece is not so much the concept itself nor the economic issues involved, but the way openness is presented by the decision makers in scholarly communication. Open Access in Greece is possible, authorized, and beneficial for all those involved. The results drawn from this work may enlighten the main trends and the current issues for Open Access in Greece. It is clear that, if a rather unexpected co-operation -fruitful for open access development- among STM publishers, the academic libraries and the scholarly community in Greece is established, this should be on the ground of well defined roles and legal arrangements. A very interesting question to be further investigated in the future arises: Is Openness a threat for small publishing markets, such as the Greek STM market? Publishers in Greece "do not perceive openness as a threat" but on the contrary, if the roles are clearly defined and their investments secured, they might build upon openness innovative novel strategies and co-operative policies with the academic libraries in Greece.

Like in every innovation, such as the concept of open access, dilemmas and/or threats arise, because we don't really know what benefits we are getting from that. Hence, we should establish the research needed to find out the technological, political consequences of open access. Furthermore, empirical results can shed light on whether publishers and academic library directors may find in Greece a common ground for improving the quality of scholarly communication in Greece and in Europe.

# Acknowledgements

The authors would like to express their gratitude to all those participated in either the pilot study or the survey presented in this work, i.e. Alivanoglou L. (director of Kleidarithmos publications), Anthi-Kalofolia E. (Ionian University, director of central library), Brindesi H. (Efgenideio foundation, library), Dimakopoulos E. (TEE library), Eleftheroudaki S. (Eleftheroudakis chain-bookstore and publications, publisher), Fragkou A. (University of Macedonia, director of central library), Georgakis K. (A.T.E.I. of Patras, director of central library), Kalambaliki M. (National Technical University of Athens, director of central library), Kamariotakis E. (TEE library, technical publications), Katsirikou A. (University of Piraeus, director of documentation centre), Kokkinos D. (National Technical University of Athens, central library), Moniou D. (Kyriakides publications), Papageorgiou V. (Metaixmio publications, publisher), Pesmatzoglou E. (University library of Attikon Hospital, director of library), Synelli K. (University of Patras, director of central library), Toraki K. (TEE, director), Vardakosta I. (Harokopio University, director of central library), Zachos G. (University of Ioannina, director of central library).

# Notes and References

[1] KOROBILI-XANTINIDOU, S.; MORELELI-CACOURIS, M.; TILIKIDOU, I. "Concepts, reality and suggestions about Greek library management education", *New Library World,* 104 (1189), 2003, pp. 203-217

[2] KOSTAGIOLAS, P. .A "Information services for supporting quality management in Healthcare", *Journal on Information Technology in Healthcare,* 4 (3), 2006, pp. 137-146.

[3] CLARK, G. *Inside Book Publishing,* third edition, London and New York: Routledge, 2006, pp. 42-56.

[4] National Book Centre of Greece.*The Book Production in Greece. 2004,* Athens: National Book Centre of Greece, May 2006, pp. 2-3.

[5] BANOU, C. (2005/2006), "Money and Taste: New roles for the Greek publishers in a changing era. A case-study of small publishing markets", *The International Journal of the Book,* vol. 3, number 2, pp. 39-46.

[6] ibid.

[7] National Book Centre of Greece.*The Book Production in Greece. 2004,* Athens: National Book Centre of Greece, May 2006, p. 39.

[8] THOMPSON, J. B. *Books in the Digital Age. The Transformation of Academic Publishing in Britain and the United States,* Cambridge: Polity Press, 2005, p. 15.

[9] LYTRAS, M.; SICILIA, M.; DAVIES, J.; KASHYAP, V. “Digital libraries in the knowledge era. Knowledge management and semantic webb technologies”, *Library Management*, vol. 26 (4/5), 2005, pp. 170-175.

[10] DOBREVA, M. “IT applications of the medieval Slavonic written cultural heritage”, *Proceedings. 1st International Conference on Typography and Visual Communication. History, Theory, Education, T*hessaloniki: University of Macedonia Press, 2004, pp. 161-170.

[11] DEAKIN, S.; MICHIE, J. “ The Theory and Practice of Contacting”, in Contracts, Co-operation, and Competition. Studies in Economics, Management and Law, edited by Simon Deakin and Jonathan Michie, Oxford University Press: Oxford, 1997, pp. 1-39.

[12] BEHRAKIS, T. *Multidimentional Data Analysis: methods and practices*. Athens: Livanis, 1999 [in Greek].

[13] HARWOOD, T.G.; GARRY, T. “An Overview of Content Analysis. *The Marketing Review*. Vol.3, 2003, pp. 479-498.

[14] XIN L., “Library as incubating space for innovations: practices, trends and skill sets”, Library Management, 27 (6/7), 2006, p. 37-378.

[15] ANTELMAN, K. Do Open-Access Articles Have a Greater Research Impact?”, *College and Research Libraries,* 65.5 (Sep. 2005), p. 372-282.

[16] SCHIFFRIN, A. *The Business of Books. How International Conglomerates Took Over Publishing and Changed the Way we Read*, London - New York: Verso, 2001.

[17] STEVENSON, I. “The liveliest of corpses”: trends and challenges for the future in the book publishing industry, *Aslib Prodeedings,* 52 (4), April 2000, pp. 133-137.

# Towards a Semantic Turn in Rich-Media Analysis

*Tobias Bürger; Georg Güntner*

Salzburg Research Forschungsgesellschaft m.b.H.
A-5020 Salzburg, Jakob-Haringer-Strasse 5/III, Austria
e-mail: {tobias.buerger, georg.guentner}@salzburgresearch.at

## Abstract

Typical application scenarios in the area of rich-media management, such as the continuous digitisation of the media production processes, the search and retrieval tasks in a growing amount of information stored in professional and semi-professional audio-visual archives, as well as the availability of easy-to-use hard- and software tools for the production of rich-media material in the consumer area, lead to an increasing demand for a meaning-based management of digital audio-visual assets. This "semantic turn" in rich-media analysis requires a semantic enrichment of content along the digital content life cycle and value chain: The semantic enrichment of content can be achieved manually (which is expensive) or automatically (which is error-prone). In particular, automatic semantic enrichment must be aware of the gap between meaning that is directly retrievable from the content and meaning that can be inferred within a given interpretative context. Each solution has its benefits and drawbacks. Our paper discusses the relevance of semantic analysis of rich-media in certain application scenarios, compares two possible approaches, a semi-automatic and an automatic approach, and presents a case study for an automatic solution. Following the observations of the case study, we come up with recommendations for the improvement of the semantic enrichment by an manual annotation step.

**Keywords:** semantic web; multimedia content management; semantic indexing

## 1 Introduction

As a motivation for the application of semantic technologies in the area of rich-media analysis we want to highlight the following application scenarios: Firstly, the continued digitisation of the media production process at professional content providers and content distributors (e.g. broadcasters, telecommunication companies) not only leads to an exponential growth of highly unstructured digital material, but also to an increased demand for a reliable classification of audio-visual material along all stages of the digital content value chain. Secondly, in the consumer area and the semi-professional area (e.g. corporate media archives, or small and medium sized audio-visual archives) the easy-to-use production tools lead to an unmanageable amount of audio-visual material, that rather later than sooner has to be managed and indexed in some way, whereas the meaning of the digital essences often is locked in the raw content. Thirdly, even if basic metadata is available to describe the content and its meaning, user-centred applications are increasingly demanding the utilisation of the benefits of the true semantic search approach, i.e. inference and reasoning, narrowing down and widening the search by using some kind of formal knowledge representation.

To exemplify the above scenarios, imagine a user who wants to find out recordings of performances of sacred music by Wolfgang Amadeus Mozart in and around the city of Salzburg during the Salzburg Festival 2005. This query is full of hidden semantics (see also figure 1 for a schematic presentation of this query and the associated knowledge model): Location based semantics (what does "in and around the city of Salzburg" mean?), time based semantics (when was "Salzburg Festival 2005"?), factual semantics: e.g. which works are considered to be "sacred works"; which musical forms are known to be "sacred works" in general (e.g. a choral, a mass); which particular works by Wolfgang Amadeus Mozart are sacred works?

The importance of the knowledge related with the query is, that nothing of it has to be encoded in the media essences or their description: all this knowledge can be modelled, described and used without any particular relation with the digital essences.

Our research group is currently investigating different approaches and methods for the combination of media content with semantic annotations and for the usage of pre-existing knowledge (i.e. the "context") for inferring further knowledge about the content automatically. We have faced the question of annotation from low-level content analysis recently, in the national research project Smart Content Factory (SCF) and we are going to address the issue of merging new content with existing knowledge in the IST project LIVE (see section 2 and 3). In another project, Smart Content Factory, we tried to automatically derive the semantics of TV news clips in order to make them browse- and searchable. To do this, we combined information extracted from raw multimedia content with domain knowledge about multimedia data.

In LIVE which deals with broadcasting of media events by integrating different videos streams with background information about these media events, we investigate how domain knowledge and background information can be efficiently combined to deduce further knowledge from broadcast live video-streams.

As we have experienced in the Smart Content Factory [1], semantic descriptions of content can enhance fast and easy navigation through audio-visual repositories. Semantics - i.e. the interpretation of the content - is important to make content machine-processable and to enable the definition of tasks in workflow-environments for knowledge workers in the content industries. Some of the recent research projects in the area of semantic (or symbolic) video annotation try to derive the semantics from the low level features of the audiovisual material or from other available basic metadata, e.g. by audio-classification of classification of camera movement. Some of the projects aim at highly automated indexing using the results of automatic speech recognition however error-prone they may be. Most of these approaches are - as also pointed out in [2] - not capable to derive the semantics of multimedia content because in many cases the results of the analysis cannot be related to the media context [3]. For humans the construction of meaning is an act of interpretation that has much more to do with pre-existing knowledge (the "context") than with the recognition of low-level-features of the content. This situation is commonly referred to as the "semantic gap" [4].

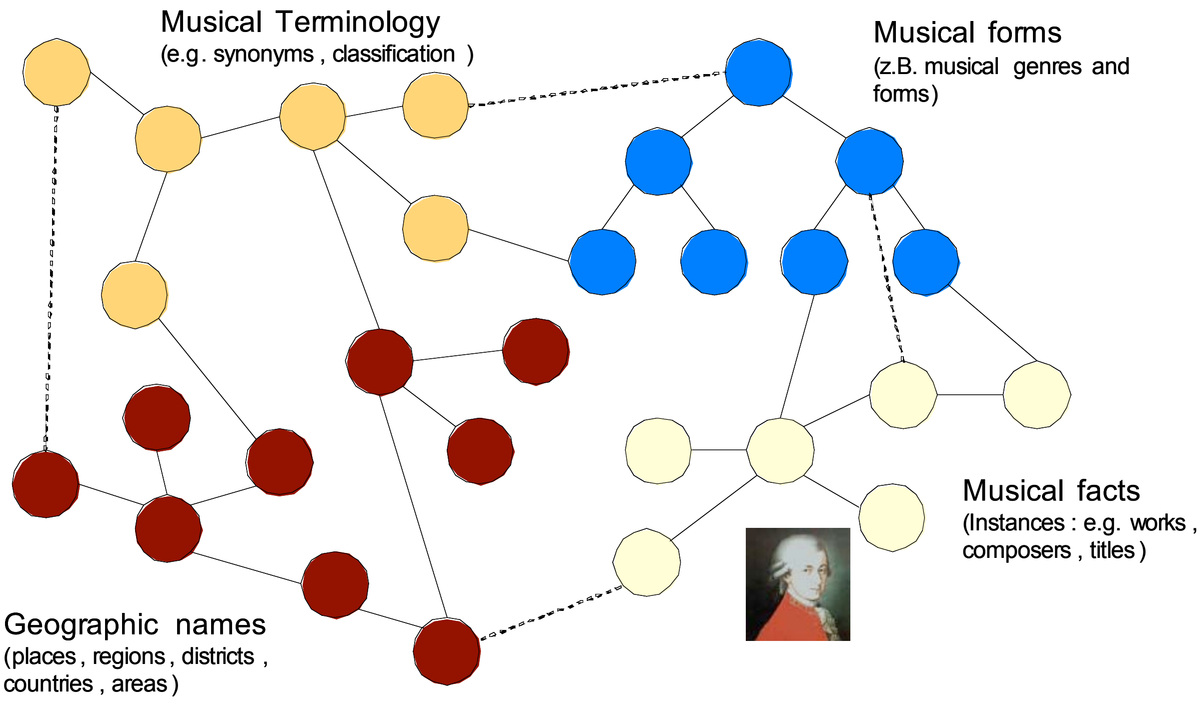

Musical Terminology
(e.g. synonyms , classification )
Musical forms
(z.B. musical genres and forms)
Musical facts
(Instances : e.g. works , composers , titles )
Geographic names
(places , regions , districts , countries , areas )

**Figure 1: The knowledge-base for the semantic query about Mozart's sacred works**

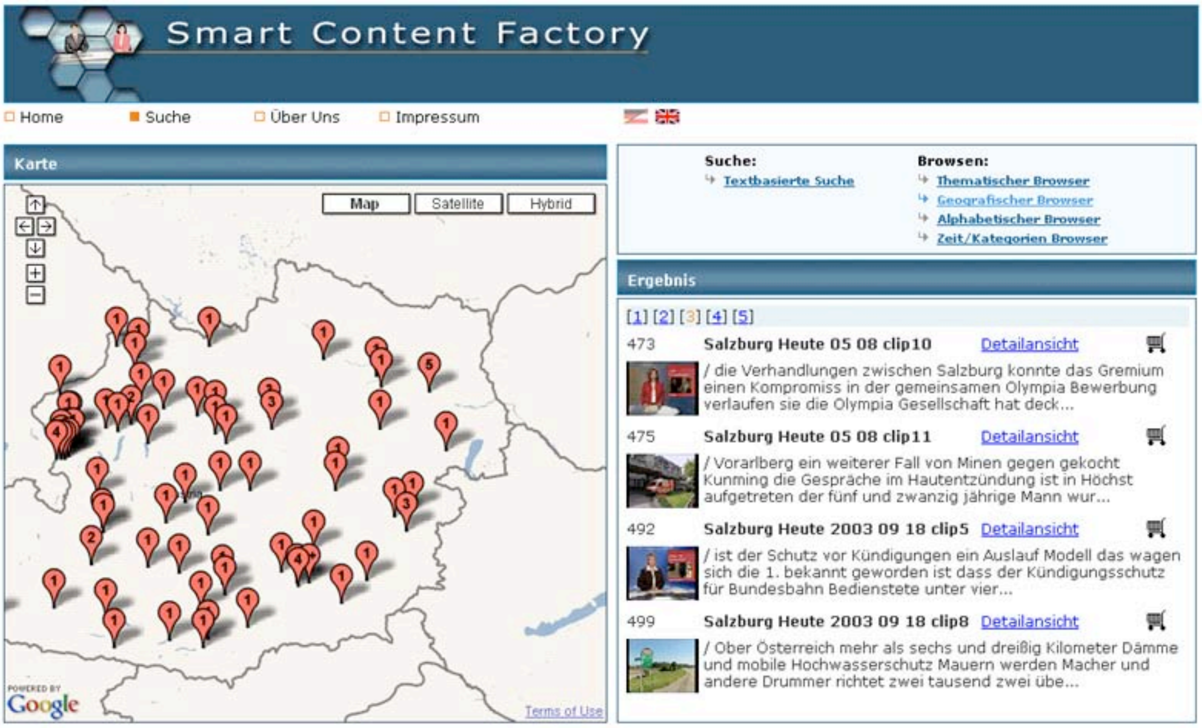
Two solution paths have emerged for this problem: The first one is to provide rich annotations created by humans as training data for the system to learn features of videos for future automatic content-based analysis. The second approach does not rely on training, but purely on analysis of the raw multimedia content. The training approach is not well suited for scenarios in which a great amount of content has to be annotated before any training and automation can be done or in which the application domain is very broad. The second approach usually only works well in settings where the relevant concepts can easily be recognized. However, most content based services demand richer semantics. As pointed out in section 4, popular examples on the Web show that there are currently many service-based platforms that make use of their users' knowledge to understand the meaning of multimedia content.

In our paper we concentrate on different approaches to close this semantic gap and provide insight into two solution paths, one automatic and one semi-automatic and demonstrate a case study on a prototypical solution for the "semantic augmentation" in the area of audio-visual archives.

## 2 Methodology: Automatic Vs. Semi-Automatic Semantic Rich-Media Analysis

In general, metadata generation systems can be classified in manual-, semi-automatic- and automatic annotation tools: The aim of automatic and semi-automatic tools for the analysis of rich-media content is to extract as much useful information from the raw media file as possible. Manual annotation tools aim to provide support for users to add metadata by hand.

Currently many systems try to expose the semantics of multimedia data by adding metadata to it. However, most of them do not derive these annotations just from the low-level features detected in the raw media data, but instead for example either analyze the different modalities of a video, analyse the usage context of the media or rely on human annotation/interpretation to derive higher-level semantic features from multimedia data.
In this section we want to introduce different approaches that we applied in two research projects making use of semantic technologies for rich-media analysis.



473 **Salzburg Heute 05 08 clip10** Detailansicht

/ die Verhandlungen zwischen Salzburg konnte das Gremium einen Kompromiss in der gemeinsamen Olympia Bewerbung verlaufen sie die Olympia Gesellschaft hat deck...

475 **Salzburg Heute 05 08 clip11** Detailansicht

/ Vorarlberg ein weiterer Fall von Minen gegen gekocht Kunming die Gespräche im Hautentzündung ist in Höchst aufgetreten der fünf und zwanzig jährige Mann wur...

492 **Salzburg Heute 2003 09 18 clip5** Detailansicht

/ ist der Schutz vor Kündigungen ein Auslauf Modell das wagen sich die 1. bekannt geworden ist dass der Kündigungsschutz für Bundesbahn Bedienstete unter vier...

499 **Salzburg Heute 2003 09 18 clip8** Detailansicht

/ Ober Österreich mehr als sechs und dreißig Kilometer Dämme und mobile Hochwasserschutz Mauern werden Macher und andere Drummer richtet zwei tausend zwei übe...

**Figure 2: Location based navigation in the Smart Content Factory**

Firstly, a project called "Smart Content Factory" (2003-2006) was driven by the idea to develop a system infrastructure for the knowledge-based search, retrieval and navigation in audio-visual archives of news clips. The approach was highly relying on an automatic feature extraction, mainly the speech-to-text transcription during the first phase. In a late phase this automatically extracted features were supported by additional reliable sources available in the digital production process at the Austrian Broadcasting Corporation (ORF). Section 3.1 describes the results and major findings of the selected approach with respect to the semantic indexing.
Secondly, an ongoing integrated project called "LIVE staging of media events" (started n 2006) is driven by an approach to combine the methods of automatic and semi-automatic detection, extraction and annotation of content with a knowledge-base under the control of a semantic based media framework. Moreover the framework propagates knowledge and contextual information to s recommender system which thus to some degree becomes aware of the meaning of the media. Section 3.2 describes the current state of this approach.

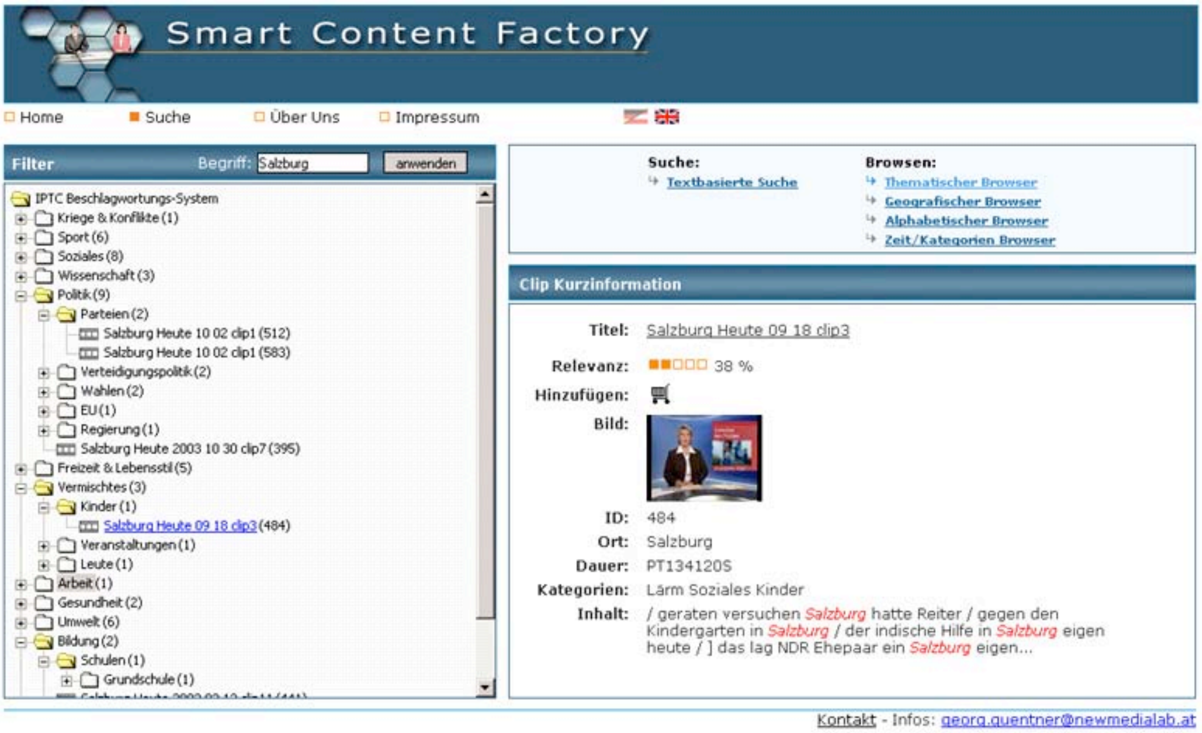
# 3 Results: Towards A Semi-Automatic Reliable Semantic Analysis Framework

This section describes the results of two research projects with respect to their selected approach for the semantic analysis.

## 3.1 Smart Content Factory – An Automatic Approach to Semantic Rich-media Analysis

In a research project called Smart Content Factory [5], we developed a prototype of a system infrastructure for the automatic knowledge-based refinement of audio-visual content repositories based on state-of-the-art digital asset management systems. In the project an automatic approach was used which was based on a two step indexing pipeline:

> In the first step a primary index is created by methods provided by state-of-the-art media analysing tools (i.e. Virage Video Logger™ and the Smart Encoding™ process). Video clips are passed to the Factory in MPEG-1 format. A polling mechanism informs the indexing and contextualisation components about newly available video clips and triggers the indexing process. The primary indexing results in the creation of key frames, the automatic detection of scenes, the transformation of speech to text, the recognition of speakers, etc.
>
> Subsequently a semantic indexing process is started which is based upon the results of the previous content based indexing and relies primarily on the speech to text transformation (i.e. "audiologging", for which different solutions have been tested in the course of the project). The dependency on the results of the audiologging is a weak point in the concept of the Smart Content Factory in so far as the subsequent semantic indexing builds on the results of a per se error-prone automatic extraction process. The semantic indexing is using the Lucene indexing framework [6] and the ontologies and thesauri forming the knowledge base of the Smart Content Factory which are accessible via "pluggable" RDF knowledge components described in the previous section. By applying and using knowledge models during the semantic indexing we create a set of "smart indices", allowing search, retrieval and reasoning along various dimensions of the information space.



Begriff: Salzburg

Titel: Salzburg Heute 09 18 clip3

Relevanz: ■■□□□ 38 %

Hinzufügen:

Bild:

ID: 484

Ort: Salzburg

Dauer: PT134120S

Kategorien: Lärm Soziales Kinder

Inhalt: / geraten versuchen Salzburg hatte Reiter / gegen den Kindergarten in Salzburg / der indische Hilfe in Salzburg eigen heute / ] das lag NDR Ehepaar ein Salzburg eigen...

**Figure 3: Category based navigation in the Smart Content Factory**

One of the key issues of our semantic indexing framework of the Smart Content Factory was the use of an extensible set of formal knowledge models (accessible via the Jena RDF framework [7]):

(1) The first knowledge model ('locations') contains a thesaurus of geographic names. The thesaurus extends the properties of gazetteers by modelling the hierarchical relations between the geographic locations (e.g. 'village' is-part-of 'political district'). The geoname thesaurus is based on data structured according to the ADL Feature Type Thesaurus [8] and is represented in RDF, the gazetteer is structured according to the ADL Gazetteer Content Standard [9]. This model allows specialisation and generalisation of search queries by location concepts. The recognition of location names is further supported by an engine for the recognition of named entities. In the Smart Content Factory we used LingPipe [10] to resolve ambiguities:

Scanning texts for occurrences of common known Austrian place names like 'Wien' (Vienna) or 'Salzburg' is rather easy. But it is a lot harder to semantically distinguish an appearance of the name of the Austrian village 'Haus' from the German word for house (the building). Due to the lack of a bigger training set, the location name recognition's precision/recall is not very high, but it serves as a starting point for distinguishing false positives from true positives. A simple tf/idf-based ranking is used to determine the 'most significant' location which a video is related to. Tf/Idf means 'term frequency-inverted document frequency' [11].

(2) To identify thematic categories we used the IPTC thesaurus (International Press and Telecommunications Council, [12]) which defines a hierarchical structure of thematic news categories (e.g. sports, policy, economy). For our purpose the IPTC thesaurus is represented in RDF according to SKOS Core 1.0, an RDF schema defined by W3C for the description of thesauri and similar types of knowledge models [13]. Similar to the location name identification process, this process identifies terms from a controlled vocabulary provided by the IPTC thesaurus.

(3) A web-based synonym service for (German) words was integrated into the semantic indexing process. The service was created and is maintained by the University of Leipzig ('Deutscher Wortschatz' [14]). The Web service interface is based on the SOAP protocol. By means of this service the index is enhanced with synonyms of non-stop words. All models are either integrated via a Web service interface or stored in a database and are accessed via the Jena RDF framework [7], which also provides a powerful inferencing mechanism (e.g. traversing of hierarchical relations).

In the Smart Content Factory, one of the objectives for the introduction of the knowledge-based index was the enhancement of search and retrieval and navigation support [1].

The main benefits of this approach were the little need for human intervention in the process as the annotations were totally generated automatically. Another benefit was that the approach was extensible as other thesauri/knowledge models could be easily plugged in to recognise for example events or dates. One drawback however was the amount of false positives locations that were recognised, which was mainly due to the bad results of the text to speech engine. Another drawback was the amount of time needed to index a video, which doesn't allow real-time indexing of the video.

Figure 2 shows the application of the location based semantics for the map based search nd navigation in the audio-visual archive. Figure 3 exemplifies the category based navigation, using the IPTC thematic thesaurus which is represented in RDF as described above. Both navigation paradigms were highly appreciated by the test user group during a user evaluation in 2006, whereas other forms, e.g. the hyperbolic tree navigation paradigm were ranked low in the users' interest profile.

## 3.2 LIVE – Real-Time Semi-Automatic Annotation of Videos with the Intelligent Media Framework

The integrated project "LIVE Staging of Media Events" (LIVE; FP6-27312, [15]) aims at the creation of novel intelligent content production methods and tools for interactive digital broadcasters to stage live media events in the area of sports, such as the 2008 Olympic Games. In the terminology of the project, "staging live media events" is a notion for the creation of a non-linear multi-stream video show in real-time, which changes due to the interests of the consumer (end user). From a technical viewpoint, this requires a transformation of raw audiovisual content into "Intelligent Media Assets". LIVE will develop a knowledge kit and a toolkit for an intelligent live content production process including dynamic human annotation and automated real-time annotation. As a consequence novel iTV video formats for live events will evolve.

In the LIVE project we applied the lessons learnt from the automatic approach of the Smart Content Factory to overcome the weaknesses of automatic metadata extraction and extended the system architecture to meet the requirements of real-time semantic indexing. In the first phase of the project we started to design an Intelligent Media Framework that is taking into account the requirements of real-time video indexing to combine several automatic and manual annotation steps. The Intelligent Media Framework thereby integrates the following components of the LIVE production support system:

> (1) The Intelligent Media Asset Information System (IMAIS) providing access to services for the storage of media, knowledge models and metadata relevant for the live staging process and providing services for the creation and management and delivery of intelligent media assets. This will be the central component of the Intelligent Media Framework and will semantically enrich incoming metadata streams;
> (2) The Recommender System, giving content recommendations to the user based on the user's personal profile and on previous user feedback;
> (3) The Metadata Generation System, dealing with the detection, extraction and annotation of knowledge from audiovisual material;
> (4) The Video Conducting System, dealing with the real-time staging of a live event.

The components of the indexing pipeline in LIVE are shown in figure 4: the Automatic Analysis Application, the Human Annotation Tool and the Intelligent Media Framework. The role of the Intelligent media Framework is to accept and handle partial information about particular media items, to add semantic information to the items and to infer and attach contextual knowledge to the items that is probably related to the event that is staged. It furthermore provides knowledge services that offer controlled vocabularies related to the current context of a stream to guarantee the unambiguousness of the terms used.

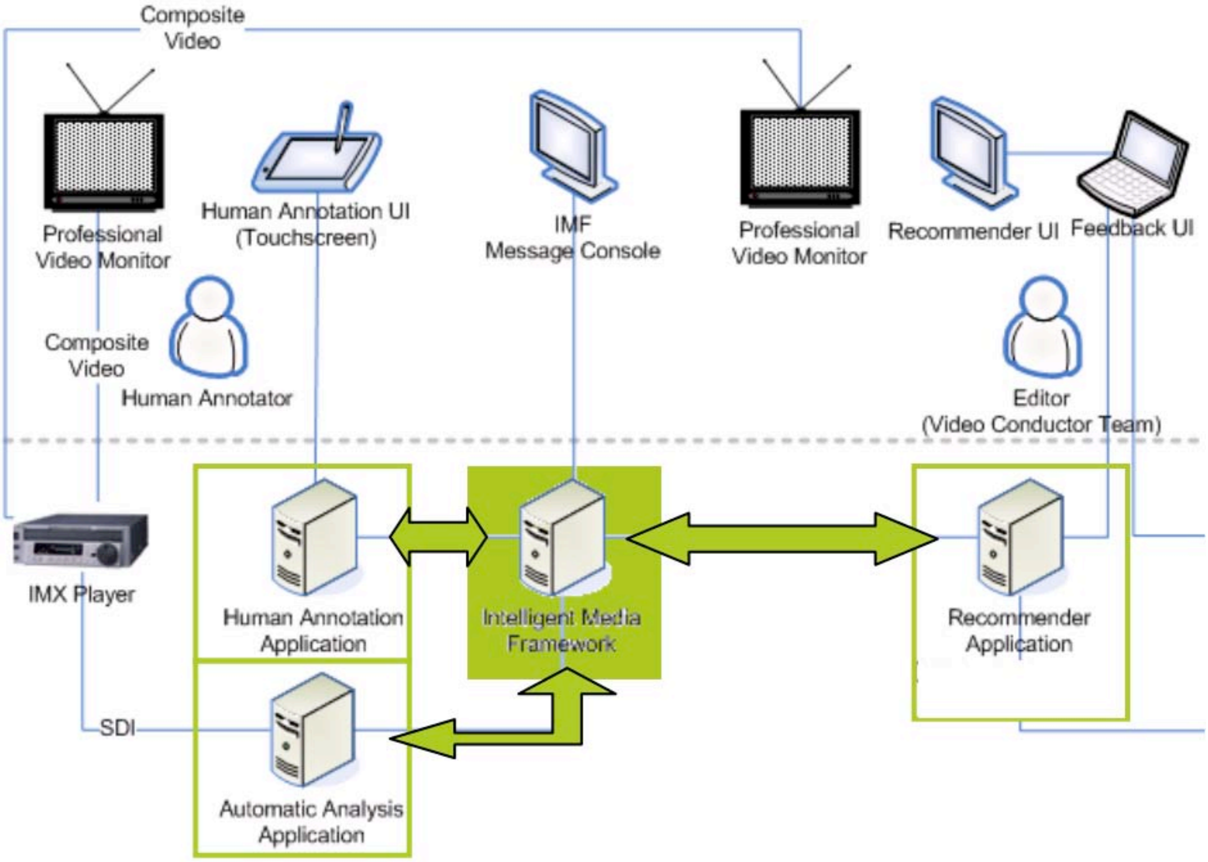

Composite Video
Professional Video Monitor
Human Annotation UI (Touchscreen)
IMF Message Console
Professional Video Monitor
Recommender UI
Feedback UI
Composite Video
Human Annotator
Editor (Video Conductor Team)
IMX Player
Human Annotation Application
Intelligent Media Framework
Recommender Application
SDI
Automatic Analysis Application

**Figure 4: Category based navigation in the Smart Content Factory**

The semantic enrichment process in LIVE is twofold: The Automatic Analysis Application detects close-ups, shots, faces, camera-motion, colour schemes, scenes and artists. This information is enriched in (1) a manual step done by the human annotator through the human annotation tool (2) in the Intelligent Media Framework that has knowledge about the context of the analysed media item. In (1) terms from the controlled vocabulary are assigned to the low-level information that was extracted in the basic analysis step. In (2) these terms are used to attach more semantic information of the current action or event to the media items that is possibly inferred by the current event schedule or other particular information that was detected in the course of this event.

This semantic indexing process is more reliable than the approach from the Smart Content Factory, because it is neither based on error-prone text transcripts nor totally relies on automatic analysis tools. One key enabler of this semantic indexing step is the use of existing information systems at the broadcaster's side that have knowledge about the staged event, the participants and so on. The most important step in this process, however, is the human annotation that is later inferred by the Intelligent Media Framework in a reliable way. This allows us to act in real-time (with a maximum delay of approx. 20ms) and provides high-level metadata helping to bridge the gap between the raw audio-visual essences and their intended meaning.

Figure 4 shows parts of a prototypical demonstration setup shown during a LIVE review meeting in Vienna in March 2007. Both, a human annotation tool and the automatic annotation system, are using a semantic aware middleware, the Intelligent Media Framework, which provides context information and the controlled vocabulary for the annotation process and propagates the detected "meaning" to a recommender system for the professional user (editor, video conductor). The video conductor team decides which live streams and switching possibilities are offered to the consumers.

## 4 Discussion

In this section we list existing approaches that try to extract knowledge from rich-media items and try to relate these approaches to the LIVE approach. Indexing and metadata generation is a common task to media analysis systems and there are sound algorithms and methods in the area of computer vision, pattern recognition, natural language processing and signal processing that can be applied, most of them applicable for extracting low-level features from the essences. Currently many systems try to expose the semantics of multimedia data by adding annotations to it. However, the common trait of all the following examples is that they do not derive these annotations just from the low-level features detected in the raw media data, but instead for example either analyse the different modalities of a video, analyse the usage context of the media or rely on human annotation/interpretation to derive higher-level semantic features from multimedia data.

Recent research efforts try to combine automatically derived features like speech-to-text transcripts with background knowledge and related information on the Web: an application called Rich News [16] deals with automatic annotation and extraction of semantics from news videos: In a first step, the system extracts text from speech and then it tries to extract the most important topics from that. With these extracted topics, the system starts a Google search to find news stories on the web that cover the same topic(s). Google did exactly the same in a research project [17] to enhance American TV news with background information from the Internet: They extract important topics from the subtitle channel that is broadcasted with every news show and give consumers the possibility to get background information about these news items by displaying links to related web pages. MediaMill [18], a system that was developed at the University of Amsterdam, uses all modalities of a video to derive the semantics of it, which is especially important when the visual content is not reflected in the associated text like close captions or speech-to-text transcripts. Besides these examples from the research community there are also some major trends in industry: Some of the established industry-strength systems come from the classical document management area with its sophisticated full text indexing and information retrieval methods. The industrial players are now extending these methods to audiovisual content (e.g. Convera's RetrievalWare [19] or Autonomy's IDOL Server [20]). Other vendors come from the digital asset management sector and extend their systems to full text indexing and knowledge management methods (e.g. Virage's VS Archive [21]).
These systems rely heavily on metadata and on full-text indexation which in turn, is based on speech-to-text extraction, but also make use of statistical methods to classify content according to taxonomies or thesauri. The leading search engines (e.g. Google or Yahoo) are currently extending their search features to audio-visual media, but they are still to a high degree relying on metadata and on text transcripts provided along with the media assets. Besides that, Google and Yahoo both released versions of their search engines to discover contents in videos: Google for example has been indexing news stories from an U.S. broad-caster. There the search index is based on the closed caption provided along with news clips. They also scan the Web for videos and images: In that case - as also described in [22] - additional metadata are generated from adjacent information on the website (e.g. text blocks or the video file name).

However, there are also other services dealing with content which are already very popular among a large user community: Two of the most popular content-based services are Flickr [23] and Last.FM [24], both of which get their users to classify (tag) images or music files within these systems. These systems do not classify the content according to predefined categories, but instead, taxonomies evolve from the users' tags (folksonomies). These tags can be regarded as user-based knowledge about certain items that on these platforms, is used to filter or recommend content to other users. The main requirements in LIVE are (1) reliability of annotations as fast decisions have to be made based on them and (2) the real-time assignment of these annotations. The mentioned approaches are shortly discussed according to their usefulness in LIVE.

As mentioned above Web sites such as Flickr [23], or also Riya [25] recently began to apply algorithms for automatic extraction of content metadata (e.g. shapes, colour or texture features) and some of them already use high level pattern recognition technology such as face detection (e.g. Riya) However the metadata provided by them is not reliable enough. Other approaches that tend to provide reliable metadata information like MediaMill [18] or the commercial tools like Virage's VS Archive [21] perform the necessary analysis not fast and accurate enough. Manual annotation tools like Vannotea [26], Advene [27] or M-Ontomat-Annotizer [28], to name just a few existing annotation tools, are to complex to be utilizable in the LIVE real-time situation. LIVE or the IMF also goes one step beyond the research as it tries to include contextual information in the analysis as much as possible, however at the moment solely in the LIVE domain.

## 5 Outlook and Conclusion

In the Smart Content Factory we experienced that it on the hand is possible to automatically extract low-level features from video and augment them through the use of semantic technology but on the other hand this information is sometimes error-prone as it has to be based on incomplete or wrongly extracted features. In LIVE wrong or incomplete information could lead to wrong decisions in the live staging process which to some extent forces us to augment extracted information manually to reach a high degree of reliability. The use of controlled vocabularies and semantically enhanced metadata throughout the whole LIVE system introduces a common language that will lead to fast and reliable decisions during the staging process. The Intelligent Media Framework is responsible for the augmentation of the automatically extracted information that is additionally manually refined. Our next steps in the development of the IMF are the further addition of contextual information to the metadata sets of single media items, such as information related to the event or the athletes.

## Acknowledgements

Smart Content Factory is one of the lead projects at Salzburg NewMediaLab (SNML [29]), the Austrian centre of excellence in the area of digital content engineering and new media technologies. Salzburg NewMediaLab is funded by the Austrian Federal Ministry of Economics and Labour and the Province of Salzburg. The project started in 2003 and ended in September 2006. It was coordinated by Salzburg Research: Further partners were ORF (the Austrian Broadcasting Corporation), X-Art ProDivision and Joanneum Research (all Austrian partners). LIVE (Live staging of media events, FP6-27312, [15]) is an integrated, multidisciplinary initiative that will contribute to the IST strategic objective "Semantic based knowledge and Content Systems" and "Exploring and bringing to maturity the intelligent vision". The project is funded by the European Commission within the 6th Framework Programme. The project started in 2006 with a duration of 45 months. It is coordinated by Fraunhofer IAIS. Further partners are Salzburg Research, the University of Ljubljana, the University of Bradford, the University of Applied Sciences in Cologne, the Academy of Media Arts Cologne, ORF, Pixelpark and ATOS Origin.

## References

[1] BÜRGER, T.; GAMS, E.; GÜNTNER, G.: *Smart Content Factory - Assisting Search for Digital Objects by Generic Linking Concepts to Multimedia Content*. In: Proceedings of the Sixteenth ACM Conference on Hypertext and Hypermedia (HT '05), 2005.

[2] BLOEHDORN, S. et al.: *Semantic Annotation of Images and Videos for Multimedia Analysis*. In: Proceedings of the 2nd European Semantic Web Conference, ESWC 2005, Heraklion, Greece, May 2005.

[3] BÜRGER, T.; WESTENTHALER, R.: *Mind the gap - requirements for the combination of content and knowledge*. In: Proceedings of the first international conference on Semantics And digital Media Technology (SAMT), December 6-8, 2006, Athens, Greece.

[4] SMEULDERS, A. W. M. et al.: *Content-Based Image Retrieval at the End of the Early Years* In: IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 22 No. 12, December 2000.

[5] Smart Content Factory project web-site: http://scf.salzburgresearch.at/ – Last visited: 10.04.2007

[6] The Apache Jakarta Project: Lucene. From: http://jakarta.apache.org/lucene - Last visited: 10.04.2007

[7] Jena – A Semantic Web Framework for Java. From: http://jena.sourceforge.net/ - Last visited: 10.04.2007

[8] University of California, Santa Barbara: Alexandria Digital Library Feature Type Thesaurus, http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/ - Last visited: 12.10.2004

[9] University of California, Santa Barbara: *Guide to the Alexandria Digital Library Gazetteer Content Standard*, http://www.alexandria.ucsb.edu/gazetteer/ContentStandard/version3.2/GCS3.2-guide.htm - Last visited: 12.10.2004

[10] LingPipe: http://www.alias-i.com – Last visited: 10.04.2007

[11] Term frequency-inverted document frequency (tf/idf): http://en.wikipedia.org/wiki/Tfidf – Last visited: 10.04.2007

[12] International Press and Telecommunications Council. From: http://www.iptc.org/ - Last visited: 10.04.2007

[13] MILES, A. J. *SKOS Core - Guidelines for Migration* .http://www.w3.org/2001/sw/Europe/reports/thes/1.0/migrate/ - Last visited: 01.04.2007]

[14] University of Leipzig, Institute of Computer Sciences. Project "Deutscher Wortschatz" (German dictionary). From: http://wortschatz.informatik.uni-leipzig.de/ - Last visited: 28.04.2005

[15] LIVE – Live staging of media events; project web-site: http://www.ist-live.org – Last visited: 10.04.2007

[16] DOWMAN, M.; TABLIN, V.; URSU, C.; CUNNINGHAM, H.; POPOV, B.: *Semantically enhanced television news through web and video integration* In Proceedings of the Workshop on Multimedia and the Semantic Web at the European Semantic Web Conference (ESWC 2005), 2005.

[17] HENZINGER, M.; CHANG, B.-W.; MILCH, B.; BRIN, S.: *Query-Free News Search*. In Proc. of the 12th World Wide Web Conference, pp. 1-10, 2003.

[18] SNOEK, C. G. M. et al.: *MediaMill - Exploring News Video Archives based on Learned Semantics*. In: Proceedings of the ACM Multimedia Conference 2005, 2005

[19] Convera - http://www.convera.com/ - Last visited: 10.04.2007

[20] Autonomy - IDOL Server: http://www.autonomy.com/content/Products/IDOL/ – Last visited: 10.04.2007

[21] Virage: http://www.virage.com – Last visited: 10.04.2007

[22] FERGUS, R. ; PERONA, P.; ZISSERMAN, A.: *A Visual Category Filter for Google Images*. In: Proceedings of the 8th European Conference on Computer Vision, Prague, Czech Republic, 2004.

[23] Flickr: http://www.flickr.com – Last visited: 10.04.2007

[24] Last.FM: http://www.last.fm – Last visited: 10.04.2007

[25] Riya: http://www.riya.com/ – Last visited: 10.04.2007

[26] Vannotea: http://liris.cnrs.fr/advene/ – Last visited: 10.04.2007

[27] Muvino: http://vitooki.sourceforge.net/components/muvino/code/index.html – Last visited: 10.04.2007

[28] M-Ontomat-Annotizer : http://www.acemedia.org/aceMedia/results/software/m-ontomat-annotizer.html – Last visited: 10.04.2007

[29] Salzburg NewMediaLab (SNML), "Kompetenzzentrum für Neue Medien" (competence centre for new media technologies and digital content engineering): http://www.newmedialab.at/ – Last visited: 10.04.2007

# Towards an Ontology of ElPub/SciX: A Proposal

*Sely M S Costa* [1]; *Claudio Gottschalg-Duque* [2]

[1] University of Brasília, Department of Information Science
Campus Universitário Darcy Ribeiro, Brasília, DF, Brazil
e-mail: selmar@unb.br
[2] University of Brasília, Department of Information Science
Campus Universitário Darcy Ribeiro, Brasília, DF, Brazil
e-mail: klauss@unb.br

## Abstract

A proposal is presented for a standard ontology language defined as ElPub/SciX Ontology, based on the content of a web digital library of conference proceedings. This content, i.e., ElPub/SciX documents, aims to provide access to papers presented at the total editions of the International Conference in Electronic Publishing (ElPub). After completing its 10th years in 2006, ElPub/SciX is now a comprehensive repository with over 400 papers. Previous work has been used as a basis to build up the ontology described here. It has been presented at Elpub2004 and it dealt with an Information Retrieval System using Computational Linguistics (SiRILiCo). ElPub/SciX ontology constitutes a lightweight ontology (classes and just some instances) and is the result of two basic procedures. The first one is a syntactic analysis carried out through the Syntactic Parser-VISL. This free tool, based on lingsoft's ENGCG parser, is made available through the Visual Interactive Syntactic Learning, a research and development project at the University of Southern Denmark, Institute of Language and Communication (ISK). The second one, carried out after that, is a semantic analysis (concept extraction) conducted through GeraOnto, an acronym that stands for "generating an ontology", which extracts the concepts needed in order to build up the ontology. The program has been developed by Gottschalg-Duque, in 2005, in Brazil. The ensuing ontology is then edited via Protégé, a free, open source ontology editor. The motivation to carry out the work reported here came from problems faced during the preparation of a paper to Elpub2006, which aimed to present data about a number of aspects regarding the ElPub/SciX collection. While searching the collection, problems with the lack of standardization of authors and institutions names and the non-existence of any control of keywords had been identified. Such problems seem to be related to an apparent absence of "paper preparation" before entering into the SciX database. Lack of preparation, in turn, has brought about the desire of finding a solution, which is expected to support the work of those interested in searching the collection to retrieve information. ElPub/SciX ontology, therefore, is seen as that helping solution to support ElPub information retrieval.

**Keywords:** ontology; Elpub conferences; information retrieval

## 1 Introduction

Electronic publishing constitutes a hot topic of discussion within the academic environment, particularly in the study of scholarly communication. Such interest is due to the opportunities provided by the web and the Internet for a document to be available world wide in electronic format. As a free, democratic environment, the Internet provides a huge amount of information, which, on the other hand, presents a challenge for those who seek relevant hits. The digital content available today actually represents a great chaos to those interested in finding relevant information for research or in scrutinising a document collection for the same purpose.

There have been a variety of approaches to help with this matter, such as those that made possible to develop thesauri, ontologies, taxonomies, topic maps and other resources. They have been developed with the aim of facilitating the intellectual work. Therefore, a well-organised collection should be supported by one of them. In this context, an ontology can be considered one of the richest resources used for the automatic treatment of electronic documents, since it constitutes a set of definitions of a *formal* vocabulary.

In this paper, we present a proposal for a standard Ontology language defined as SciX ontology, based on the content of SciX, a digital library of conference proceedings. SciX can be viewed as a response to the need of organising and making available a collection of papers presented in an annual international conference. It is actually a web digital library that provides access to papers presented at the International Conference in Electronic Publishing (ElPub). The conference completed its 10th years in 2006, and SciX is now a comprehensive repository with over 400 papers. The collection comprises papers presented at the two traditional tracks of sessions, namely, general and technical, as well as abstracts of keynote speeches, workshops, round tables and special sessions presentations as well as other kinds of contribution.

## 2 Motivation and Expectations

During ElPub2006, the 10th version of the conference, a short paper was published with observations on a few quantitative data [1]. The analysis of papers from the 10 years conferences showed that SciX content does needs a standardisation process of its data in order to improve search and retrieval for research purposes.

Since the work of Paul Otlet and Henri La Fontaine, regarding documentation, retrieval of relevant content of a document is deemed the key factor of success in any information service/product. In this regard, ontologies have the capability of significantly improving retrieval needed in information services. The proposed ontology will certainly help the exploration of SciX content in both quantitative and qualitative approaches, in the extent that ontologies constitute a set of *classes* (for example, author, title, key-words), *individuals* (for example, Leslie Chan, University of Toronto) and *properties* (http://www.utsc.utoronto.ca) that allows a sounder work on the data available.

It is interesting to note that a vocabulary ontology expressed in a formal specification, such as the Web Ontology Language (OWL), makes possible machine processing of information (in a very basic level), rather than simple data, adding expediency to web content search and retrieval. Based on this understanding, the authors of this paper have decided to develop an ontology to help the work of researchers or practitioners interested in using SciX data for research. The leading objective is to provide an “information resource”, allowing the generation of a richer domain-specific knowledge, which is a formal specification of a controlled vocabulary.

The work carried out on the ElPub/SciX collection in 2006 has actually allowed the standardisation of authors and institutions names. The output, however, has not been aggregated to that collection until now. Taking into account that the digital collection so far reproduces the information provided by authors themselves, the work on keywords standardization requires a controlled vocabulary, gradually built up while processing SciX collection. Nevertheless, despite more than 10 years, it seems possible to create this control and help future authors to rely on the output for better stating keywords pertaining to their papers.

The ontology makes it possible to define nodes of semantic relationships and make inferences concerning the topics covered by authors. In addition, a number of relationships between concepts are possible to identify, which, in turn, can respond to the need of standardising them.

This paper, therefore, reports the experience of developing a semi-automated process of extracting concepts from an electronic document collection, in order to create an ontology. It is semi-automatic because of the non-automated procedures concerning part of the data extracted from the database. The idea is to develop an optimised, interesting output of a scholarly papers collection, with the aim of making it easier to be handled by researchers. Through its semantic net, ElPub/SciX ontology is intended to provide better conditions for the user to find the information needed more efficiently. These standard metadata, besides other possibilities, can contribute to define a new environment based on ElPub/SciX Ontology concepts.

It is noteworthy to call attention to the fact that authors have always used different ways of informing both their own names and their institutions names wherever, and whenever they publish. Moreover, different authors define the same topic differently. Concerning Elpub, another aspect that deserves attention is related to the conference sessions’ title, which do not always represents a 'core topic’ to which the content of those session papers converge. This, in turn, makes an accurate content analysis difficult to carry out.

Such ambiguities and inconsistencies are probably related to the way Elpub has been conducted. It is observable that, as the conference progressed, a number of procedures started to be implemented to make it more organised. At least three of these improvements are clearly identified. Firstly, preoccupation with well-formatted/presented papers based on a well-planned template, along with guidelines about what is expected from authors, has helped improve the content entered into SciX. Secondly, requirements of an abstract, keywords and other data, not present in some of the first editions of the conference, seem to have had an impact on such improvement. Thirdly, the definition of the conference sessions’ titles, over the last editions (and 1997’s!!), well depicts the content of papers presented in those sessions. Nevertheless, for the enhancement of the access and use of this content, a standardisation is urgently needed. ElPub/SciX ontology, with no doubt, can definitely contribute to that. In spite of still being in an incipient stage as yet, the ontology has the potential of accomplishing the aforementioned purpose.

# 3 Methodology (Theoretical Approach and Methodological Procedures)

In order to carry out the study, a number of procedures have been performed, in accordance to what has been stated in theories that underlie content analysis and retrieval in a web environment. As it has been asserted, *the use of ontology as a formal explicit specification of a shared conceptualisation, can help to solve the problem of inefficiency, overloaded "fake information", ambiguity and chaos* [2]. These authors draw attention to the fact that the use of automatic semantic analysers (despite its earlier own approaches with some incipient, but encourage results) makes possible extract the conceptual structure, describe phrases and use semantic relationships between words and concepts to establish connections between them. This structure, which constitutes a meta-level description, is a representation that brings order to the collection of documents, so it can be understood as an ontology, in the sense defined by Gruber [3] and others. That is, in computer science, the term 'ontology' expresses explicit formal specifications of terms in a domain and the relationships among them.
Notwithstanding the improvement afforded by an ontology, two problems remain, ambiguity and inconsistencies. Names ambiguity have been approached as one of the major problems in retrieving information from a database. As observed by Han et al [4], "*because of name variations, identical names, name misspellings or pseudonyms, two types of ambiguities in research papers and bibliographies can be observed*". They are authoring multiple name labels and multiple authors sharing the same name label. The same authors point out that "*it may affect the quality of scientific data gathering, can decrease the performance of information retrieval and web search, and even may cause incorrect identification of and credit attribution to authors*".

The solution, that is, name disambiguation, has been approached in a variety of ways and has always been related to the creation of authority files. Auld, cited by French et al [5], stressed that this sort of strategy has been called 'authority work' and have mostly benefited from computational procedures.

It is interesting to emphasise that name ambiguity can be related to a number of entities, such as authors, institutions, journal or conference titles and so forth. Ambiguities in institutions names have been approached by French et al, who looked at techniques to aid in detecting variant forms of strings in bibliographic databases. They highlight Taylor's approach [6], whose first principle of authority control is concerned with all variants of a name being "brought together under a single form so that once users find that form, they will be confident that they have located everything relating to the name". This 'single form' has been defined as 'canonical name' and, in the work of French and his colleagues, consisted of deciding on a set of canonical affiliation strings and then, assigning each affiliation string in the database to one of these canonical strings.

As can be inferred, disambiguation of names is crucial to the work proposed here, as the simple creation of the ontology itself could not solve this sort of problem by itself. It has been partially and preliminary performed by Costa et al [7] in order to carry their analysis out. Nevertheless, it has been a non-automated process in the sense that no computational procedure was used.

The result, however, corresponds to Taylor's first principle of authority control and is used in this second work upon ElPub/SciX collection. That is why it is still a 'semi-automatic' extraction process. Further work will develop an automated procedure for the creation of canonical names of both authors and institutions.

As regard the ontology and the procedures developed in order to create it, the work was based on the previous model developed by Gottschalg-Duque [8], adapted for this application (Figure 1), as it does not yet include the indexing module. As can be observed, the whole process consists of the stages *file conversion, natural language processing and ontology creation and editing*. The implementation of the modules and sub-modules (figure 2 shows a detailed view of the natural language processing module, comprised of two sub-modules, syntactic and semantic) is done by means of three programs, which are Syntactic Parser, GeraOnto and Protegé.

The stages involved in the analyses and in the ontology creation are succinctly described further, and show how each of them is performed, along with the indication of the software used. It is important to highlight that GeraOnto, because of patent problems, does not allow giving any detail.

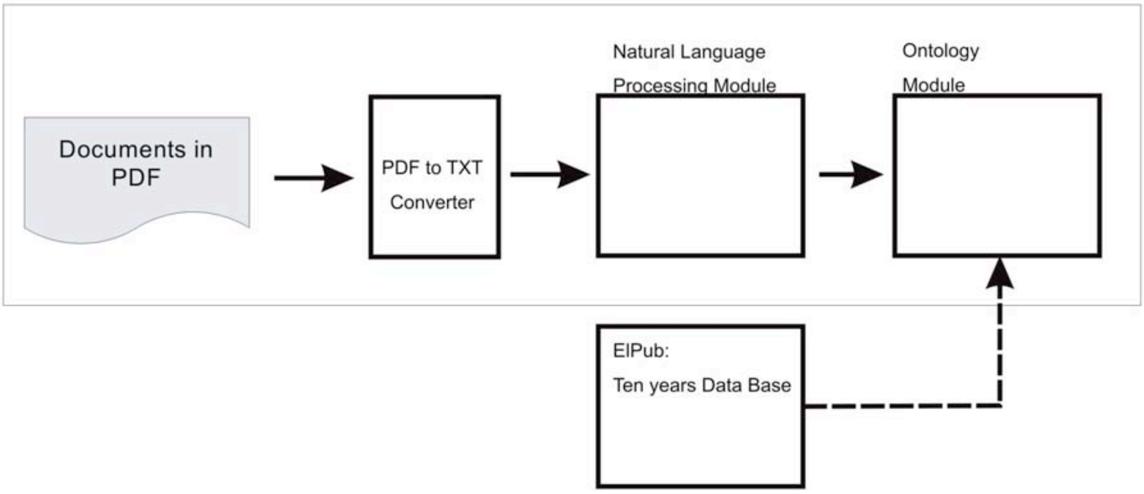


**Figure 1: The process of creating the ElPub/SciX Ontology**

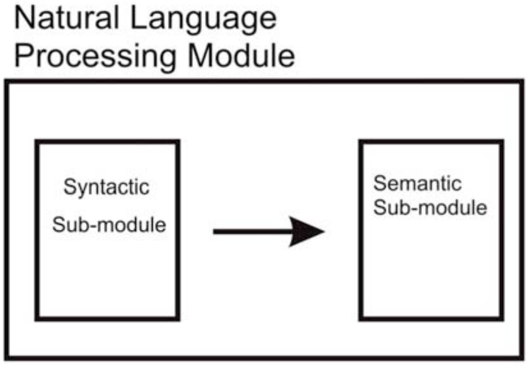


**Figure 2: Detail of the two sub-modules of the natural language processing**

The ontology construction policy adopted pointed to the definition of what constitutes the relevant concepts that should compose the ontology structure (Figure 3).

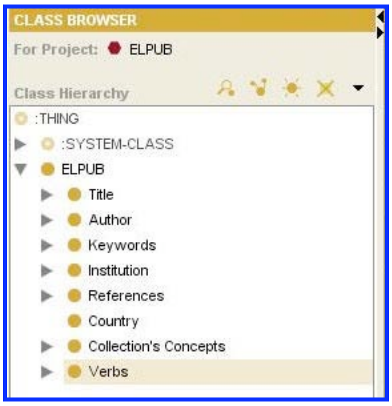
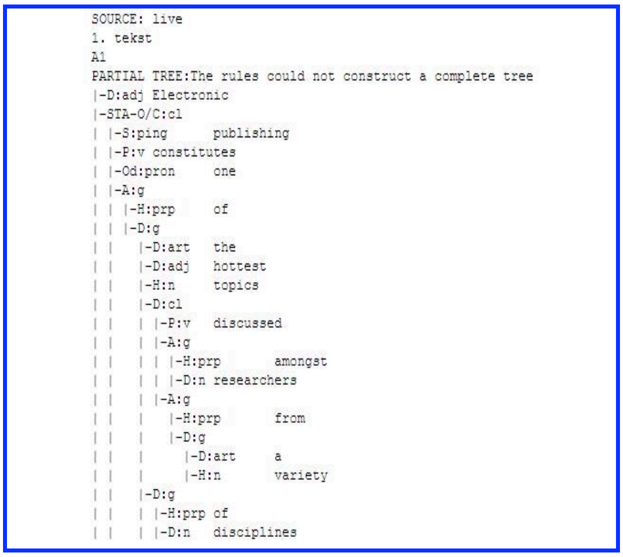


**Figure 3: The ontology structure**

# 4 Results

The procedures carried out comprised the following steps and produced results as exemplified in figures 4, 5 and 6:

- Visit SciX site and collect the entire collection of ElPub papers;
- Transfer the collection into a native database;
- Manually extract titles, author's and institution's names, as well as keywords;
- Replace authors and institution names in the native database by the canonical names created by Costa et al [11]. It is interesting to point out that, for institution names, canonical affiliation strings have been created by applying the rule of putting names given by authors in a standard order. That is: university, faculty/school/institute, department and programme/project, whatever appears. For authors names, the rule was to adopt the most complete form of a name;
- Convert all pdf files into txt files;
- Send the texts (from the introduction to just before the references) to a syntactic analyser (**Syntactic Parser - VISL**), which automatically performs the analysis and generates a syntactic tree with all syntactic tags (example in figure 4);

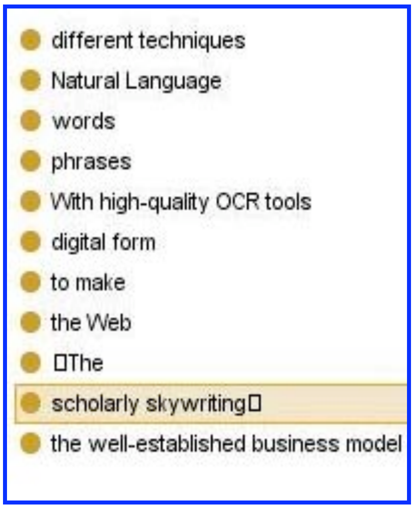


```
SOURCE: live
1. tekst
A1
PARTIAL TREE:The rules could not construct a complete tree
|-D:adj Electronic
|-STA-O/C:cl
| |-S:ping      publishing
| |-P:v constitutes
| |-Od:pron     one
| |-A:g
| | |-H:prp     of
| | |-D:g
| |   |-D:art   the
| |   |-D:adj   hottest
| |   |-H:n     topics
| |   |-D:cl
| |   | |-P:v   discussed
| |   | |-A:g
| |   | | |-H:prp       amongst
| |   | | |-D:n researchers
| |   | |-A:g
| |   |   |-H:prp       from
| |   |   |-D:g
| |   |     |-D:art     a
| |   |     |-H:n       variety
| |   |-D:g
| |   | |-H:prp of
| |   | |-D:n   disciplines
```

**Figure 4: Syntactic Parser output**

- Send the syntactic tree to *GeraOnto*, which extracts the semantic elements (noun phrases and verbs) of interest for the construction of the ontology. These are concepts that can or cannot be composed of more than one term or concept;

- Insert these concepts into *Protegé*, which edits the ElPub/SciX Ontology, using SciX's record identifiers as its slots (examples of the output are shown in figures 5, 6 and 7).



- different techniques
- Natural Language
- words
- phrases
- With high-quality OCR tools
- digital form
- to make
- the Web
- The
- scholarly skywriting
- the well-established business model

**Figure 5: Concepts automatically extracted from a text**

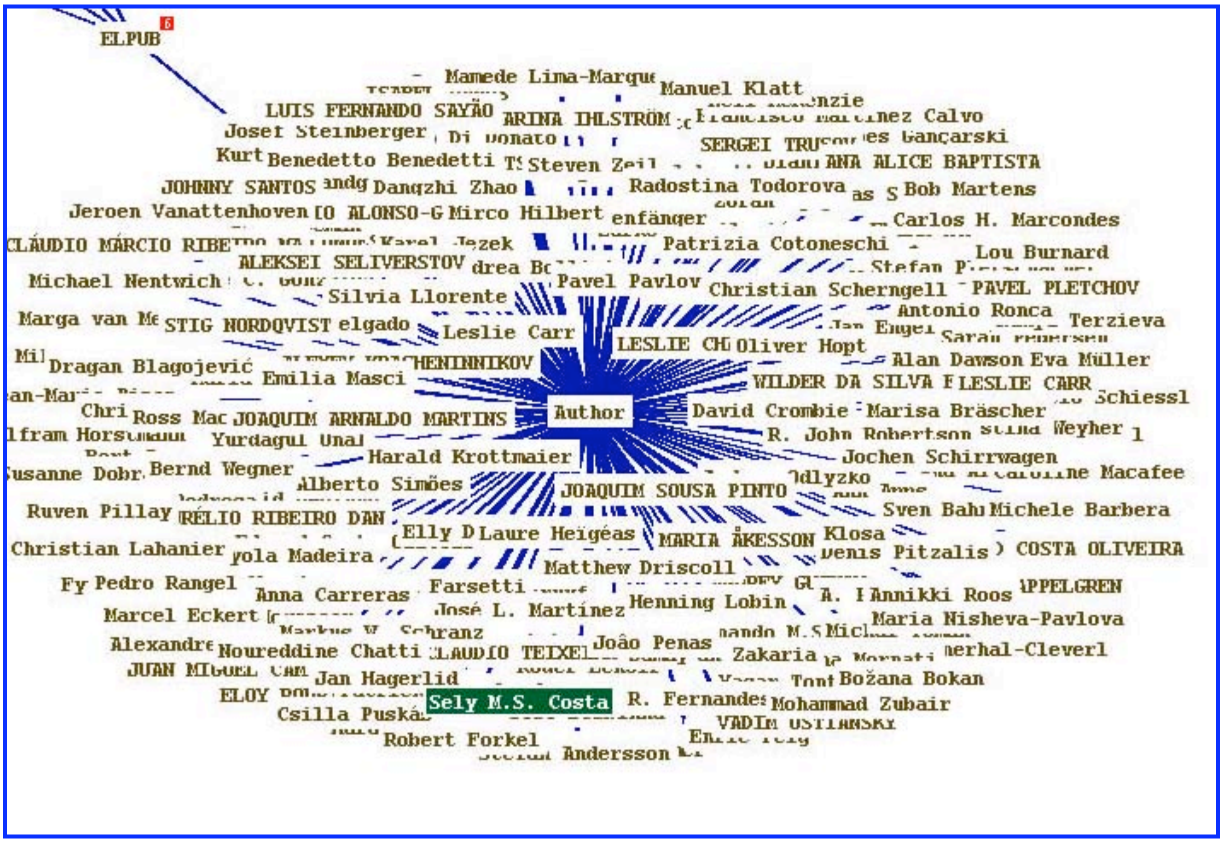


**Figure 6: Graphic presentation of the 'author' super class and its sub-classes**

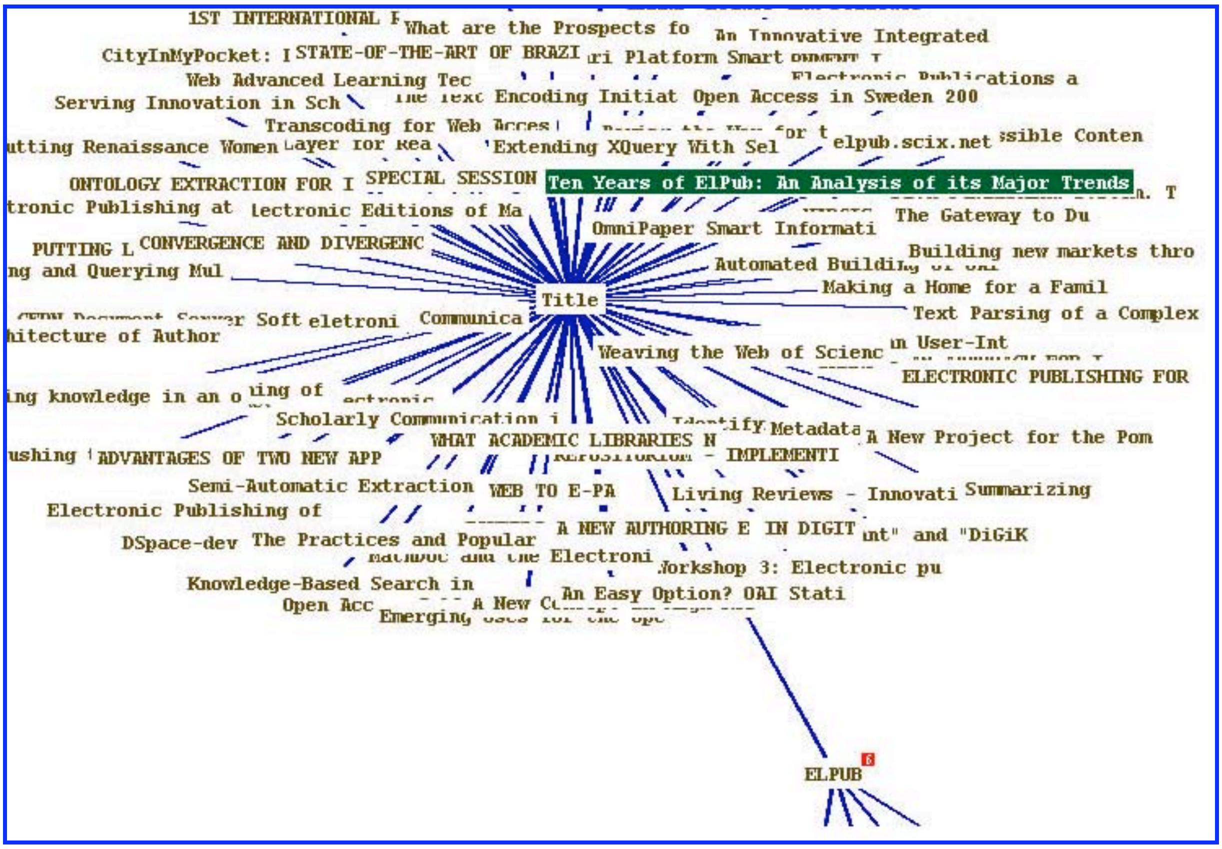

1ST INTERNATIONAL F
What are the Prospects fo
An Innovative Integrated
CityInMyPocket: I
STATE-OF-THE-ART OF BRAZI
ri Platform Smart
Web Advanced Learning Tec
Electronic Publications a
Serving Innovation in Sch
The Text Encoding Initiat
Open Access in Sweden 200
Transcoding for Web Acces
utting Renaissance Women
Layer for Rea
'Extending XQuery With Sel
elpub.scix.net
ssible Conten
ONTOLOGY EXTRACTION FOR I
SPECIAL SESSION
Ten Years of ElPub: An Analysis of its Major Trends
tronic Publishing at
lectronic Editions of Ma
OmniPaper Smart Informati
The Gateway to Du
PUTTING L
CONVERGENCE AND DIVERGENC
Building new markets thro
ng and Querying Mul
Automated Buildin
Making a Home for a Famil
Title
Soft eletroni
Communica
Text Parsing of a Complex
hitecture of Author
Weaving the Web of Scienc
User-Int
ELECTRONIC PUBLISHING FOR
ing knowledge in an o
Scholarly Communication i
Identify Metadata
A New Project for the Pom
ushing
ADVANTAGES OF TWO NEW APP
WHAT ACADEMIC LIBRARIES N
IMPLEMENTI
Semi-Automatic Extraction
WEB TO E-PA
Living Reviews - Innovati
Summarizing
Electronic Publishing of
A NEW AUTHORING E
IN DIGIT
nt" and "DiGiK
DSpace-dev
The Practices and Popular
Electroni
Workshop 3: Electronic pu
Knowledge-Based Search in
Open Acc
A New C
An Easy Option? OAI Stati
Emerging
ELPUB
6

**Figure 7: Graphic presentation of the 'title' super class, with its sub-classes**

Results obtained so far have extracted more than 4,000 concepts. Some of them, especially those related to keywords, need human interference in order to be refined and standardised. Nevertheless, as a work in progress, a number of improvements are still taking place. One of them is to automatically generate the authority names file by creating authors and institutions names as super classes of the ontology, and all their variant names as their sub-classes.

# 5 Conclusions and Recommendations

This proposal minimises, rather than completely eliminates noises identified in information retrieval from ElPub/SciX collection, by creating an ontology. Besides the creation of authority names control, it does reduce ambiguities by means of the syntactic analysis. Instances (SciX record identifiers, such as ELPUB2004_11elpub2004.content.pdf) identified help to prevent ambiguity of explicit repetition of terms. The combined use of Syntactic Parser-VISL, GeraOnto and Protegé has proved to be very helpful and useful for this work.

The resultant ontology will be available at SciX, as well as a tutorial that is intended to be developed and made available for those interested in it. The major learning, however, has been the solutions and strategies identified so far in order to deal with the problem studied. It appeared very clear that the use of such ontology indeed enhances the information retrieval from a collection like ElPub/SciX, particularly because it reduces ambiguities. Regarding keywords, the creation of a controlled vocabulary is highly recommended, based on what is already available on ElPub/SciX collection, in order to guide prospective authors in defining them.

## Acknowledgements

The first author of this work has been generously sponsored by *Finatec* (http://www.finatec.org.br). We are grateful to Rafael Odon de Alencar, who developed java implementations that allowed the development of the ontology with a collection in English, since GeraOnto was created for a Portuguese collection of texts and needed some adaptations.

## Notes and References

[1] COSTA, S. M. S.; BRÄSCHER, M; MADEIRA, F.; SCHIESSL, M. Ten years of ElPub: an analysis of its major trends. In: Martens, B.; Dobreva, M. (Eds.) *Digital spectrum: integrating technology and culture*. Proceedings of the Elpub conference. Bansko : FOI-COMMERCE, 2006. pp. 395-399.

[2] GOTTSCHALG-DUQUE, C. ; LOBIN, H. Ontology extraction for index generation.. In: COSTA, Sely M. S.; ENGELEN, Jan; MOREIRA, A. C. S. (Eds.) *Building digital bridges: linking culture, commerce and science*. Proceedings of the 8th ICCC International Conference on Electronic Publishing. Brasília, 2004. pp. 111-120.

[3] GRUBER, T. *What is an ontology*, 1996. Available at: http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

[4] HAN, H.; ZHA, H.; GILES, C. L. Name disambiguation in author citations using a K-way spectral clustering method. In: *International Conference on Digital Libraries archive*. Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries. New York : ACM Press, 2005.

[5] FRENCH, J. C.; POWELL, A. L.; SCHULMAN, E. Using clustering strategies for creating authority files. *Journal of the American Society for Information Science*, 2000, vol. 51, no. 8, pp. 774-786.

[6] FRENCH, J. C.; POWELL, A. L.; SCHULMAN, E.; PFALTS, J. L. cite the work of Taylor, published in 1984. Their article is about authority files and has been published in 1997, in the proceedings of the ECDL 1997.

[7] COSTA, ref. [1]

[8] GOTTSCHALG-DUQUE, C. *SiRILiCO uma proposta para um Sistema de Recuperação de Informação âbaseado em Teorias da Lingüística Computacional e Ontologia*. Belo Horizonte, 2005.

# On the Evolution of Computer Terminology and the SPOT On-Line Dictionary Project

*Jiri Hynek; Premek Brada*

Department of Computer Science & Engineering, Faculty of Applied Sciences, University of West Bohemia
Univerzitní 22, 306 14 Pilsen, Czech Republic
e-mail: {jhynek, brada}@kiv.zcu.cz

## Abstract

In this paper we discuss the issue of ICT terminology and translations of specific technical terms. We also present SPOT – a new on-line dictionary of computer terminology. SPOT's web platform is adaptable to any language and/or field. We hope that SPOT will become an open platform for discussing controversial computer terms (and their translations into Czech) among professionals. The resulting on-line computer dictionary is freely available to the general public, university teachers, students, editors and professional translators. The dictionary includes some novel features, such as presenting translated terms used in several different contexts – a feature highly appreciated namely by users lacking technical knowledge for deciding which of the dictionary terms being offered should be used.

**Keywords:** terminology; dictionary; language; lexicography; translation; wiki; information technology

## 1 Introduction

Ordinary users and experts alike get the feeling that Czech translations of computer documentation (including books, help files, reference guides) are of inferior quality [1,2]. It is likely that the same applies to other languages. Readers often prefer original documents in English, which discriminates against those readers that either have no access to these materials, or lack the required linguistic knowledge.

It is generally believed that what makes a translator's work difficult is the technical terminology. But in reality, the main problem is to grasp the *meaning* of specific terms, the actual thing that is denoted by the term, use of the term in the *context*, occurrence of the term with other terms, and the term's stylistic features.

The situation is complicated by the fact that translations are created (literally made up, invented) by technical staff and software developers who lack linguistic skills and „feeling" for the natural language; or at the opposite end of the spectrum, by „professional" translators who lack skills and terminology in the specific ICT (information and communication technology) sub-domain.

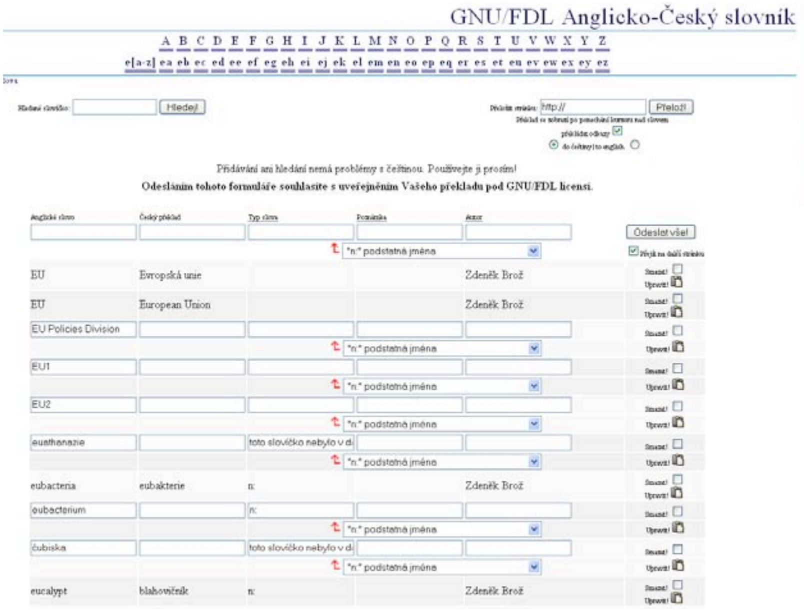
Complaints about the poor quality of translations are often backed by the lack of uniformity in addressing specific technical terms. In other words, different authors give the same thing different names. Activities aimed at standardizing computer terminology are very rare; one of the very few is Microsoft Terminology Translations [3] available for 59 different languages at the time of writing. The glossary provided in the form of a CSV file contains more than 12,000 English terms plus the translations of the terms. Unfortunately, translated terms are not provided for all the terms in the list, and some translations seem to be controversial.

In this paper we would like to discuss the issues which are encountered when translating terms arising in a rapidly developing domain like the ICT, including taxonomy of the terms in view of their development status. Secondly we present a project of an on-line dictionary aiming at supporting the work of translators and localization developers. The structure of the paper is as follows. In the following section we discuss the work of translators and the tools they have available. Section 3 describes the taxonomy of terms and some issues with respect to translation. Section 4 presents the SPOT on-line dictionary, the goals of the project, key distinguishing characteristics of the tool, and its current status. The paper is finished with a conclusion.

## 2 Dictionaries and Resources for Translators

We are living in the age of an information explosion. We would need more than one year to study the amount of information produced worldwide within a single day. It is up to translators to cope with all the changes that take place in the field of their specialty. Searching for technical terminology is one of the most time-consuming tasks of every professional translator. Fortunately, the majority of support data can be found by means of search engines, such as Google. Online dictionaries are also available, although the contents of their free versions are often limited, or they are too general for specific purposes. We are happy to observe the trend of utilizing "collective wisdom" in special Internet projects (Wikipedia being the best example), with users selflessly sharing their expertise with their fellows. Our project is one of those.

The SPOT online dictionary presented here (see http://spot.zcu.cz/) is not the only project of its kind at our University. There is the English-Czech GNU/FDL dictionary available at http://slovnik.zcu.cz/online/ and www.wordbook.cz, which is based on the i-spell database [4] – see Figure 1 below.



GNU/FDL Anglicko-Český slovník

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

e[a-z] ea eb ec ed ee ef eg eh ei ej ek el em en eo ep eq er es et eu ev ew ex ey ez

Hledané slovíčko: Hledej!

Překlad stránky: http:// Přelož!

Překlad se zobrazí po ponechání kurzoru nad slovem

překládat odkazy ☑

● do češtiny | to english ○

Přidávání ani hledání nemá problémy s češtinou. Používejte ji prosím!

**Odesláním tohoto formuláře souhlasíte s uveřejněním Vašeho překladu pod GNU/FDL licensí.**

| Anglické slovo | Český překlad | Typ slova | Poznámka | Autor | |
|---|---|---|---|---|---|
| | | "n:" podstatná jména | | | Odeslat vše!<br>☑ Přejít na další stránku |
| EU | Evropská unie | | | Zdeněk Brož | Smazat! □<br>Upravit! |
| EU | European Union | | | Zdeněk Brož | Smazat! □<br>Upravit! |
| EU Policies Division | | "n:" podstatná jména | | | Smazat! □<br>Upravit! |
| EU1 | | "n:" podstatná jména | | | Smazat! □<br>Upravit! |
| EU2 | | "n:" podstatná jména | | | Smazat! □<br>Upravit! |
| euathanazie | | toto slovíčko nebylo v d<br>"n:" podstatná jména | | | Smazat! □<br>Upravit! |
| eubacteria | eubakterie | n: | | Zdeněk Brož | Smazat! □<br>Upravit! |
| eubacterium | | n:<br>"n:" podstatná jména | | | Smazat! □<br>Upravit! |
| čubiska | | toto slovíčko nebylo v d<br>"n:" podstatná jména | | | Smazat! □<br>Upravit! |
| eucalypt | blahovičník | n: | | Zdeněk Brož | Smazat! □<br>Upravit! |

**Figure 1: English-Czech dictionary based on the i-spell database**

Of course, there are various general dictionaries and encyclopaedias freely available online. These represent an excellent source of linguistic information as well as wisdom, but they are unsuitable for our objective. You can visit the following:

- Merriam-Webster OnLine at: www.webster.com
- Encyclopaedia Britannica online at: www.britannica.com
- Oxford Advanced Learner's Dictionary at: www.oup.com/elt/oald/
- Wikipedia at: http://wikipedia.org/
- Citizendium at: http://en.citizendium.org/
- Dictionary, Thesaurus and Encyclopedia at: www.reference.com
- EuroWordNet (multilingual database with wordnets) at: www.illc.uva.nl/EuroWordNet/
- Free On-Line Dictionary of Computing at: http://foldoc.org/

A comprehensive list of on-line dictionaries for multiple languages suitable for both general and specific purposes can be found at http://a-z-dictionaries.com/online-dictionary.html, or www.yourdictionary.com.

## 2.1 CAT Tools for Professionals

Today's ICT localization projects often involve millions of words of software documentation, help files, warnings, error messages and other texts to be translated. In order to keep the documentation consistent, memory-based computer-aided translation (CAT) tools are a must. They have a substantial impact on both the translation quality and the productivity of all people involved in the localization project. The roll-out of CAT tools dates back to the beginning of the 1990s. Here are a few examples of today's most popular CAT tools (listed alphabetically):

- Deja Vu (www.atril.com)
- IBM Translation Manager (www.ibm.com)
- MetaTexis (www.metatexis.com)

- SDL Trados (www.trados.com)
- SDLX (www.sdl.com)
- Star Transit (www.star-ag.ch)
- Systran (www.systransoft.com)
- WordFast (www.wordfast.net)

Additional resources for translators and linguists can be accessed via www.multilingual.com, where you can find a collection of more than 1600 links divided into 41 categories (such as Automated Translation, Dictionaries, Internationalization Tools). There are also conferences dedicated to translation and localization industries, such as Localization World (www.localizationworld.com). The localisation community is also supported by The Localisation Research Centre at University College Dublin (www.localisation.ie). Useful links to various translation resources can be found at www.translation.net/links.html. You can download various glossaries at the translators' directory Go Translators (www.gotranslators.com). An example of a free memory translation database for multiple languages can be accessed at www.open-tran.eu.

## 3 English ICT Terms: From Old-timers to Troublemakers

We are witnessing rapid development in the ICT domain, with quickly changing terminology as one of the consequences. And the language of ICT professionals is suffering badly. The impact on other languages is serious – in the case of Czech, for example, the majority of terms are more or less adopted directly from English. Some terms find their Czech equivalent immediately, some are developing, and for some we still have no equivalent. For those that enjoy categorization, the following classes of computing terms can be identified.

### 3.1 Old-timers

This category includes many terms dating back to the 1960s. Their meaning has settled, as well as their translation into other languages. We can subdivide into:

- Morally obsolete terms (such as *bubble memory*, *punch tape*, *punch card*, and recently, also *floppy disk* and *diskette*);
- Stabilized and commonly used terms (such as *display*, *plotter*, *button*, *printer*, *mail server, dialog window*, *RAM, search engine*).

### 3.2 Novas and Supernovas

English is quite a flexible and even playful language with the ability to create and absorb new words (e.g. *text me* for "send me an SMS message", or abbreviations like *B4*), and new terms therefore appear very quickly, sometimes with an associated hype. Other languages need not be that flexible, which then poses problems in translating these new words, often created ad hoc. Unfortunately, this happens more than often in our mother language.

Compared to the category above, the number of terms labelled as Novas is relatively small. More difficult is the decision taken by a translator or a publication/magazine editor as to how to have these localized. This category includes: *cookie*, *spam*, *phishing*, *blog, freeware*, *emoticon*, *code closure, social bookmarking* and so on.

### 3.3 Troublemakers

The category of Troublemakers is relatively large. Thanks to Troublemakers, computer dictionaries get published and sold in large volumes, electronic dictionaries flourish on the Internet, and long debates among academics and language purists are held. It is up to computer users, editors and authors which terms shall prevail and which shall become obsolete. We can, again, subdivide into:

- **Homonymous Troublemakers** – identical computer terms with different meanings, such as *collector* (of a transistor) and (software) *collector* (e.g. data collector, portable collector representing a SW component), or *plug-in* (meaning an amplifier) vs. *plug-in module*, a SW component);

- **Synonymous Troublemakers –** The same (or virtually the same) meaning expressed differently, sometimes erroneously (such as *cross reference* vs. *cross-index* vs. *link*, *local menu* vs. *context menu* vs. *pop-up menu* vs. *shortcut menu*, *pull-down menu* vs. *drop-down menu* (or a *list*) *submenu* vs. *child* *menu, tool palette* vs. *toolbox, custom* vs. *personal, scroll bar* vs. *slide bar, clickable map* vs. *interactive map*).

### 3.4 Terms on the Move

- Over time, some IT terms **shift and/or extend their meaning**, such as the former *monitor*, meaning today's *screen*, and which now commonly refers to SW utilities; *link,* formerly used in the context of network connections (e.g. *link control protocol*) is now more often used for *web links* or *object linking*;

- **Everyday terms acquire new meaning** in the context of IT: e.g. *signature* is now commonly used in *method signature*, *virus signature*, or *digital signature*; *little endian* and *big endian* (adopted from Gulliver's Travels by Jonathan Swift) now refer to the method of storing multi-byte data; *pool* acquired the meaning of *fund*, i.e. a source of something (*thread pool*, *resource pool*); *key* becomes an "identifier" (*database primary key*); *root* (such as in *tree root*) becomes a type of user, or is used for *root folder*s; *builder* has a new meaning, such as in *application builder*, *list builder*, or *expression builder;* the word *seamless* has become a buzzword in the context of application integration; *heap* currently represents a type of memory; *thread* today refers to a sub-process (such as in *multithreading*); *stamp* has acquired a more abstract sense; *field* (such as in *sports field* or *mine field*) got the meaning of *entry/item*; *host* no longer refers just to persons, but also servers (either hardware or software); *wizard* is commonly used in *installation wizard* or *test wizard*; *garbage* now refers to meaningless data or data no longer needed; *docking* is used for toolbars or laptops instead of (space) ships;

- We are witnessing the process of **heavy verbalization**, e.g. *to right-click* (in place of "click the right mouse button"), *to cache* (in place of "save in cache"), *to host* ("to act as a host"), *personalize*, *televise*, *deserialize,* or *visualize*.

### 3.5 Esoteric Terms

- **Esoteric IT terms** are used by a relatively narrow group of IT specialists; these include, for example: *setter* and *getter* (in object-oriented programming), *undeploy*, *abstract factory*, *uptime*, *design pattern*, *marshalling*, *serialization*, *proxying*, *entity bean*, *refactoring*, *tight coupling*, *box model*, *locale*, and *hot-swap*;
- **Application- or corporation-specific terms**: these are quite special purpose, found in a limited number of applications, or used only within specific corporations or teams. For example: *rolling period*, *purge*, *context root*, *governance, updater, Apple menu,* and *Start menu*. Please note that these terms may have different meanings depending on the organization/application.

Both situations in which esoteric terms arise, summed by the "narrow user group" characteristic, make their localization complicated, since specialists are happy to use the original (non-translated) terms, while translators lack the technical background knowledge to make up suitable translations.

## 4 SPOT On-Line Dictionary

Aware of the issues discussed above, the authors started a new project to overcome the difficulties in creating suitable translations in the rapidly changing ICT domain. Our objective is to help the community of translators (plus editors and IT professionals) to either quickly find the correct translation for new, unusual, or tricky terms, or – if no such translation is known – to create one with the assistance of the "collective wisdom" of their colleagues.

Our work is novel in several aspects. Although there are various general-purpose dictionaries available that help in achieving the first objective, we are focusing strictly on the area of ICT, which undergoes frequent changes from the linguistics point of view. The SPOT dictionary is initially starting with a well-established terminology corpus, and its further development will be supported by our extensive experience in localization projects, besides the knowledge of leading IT professionals and editors.

Our approach is also new in that we "settle" the final version of Czech translations within the community of users, under the supervision of ICT specialists. It is anticipated that we will capture the latent interest of the community of translators, readers, editors and companies involved in localization projects (such as translations of computer publications, documentation and help files, translation of game scripts, and office software localization).

While we believe that professional editorial supervision is absolutely essential for acceptable "settling" of translated terms, the role of the mass of users is seen as a key differentiating point to classical approaches in translation and also as a critical success factor. Similarly to several successful Web 2.0 projects like Wikipedia (www.wikipedia.org), the SPOT project thus hopes to bring into fruitful cooperation professionals and users alike.

Finally, we also offer new features facilitating the work of large or distributed localization teams. Controversial terms are always on-line, rather than stored locally with a localization team member. It is essential that everyone uses the latest version of translated terms. SPOT also eliminates the need to redistribute up-to-date versions of dictionaries, which gets time-consuming in the case of large translation projects.

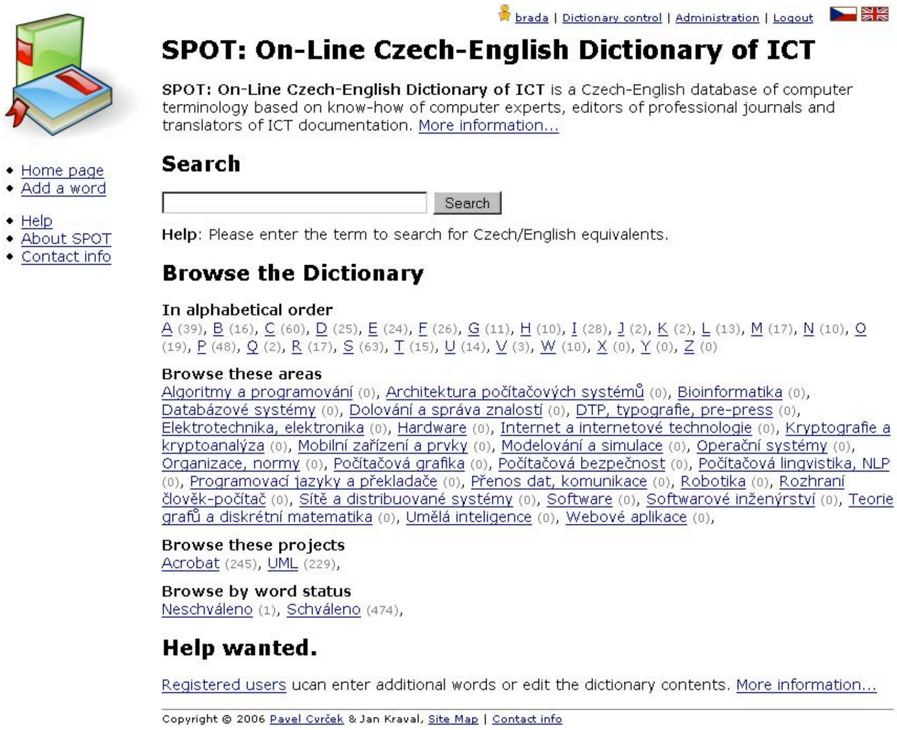
Last but not least, the added value of our dictionary is greatly enhanced by showing translations in various *contexts* based on on-line search of the Internet. The user can see which of the translations offered should be used, as it can be derived from the context information shown (see Figures 3, 4, 5 below).

## 4.1 The Dual Role of SPOT

As suggested by the previous paragraphs, SPOT will serve two complementary purposes: a reference dictionary of computer terms, and a platform for "settling" these terms.

### 4.1.1 Reference Dictionary

The basic function of this Internet dictionary is to provide Czech translations for specific ICT terms, either by browsing or by searching (see Figure 2 below). The quality of the translations is guaranteed by the initial English-Czech corpus based on the English-Czech ICT Dictionary written by the first author. Opportunities to provide valuable add-ons on top of the initial corpus are given by the fact that the dictionary is on-line.



SPOT: On-Line Czech-English Dictionary of ICT

**SPOT: On-Line Czech-English Dictionary of ICT** is a Czech-English database of computer terminology based on know-how of computer experts, editors of professional journals and translators of ICT documentation. More information...

Search

**Help**: Please enter the term to search for Czech/English equivalents.

Browse the Dictionary

In alphabetical order
A (39), B (16), C (60), D (25), E (24), F (26), G (11), H (10), I (28), J (2), K (2), L (13), M (17), N (10), O (19), P (48), Q (2), R (17), S (63), T (15), U (14), V (3), W (10), X (0), Y (0), Z (0)

Browse these areas
Algoritmy a programování (0), Architektura počítačových systémů (0), Bioinformatika (0), Databázové systémy (0), Dolování a správa znalostí (0), DTP, typografie, pre-press (0), Elektrotechnika, elektronika (0), Hardware (0), Internet a internetové technologie (0), Kryptografie a kryptoanalýza (0), Mobilní zařízení a prvky (0), Modelování a simulace (0), Operační systémy (0), Organizace, normy (0), Počítačová grafika (0), Počítačová bezpečnost (0), Počítačová lingvistika, NLP (0), Programovací jazyky a překladače (0), Přenos dat, komunikace (0), Robotika (0), Rozhraní člověk-počítač (0), Sítě a distribuované systémy (0), Software (0), Softwarové inženýrství (0), Teorie grafů a diskrétní matematika (0), Umělá inteligence (0), Webové aplikace (0),

Browse these projects
Acrobat (245), UML (229),

Browse by word status
Neschváleno (1), Schváleno (474),

Help wanted.

Registered users ucan enter additional words or edit the dictionary contents. More information...

Copyright © 2006 Pavel Cvrček & Jan Kraval, Site Map | Contact info

**Figure 2: SPOT interface**

Firstly, editors can assign categorization *tags* to individual terms, in a manner similar to some popular web services such as Flickr for photographs [5] or StumbleUpon for website links [6]. Users will thus be able to confine their search to a specific ICT area, or study other terms related to their area of interest:

- Algorithms and Programming
- Artificial Intelligence

- Communications
- Computer Graphics
- Computer Linguistics, NLP (Natural Language Processing)
- Computer Modelling and Simulations
- Computer Networks and Distributed Systems
- Cryptography, Cryptanalysis
- Data and knowledge mining
- Database Systems
- DTP, Pre-press
- Electrical Engineering and Electronics
- Hardware
- Internet and Web Technologies
- Man-machine Interface
- Mobile Devices
- Operating Systems
- Programming Languages
- Robotics
- Software

Secondly, we can enrich the information about the terms and their translations by displaying their occurrences in various *contexts* based on on-line search of the Internet. We use *Google search API* to obtain on-demand results, configured so that only sites with high relevance to ICT are searched. The probability that the context results will be meaningful is thereby increased. Having spent many years working on large localization projects, we are aware that showing a term's usage in the context is very useful and highly appreciated, especially by translators.

In its current implementation, SPOT can display the following context information (see Figures 3, 4, 5):

- Web sites, no restriction,
- IT webs only,
- Wikis,
- On-line dictionaries,
- Blogs (to be implemented).

A list of authoritative IT web sites, wikipedias, on-line dictionaries and blogs is maintained by the system's administrator.

**Disadvantages of the web-based Corpus**

Corpus gathered from the web is not always reliable and Google cannot provide information that linguists would like to have. In addition,

- Users have no control over the content,
- Web sites are full of metadata that are unnecessary for corpus building,
- We need to focus on pages in a specific language only,
- Specific jargon is used in web blogs and chat rooms.

In spite of the above, we believe that the advantages of providing the user with context information outweigh the drawbacks. Searching for context information via Google is exactly what most translators groping in the dark would do in the first place.

**Context information prepared ahead – the Database Corpus**

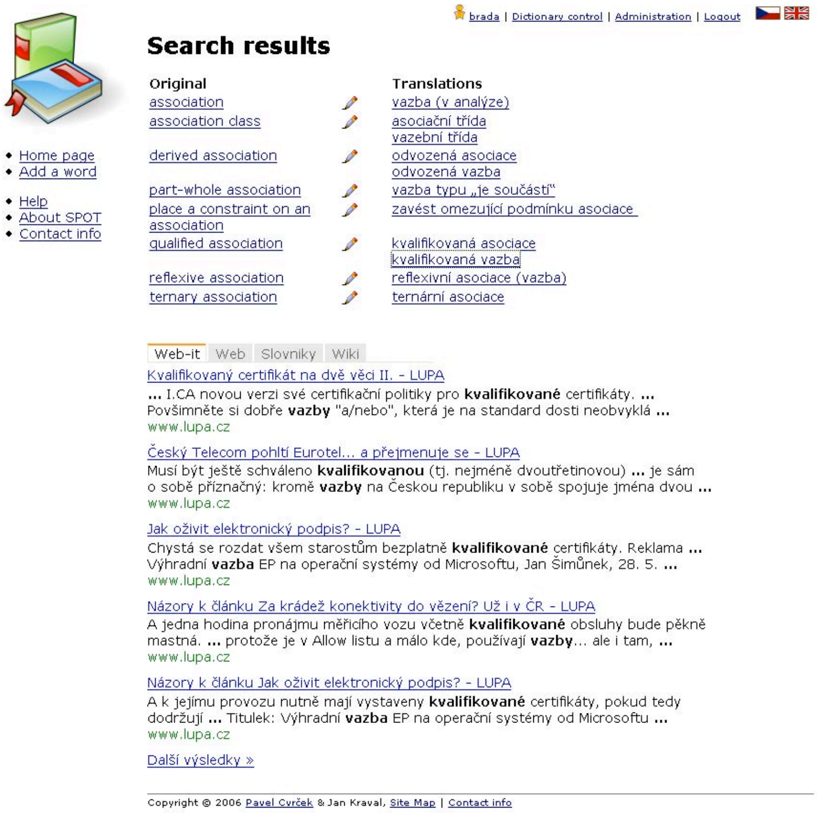
Initially, we were planning to build the corpus manually by collecting context information from the web, selecting the best samples manually and storing the resulting corpus in a text database. We were attracted by the design of the WebBootCaT (see http://corpora.fi.muni.cz/bootcat/ or http://sslmit.unibo.it/~baroni/), which is a tool for building domain-specific corpora to support translators [7].

For more information on web corpora building, see also KWiC Finder (www.kwicfinder.com) ("Web as Corpus"), WebCorp (www.webcorp.org.uk), or Web as Corpus Toolkit (www.drni.de/wac-tk/).

There are two reasons why we rejected the idea of corpus building within a database:

1) SPOT users are on-line and it is unlikely that corpus building on the fly would cause any time delays; context information gathered via Google search API is presented instantly; using a database involves additional overhead plus necessitates database maintenance.

2) ICT terminology evolves very quickly and new terms are created on virtually a daily basis; it would be prohibitive to maintain database corpus up-to-date for this field of specialty.



Search results

| Original | | Translations |
|---|---|---|
| association | ✎ | vazba (v analýze) |
| association class | ✎ | asociační třída<br>vazební třída |
| derived association | ✎ | odvozená asociace<br>odvozená vazba |
| part-whole association | ✎ | vazba typu „je součástí" |
| place a constraint on an association | ✎ | zavést omezující podmínku asociace |
| qualified association | ✎ | kvalifikovaná asociace<br>kvalifikovaná vazba |
| reflexive association | ✎ | reflexivní asociace (vazba) |
| ternary association | ✎ | ternární asociace |

Kvalifikovaný certifikát na dvě věci II. - LUPA
... I.CA novou verzi své certifikační politiky pro **kvalifikované** certifikáty. ... Povšimněte si dobře **vazby** "a/nebo", která je na standard dosti neobvyklá ...
www.lupa.cz

Český Telecom pohltí Eurotel... a přejmenuje se - LUPA
Musí být ještě schváleno **kvalifikovanou** (tj. nejméně dvoutřetinovou) ... je sám o sobě příznačný: kromě **vazby** na Českou republiku v sobě spojuje jména dvou ...
www.lupa.cz

Jak oživit elektronický podpis? - LUPA
Chystá se rozdat všem starostům bezplatně **kvalifikované** certifikáty. Reklama ... Výhradní **vazba** EP na operační systémy od Microsoftu, Jan Šimůnek, 28. 5. ...
www.lupa.cz

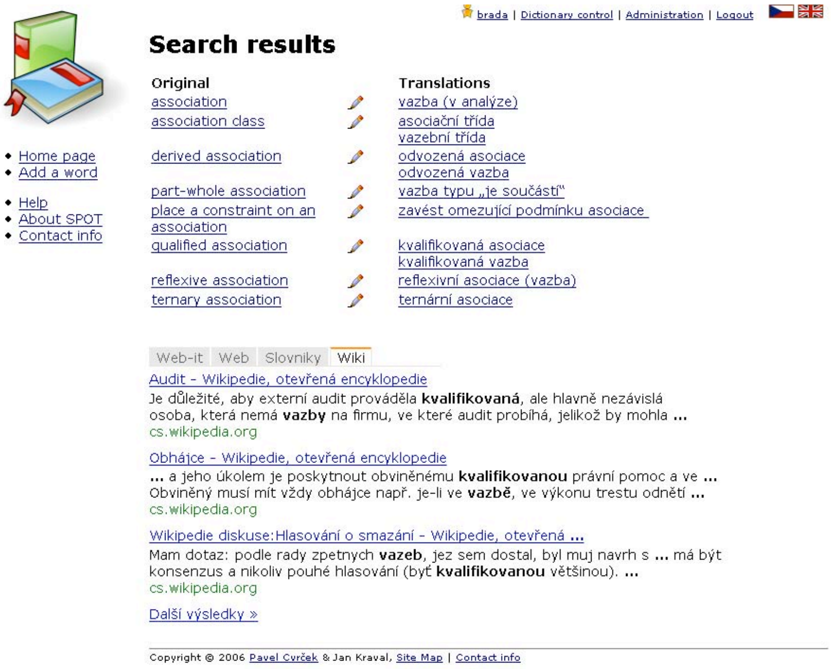
Názory k článku Za krádež konektivity do vězení? Už i v ČR - LUPA
A jedna hodina pronájmu měřícího vozu včetně **kvalifikované** obsluhy bude pěkně mastná. ... protože je v Allow listu a málo kde, používají **vazby**... ale i tam, ...
www.lupa.cz

Názory k článku Jak oživit elektronický podpis? - LUPA
A k jejímu provozu nutně mají vystaveny **kvalifikované** certifikáty, pokud tedy dodržují ... Titulek: Výhradní **vazba** EP na operační systémy od Microsoftu ...
www.lupa.cz

Další výsledky »

Copyright © 2006 Pavel Cvrček & Jan Kraval, Site Map | Contact info

**Figure 3: Term translations shown in the context of the web, confined to IT sites only**



Search results

| Original | | Translations |
|---|---|---|
| association | ✎ | vazba (v analýze) |
| association class | ✎ | asociační třída<br>vazební třída |
| derived association | ✎ | odvozená asociace<br>odvozená vazba |
| part-whole association | ✎ | vazba typu „je součástí" |
| place a constraint on an association | ✎ | zavést omezující podmínku asociace |
| qualified association | ✎ | kvalifikovaná asociace<br>kvalifikovaná vazba |
| reflexive association | ✎ | reflexivní asociace (vazba) |
| ternary association | ✎ | ternární asociace |

Audit - Wikipedie, otevřená encyklopedie
Je důležité, aby externí audit prováděla **kvalifikovaná**, ale hlavně nezávislá osoba, která nemá **vazby** na firmu, ve které audit probíhá, jelikož by mohla ...
cs.wikipedia.org

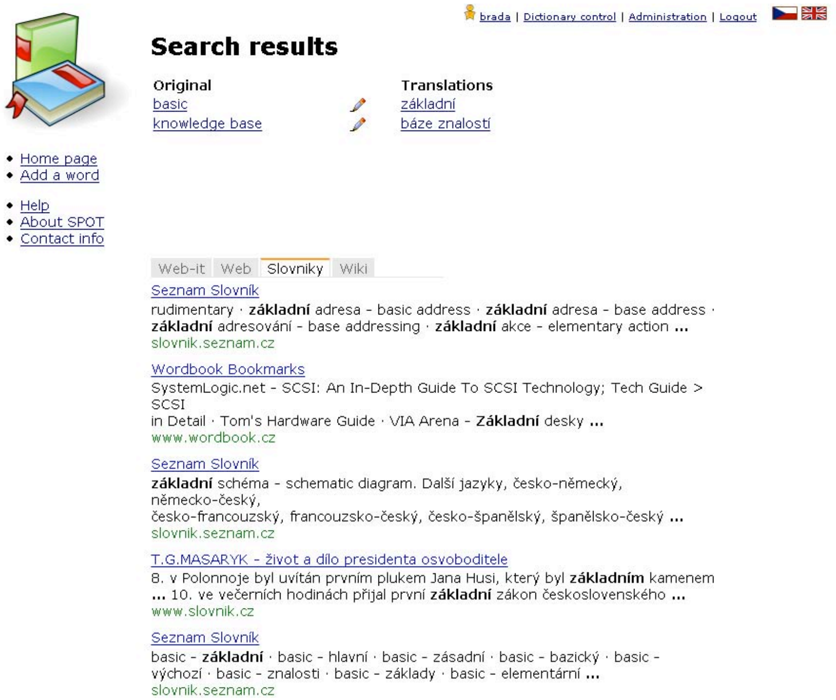
Obhájce - Wikipedie, otevřená encyklopedie
... a jeho úkolem je poskytnout obviněnému **kvalifikovanou** právní pomoc a ve ... Obviněný musí mít vždy obhájce např. je-li ve **vazbě**, ve výkonu trestu odnětí ...
cs.wikipedia.org

Wikipedie diskuse:Hlasování o smazání - Wikipedie, otevřená ...
Mam dotaz: podle rady zpetnych **vazeb**, jez sem dostal, byl muj navrh s ... má být konsenzus a nikoliv pouhé hlasování (byť **kvalifikovanou** většinou). ...
cs.wikipedia.org

Další výsledky »

Copyright © 2006 Pavel Cvrček & Jan Kraval, Site Map | Contact info

**Figure 4: Term translations shown in the context of pre-defined wikis**



Search results

| Original | Translations |
|---|---|
| basic | základní |
| knowledge base | báze znalostí |

Seznam Slovník
rudimentary · **základní** adresa - basic address · **základní** adresa - base address ·
**základní** adresování - base addressing · **základní** akce - elementary action **...**
slovnik.seznam.cz

Wordbook Bookmarks
SystemLogic.net - SCSI: An In-Depth Guide To SCSI Technology; Tech Guide >
SCSI
in Detail · Tom's Hardware Guide · VIA Arena - **Základní** desky **...**
www.wordbook.cz

Seznam Slovník
**základní** schéma - schematic diagram. Další jazyky, česko-německý,
německo-český,
česko-francouzský, francouzsko-český, česko-španělský, španělsko-český **...**
slovnik.seznam.cz

T.G.MASARYK - život a dílo presidenta osvoboditele
8. v Polonnoje byl uvítán prvním plukem Jana Husi, který byl **základním** kamenem
**...** 10. ve večerních hodinách přijal první **základní** zákon československého **...**
www.slovnik.cz

Seznam Slovník
basic - **základní** · basic - hlavní · basic - zásadní · basic - bazický · basic -
výchozí · basic - znalosti · basic - základy · basic - elementární **...**
slovnik.seznam.cz

**Figure 5: Term translations shown in the context of other on-line dictionaries defined by the system's administrator**

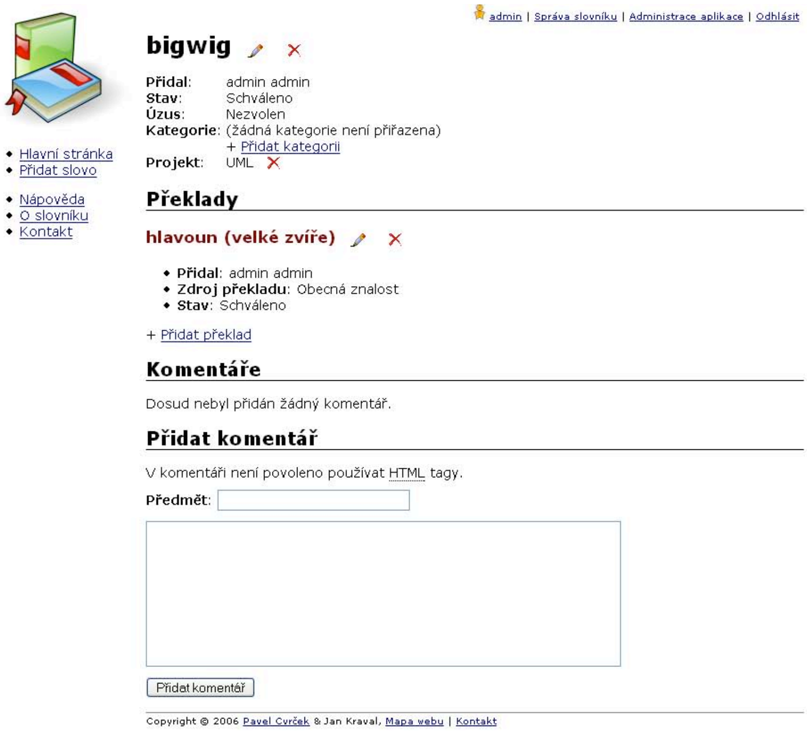
Last but not least, there is a set of *minor features* that can be helpful for SPOT users. Apart from the translation(s) deemed correct, the dictionary can show incorrect or unsuitable translations of a term when such were labelled by the editors. This information can be a valuable guide for translators and learners alike, helping them avoid common mistakes. Also, since registered users are able to vote for the translations, popular and well-accepted Czech equivalents become easily visible.

### 4.1.2 Platform for Terminology "Settling" by Voting

The main advantage of SPOT is that it can act as a platform for the natural development of quality Czech equivalents to the original English terms. With very little effort it can be adapted for any language.

Internet facilitates quick and efficient communication, and it is the inherent property of "collective wisdom" that we are planning to utilize. SPOT will let its users propose Czech translations of unlisted or "unsettled" terms, vote for these translations and discuss them. Based on our long-term observations, users that are likely (and willing) to suggest language equivalents are those that are truly concerned about their form, such as professional translators, professional engineers authoring technical documentation, or academics.

As a necessary complement, the dictionary also includes features for editors to decide upon the final version of individual translations (see Figure 6 below). The final choice on the correct vs. wrong translations will be taken by a small team of editors chosen from the most renowned professionals in the field of ICT, possessing sufficient linguistic knowledge.



admin | Správa slovníku | Administrace aplikace | Odhlásit

- Hlavní stránka
- Přidat slovo

- Nápověda
- O slovníku
- Kontakt

bigwig

Přidal: admin admin
Stav: Schváleno
Úzus: Nezvolen
Kategorie: (žádná kategorie není přiřazena)
+ Přidat kategorii
Projekt: UML

Překlady

hlavoun (velké zvíře)

- Přidal: admin admin
- Zdroj překladu: Obecná znalost
- Stav: Schváleno

+ Přidat překlad

Komentáře

Dosud nebyl přidán žádný komentář.

Přidat komentář

V komentáři není povoleno používat HTML tagy.

Předmět:

Copyright © 2006 Pavel Cvrček & Jan Kraval, Mapa webu | Kontakt

**Figure 6: Editing, updating and commenting on term translations**

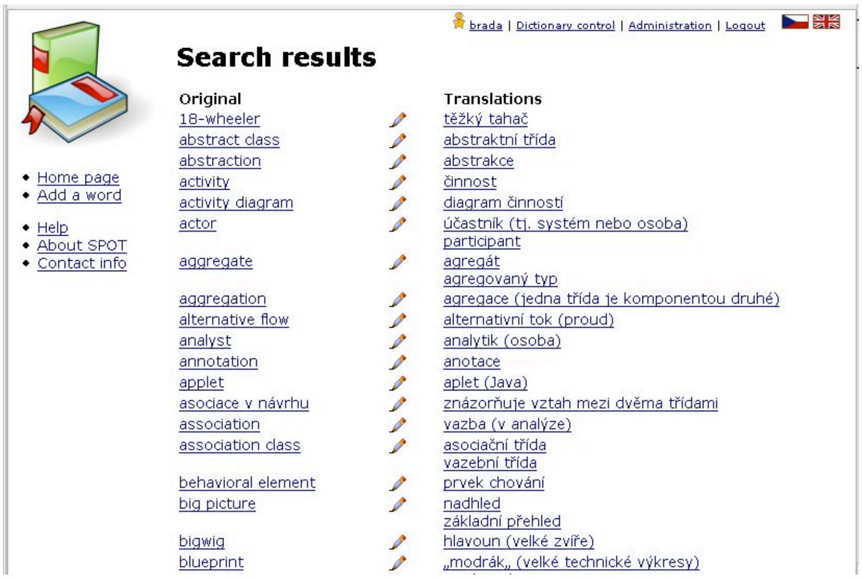
The process of establishing an accepted term translation is as follows:

1. Users add (or import from a CSV file) original terms (in English), assigning them to specific categories;
2. Other users propose additional suitable translations, possibly with references to sources of occurrence, and explanatory comments;
3. Registered users can vote on translations and discuss controversial ones;
4. When the discussion has settled, the editor marks the most suitable translation as "Official", and the remaining versions as either "Usable" or "Unusable / Non-recommended".

Since anyone can become a registered user, SPOT supports the idea that translations may become shared work, and consequently a shared responsibility of the "ICT general public". This is a different approach than the prevailing practice, where a few linguists work on the official localization terms in isolation.

SPOT will also find practical use amongst members of localization teams while creating specialized dictionaries for specific translating projects (see Figure 7). The development of a local corpus proceeds as follows:

1. A special section in SPOT is reserved for the localization team (specific project / customer / product);
2. On-line "settling" of controversial translations (editing, voting) takes place within the team;
3. Terms are propagated instantly into the rest of the dictionary corpus; nonetheless, project-specific terms are designated as such. There is a feature to export this partial corpus to other formats, such as CSV, and to import the terms into the team's Computer Aided Translation tool – CAT.



Search results

| Original | Translations |
|---|---|
| 18-wheeler | těžký tahač |
| abstract class | abstraktní třída |
| abstraction | abstrakce |
| activity | činnost |
| activity diagram | diagram činností |
| actor | účastník (tj. systém nebo osoba)<br>participant |
| aggregate | agregát<br>agregovaný typ |
| aggregation | agregace (jedna třída je komponentou druhé) |
| alternative flow | alternativní tok (proud) |
| analyst | analytik (osoba) |
| annotation | anotace |
| applet | aplet (Java) |
| asociace v návrhu | znázorňuje vztah mezi dvěma třídami |
| association | vazba (v analýze) |
| association class | asociační třída<br>vazební třída |
| behavioral element | prvek chování |
| big picture | nadhled<br>základní přehled |
| bigwig | hlavoun (velké zvíře) |
| blueprint | „modrák„ (velké technické výkresy) |

**Figure 7: Listing terms filtered by a translation project**

The advantages this usage of SPOT brings to the localization teams are manifold, rooted mainly in its on-line implementation:

- Controversial terms are always either on-line, or directly in CAT, rather than stored locally with a localization team member;
- Everyone uses the latest (i.e. currently agreed) version of terms;
- No need to redistribute up-to-date versions of dictionaries.

Below you can see an extract from the MARTIF file used for interchanging newly translated terms among team members (export from Termstar CAT tool while working on a localization project):

```
<langSet lang='eng-us'>
<ntig><termGrp>
  <term>Add a Favorite Place...</term>
  <termNote type='termType'>full form</termNote>
  <termNote type='TS_CreateId'>Jiri Hynek</termNote>
  <date type='origination'>20060917T161007Z</date>
  <termNote type='TS_UpdateId'>Jiri Hynek</termNote>
  <date type='modification'>20060917T161007Z</date>
</termGrp></ntig>

</langSet>

<langSet lang='ces-cz'>
<ntig><termGrp>
  <term>P□idat oblíbené místo</term>
  <termNote type='termType'>full form</termNote>
  <termNote type='TS_CreateId'>Jiri Hynek</termNote>
  <date type='origination'>20060917T161007Z</date>
  <termNote type='TS_UpdateId'>Jiri Hynek</termNote>
  <date type='modification'>20060917T161007Z</date>
</termGrp></ntig>
```

## 4.2 SPOT Implementation – The Current Status

We have currently implemented basic features that include searching, adding new translations, tagging and commenting, dictionary administration by the editorial team, and namely showing translated terms in the context. Under way is the implementation of discussion forums to translations, as well as voting for individual terms suggested by users or editors. Cross-referencing ("see also" links) will be featured in a subsequent release as well. The application is implemented on the Java 5 platform, utilizing the Spring framework and PostgreSQL database system.

The SPOT dictionary can be accessed at http://spot.zcu.cz, where interested readers will also find links to project-related pages. For the data model of the current implementation, see Figure 8 below.

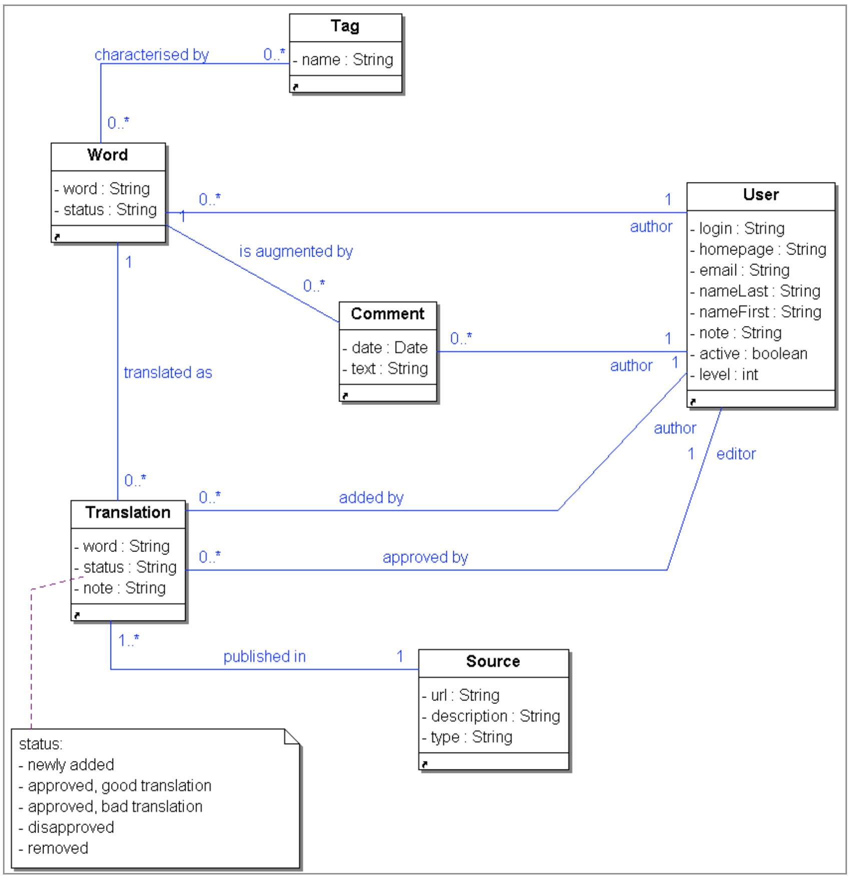


**Figure 8: SPOT data model**

# 5 Conclusion

It is not the task of university teachers or pure linguists to propose suitable equivalents of specialized English terms in their native language. Rather than these people, it should be up to the editors of computer magazines, students, and translators of IT documentation. Terminology must adapt to users, not the other way round.

We believe that by implementing the solution proposed herein, we can not only improve the quality of computer documentation translated into Czech, but also the general culture in this area. Indirectly, this may help to increase readers' preferences for translated publications, while curtailing discrimination against those who either do not have the means to obtain the original books, or the language skills to read them.

Hopefully, SPOT will help in at least the partial standardization of terminology being used, and become a useful source of information for persons involved in the technical or academic writing process.

# Acknowledgements

This work was partly supported by the Ministry of Education of the Czech Republic under Grant No. 2C06009 within the National Program for Research II.

## Notes and References

[1] VIRIUS M. On the quality of Czech translations of technical literature (in Czech: Nad kvalitou českých překladů odborné literatury). Proceedings of OBJEKTY 2002. Praha 2002. ISBN 80-213-0947-4.

[2] MONDSCHEIN P. Oficially dishonoured Czech (in Czech: Oficiálně przněná čeština.) http://games.tiscali.cz/specials/prznenacestina/index.asp. Published 8th September 2004.

[3] Microsoft Terminology Translations, available at: www.microsoft.com/globaldev/tools/MILSGlossary.mspx

[4] The i-spell database, available at: www.lasr.cs.ucla.edu/geoff/tars/

[5] Flickr – Online photo management and sharing application, available at: www.flickr.com/

[6] Stumbleupon – Personalized recommendations of websites, videos, pictures and more. Available at: www.stumbleupon.com/

[7] BARONI, K.; POMIKÁLEK, R. 2006. WebBootCaT: instant domain-specific corpora to support human translators. In: Proceedings of EAMT 2006 - 11th Annual Conference of the European Association for Machine Translation. Oslo, Oslo University (Norway), ISBN 82-7368-294-3.

# Scientific Heritage in Bulgaria Makes First Digital Steps

*Milena Dobreva*[1]; *Nikola Ikonomov*[1,2]

[1] Digitisation of Scientific Heritage Dept., Institute of Mathematics and Informatics
bl. 8, Acad. G. Bonchev St., Sofia 1113 Bulgaria
e-mail: dobreva@math.bas.bg
[2] Phonology and Speech Communications Lab, Institute for Bulgarian Language
Shipchenski Prohod 52, Sofia 1113 Bulgaria
e-mail: nikonomov@ibl.bas.bg

## Abstract

The paper presents recent initiatives in creation, delivery and management of scientific heritage digital resources in Bulgaria. The local and international tendencies will be sketched. Then the work of the Department for Digitization of Scientific Heritage at the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences and more specifically the joint projects with the State Department of Archives and the Central Library of the Bulgarian Academy of Sciences will be described. Finally, we will present a SWOT analysis of the local situation and suggestions for urgently needed actions.

**Keywords:** digitization; BulDML; musical periodicals; DIGMAP

## 1 Introduction

The high expectations about the digital libraries are seen from their listing amongst the three flagship initiatives of the strategy "i2010 – A European Information Society for growth and employment" (the other two areas are caring for people in an ageing society; and the intelligent car). Amongst the basic aims in the joint European efforts are to *avoid duplication*, to *cooperate in networking and standards*, as well as in *developing common and more cost-effective solutions* [1].

Without any doubt, the availability of high quality digital content is in the basis of humanitarian and social research. Most countries from the South-Eastern Europe are still far from moving in line with the European Commission (EC) guidelines in this area. The *EC Recommendation of 24 August 2006 on the digitization and online accessibility of cultural material and digital preservation* [2] emphasizes the importance of setting up of large and sustained digitization facilities, encouraging partnerships between cultural institutions and the private sector, solving the problems around orphan works as well as developing clear quantitative targets for digitization efforts. Furthermore, the *Council Conclusions on the Digitisation and Online Accessibility of Cultural Material and Digital Preservation* (2006/C 297/01) [3] suggest an action plan which can not be followed in countries where there is neither national strategy nor large scale digitisation facilities or recognized competence centres promoting digitization activities.

Recently at the closing of Conference on Scientific Publishing in the European Research Area Access, Dissemination and Preservation in the Digital Age (Brussels, 16 February 2007), Ms. Viviane Reding, EU Commissioner on IS & Media stressed:

> *"... if we do not actively pursue the preservation of digital material now, we risk having* ***a gap in our intellectual record****. If you allow me another historical reference, we do not want to experience the digital equivalent of the destruction of the Alexandria Library. Scientific assets are just too valuable to be put at risk."* [4].

Countries like Bulgaria, which still do not have a national framework for digitization of cultural and scientific heritage, are even in more danger of deeper digital divide. Not only the structured effort to digitize and preserve is missing, but there is yet another danger. The extensive brain drain causes a gap in the community of those who could work on content provision, the experts expected not only to take care of digitisation of cultural and scientific material, but also to place it in the local and wider European context. The experienced researchers which are still active in their profession do not have to whom to transfer their knowledge because local research career is not attractive to the young generation.

## 2 The Current Situation in Bulgaria

As it was already mentioned, the concerns communicated on the top EC level are not exactly matching the local Bulgarian situation. However, common and more cost-effective solutions, cooperation in standards and practical work, and care to avoid duplication should be strictly followed in a country with a population of 7, 2 million which hosts over 5 million cultural and scientific heritage objects objects.

Another important issue of the cultural and scientific heritage institutions nowadays for Bulgaria is the adoption of brand-new IT applications in the sector. Especially new and emerging technologies which are presented in [5] if used at all in Bulgaria appear just in small demonstration projects. Amongst the current problems in Bulgaria, we should mention:

- The absence of a *national strategy*, which leads to lack of co-ordination between separate local initiatives;
- The lack of understanding and practical solutions on importance of such issues as *common quality standards* and *interoperability*;
- The gaps in the local laws and *legislative regulations* related to digitization lead to difficulties for the decision makers in the cultural and scientific heritage sector institutions;
- The need for better *co-operation on regional and European level*, since most of the cultural heritage is one we all share;
- The ambiguity of legal *copyright issues* which leads to serious problems in persuading researchers to share their knowledge in digitization projects affecting the level of presentation of materials, and restricting the depth of presentation. Copyright issues are related to the primary sources on the one hand; on the other hand the issues of legally using the results of research work during digitization are completely unclear.

If we would have to summarize the current situation in Bulgaria, we could draw the following basic conclusions:

- Bulgarian collections are of European importance but they still are not accessible in electronic form;
- Experience exists basically in the pre-digitisation stages of work such as cataloguing, and text encoding, but mass digitisation projects are just about to start in several libraries;
- Digitisation work per se has not been done, thus the country does not match current EC priorities;
- No regular governmental programme (respectively, funding) is available, digitization in Bulgaria strongly depends on external financial support. The Ministry of Education and Science had one stand-alone call for projects on cultural heritage in 2006. Through it several libraries in Bulgaria currently start their own digitisation projects which are not interoperable. Currently, the State Agency for Information and Communication Technology is working on a national strategy for the accelerated development if information society in Bulgaria in 2007-2010 and digitization is included amongst the priority areas as a general topic. This is a positive sign that the field of work becomes officially recognised, but this is not sufficient for the success of the efforts of various institutions;
- Regional cooperation in the field is realistic. SEEDI (South Eastern Europe Digitization Initiative, [6]) is a joint effort to develop awareness about digitization of cultural and scientific heritage in the region along the Lund Principles of the European Union. It is based on the acceptance that researchers and institutions from the region face common problems and share common scientific and cultural heritage, which still cannot be widely accessed in electronic form. The cooperation within SEEDI is bringing together researchers from regional and European centres with similar scientific and practical interest in digitization and by supporting cooperation between them. For that purpose core groups of specialists are created in order to consult, assist, monitor and develop innovative technologies and digitization projects in collaboration with the local cultural and scientific heritage institutions.

Thus, currently there is not only a niche but an urgent need for several interconnected efforts: creating a national framework, boosting wide-scale digitization work, promoting cooperation of local institutions and improving the excellence in the profession.

# 3 The Experience of the Digitization of Scientific Heritage Department

In 2004, the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences created the Digitization of Scientific Heritage department which hosts the first and still unique in the country Bulgarian digitization line for scanning books and archival documents. The department has set the ambitious goal not only to do research work but to serve as a competence centre, providing following basic activities:

- assistance to technology and content providers;
- development of state of the art workflows and best practices in various areas of digitisation of scientific and cultural heritage;
- implementation of new technologies related to the digitization of cultural and scientific heritage:
- organization of specialised trainings;
- active participation at the elaboration of a National strategy for digitization of cultural and scientific heritage;
- contribution to the international cooperation within regional and European initiatives;
- representation of the country at international fora;
- methodological guidance and practical implementation within participation in various projects.

# 4 Current Activities

## 4.1 Cooperation with the General Department of Archives

The Digitisation of Scientific Heritage department cooperates with the General Department of Archives at the Council of Ministers of Bulgaria (GDA) since 2004. GDA is contributing with defining the priorities for selecting materials for digitization; developing the strategy for preparing descriptions and metadata; preparing specification of the search tools and their future improvement. The Digitization of Scientific Heritage department is contributing with providing its know-how for scanning and optical character recognition, workflow choice, digital image processing and with ensuring the necessary equipment and qualified personnel.

The selection and preparation of documents for digitization (single documents, parts of the archival funds and complete archival funds) is based on the holdings of the Regional Unit “State Archive” – Sofia and includes interesting materials, related to the management of Sofia Municipality, the history of Sofia University, the archives of the former Bulgarian communist party, archival funds of the Monarchy Institute, The Parliament, the Council of Ministers, etc. Selected materials contain valuable manuscripts and printed documents, photographs, sketches, geographical maps and rare books, etc.

The joint work has already brought practical outcomes. In March 2005 both institutions released a multimedia disk “Sofia. Religious spaces”, containing items displayed during an exhibition of the same name, and including scanned documents, digital copies of canonical and dogmatic books, and photographs of paintings, ritual clothes and cult objects.

Recently GDA and the Digitization of Scientific Heritage department started another joint project aimed at the electronic publishing of archive documents related to the Temporary Russian Governance which was established after the liberation of Bulgaria and ruled in the period 1878–1879. This collection of documents is being prepared as a combination of digitised images and full text. It will be organised as a semantic web portal. Scientists and general public will benefit from the availability of full-text transcriptions and tools for semantic search of the archival documents, mostly hand-written sources in Bulgarian and Russian.

Another ongoing effort is aimed at building an electronic archive of documents issued by the Bulgarian Ministry of Education in the 40ies and 50ies of the 20th century [7]. The department provides the methodological guidance in this project. The collection of documents is stored in the Archive of the Ministry of the People’s Education within the State Archival Fund of the General Department of Archives. This is a mixed collection which contains quite diverse documents - official documentation which follows specific templates; letters; notes, certificates; photographs; newspapers, etc. The text documents are printed, typewritten or handwritten. The basic aim of this work is to provide access to different users (specialists in education, historians, and the citizens) to the educational documentation of this historical period. The long-term goal is to build a joint collection of such documents from Bulgaria and Greece.

## 4.2 Joint Work with the Scientific Archive of the Bulgarian Academy of Sciences

Both institutions cooperate since 2005. The Scientific Archive of the Bulgarian Academy of Sciences stores precious documents, related to the history of the Academy. A pilot project was initiated in December 2006, aimed at the digitization of personal archives of famous Bulgarian scientists. As a kick-off both institutions selected the archive of Marin Drinov, one of the founders of the Academy of Sciences. It contains valuable documents, letters, personal notes and pictures. All of them were digitized and then prepared for electronic publishing. Thus scientific archives will become easily accessible both for the wide public and the researchers.

## 4.3 Involvement in the Digmap Project

DIGMAP (Discovering our Past World with Digitised Maps) is a project which is supported through the eCONTENT*plus* programme [8]. It proposes to develop solutions for geo-referenced digital libraries, especially focused on historical materials and in the promoting of our cultural scientific heritage. The final results of the project will consist in a set of service available in the Internet, and in open-source software solutions that will be able to be reused in other services. The main service will be a specialized digital library, reusing metadata from European national libraries, to provide discovery and access to contents. Also, relevant metadata from third party sources will be reused, as also descriptions and references to any other relevant external resource. Ultimately, DIGMAP will pursue the purpose to become the main international information source and reference service for old maps and related bibliography. DIGMAP will develop solutions for georeferenced digital libraries, especially focused on historical materials and in the promoting of our cultural and scientific heritage. The final results of the project will consist in a set of services available in the Internet, and in reusable open-source software solutions.

The project will make a proof of concept reusing and enriching the contents from the **National Library of Portugal (BNP)**, the **Royal Library of Belgium (KBR/BRB)**, the **National Library of Italy in Florence (BNCF)**, and the **National Library of Estonia (NLE)**. In a second phase, that will be complemented with contents and references from other libraries, archives and information sources, namely from other European national libraries members of TEL – The European Library [9]. DIGMAP might became an effective service integrated with TEL - in this sense the project is fully aligned with the vision "**European Digital Library**" as expressed in the "i2010 digital libraries" initiative of the European Commission.

The technology will be developed by the Department of Information Systems and Computer Engineering of the **Instituto Superior Técnico — Lisbon (IST)**, in cooperation with the Group MERCATOR of the **Polytechnic University of Madrid (UPM)**. The project started in October 2006, and will have the duration of 24 months. The project coordinator is Prof. José Borbinha, of the IST. The technical work will be co-ordinated by the IST, UPM and KBR. The evaluation of the results will be co-ordinated by the NLE. The liaisons with external entities and advising groups will be coordinated by the BNCF. The dissemination will be co-ordinated by the BNP.

The **Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences (IMI)** will provide assistance to the evaluation, liaison with the Southern East Europe, and to the dissemination. In particular, the Digitisation of Scientific Heritage department already discussed what are the available maps and books relevant to the project in the collections of the General department of archives, Central Library of the Bulgarian Academy of Sciences, and National Library 'Ivan Vazov' in Plovdiv. Electronic records on these objects are not available yet and the participation in this project would be a chance for exposure of local materials.

## 4.4 Digitisation of Bulgarian Mathematical Heritage

One of the principal activities of the department is the digitization of mathematical publications of Bulgarian mathematicians, with the aim to build BulDML – Bulgarian digital mathematical library. Following journals and documents are currently being digitised:

1. Physical-Mathematical Journal (in Bulgarian), 1958 – 1991, 1993;
2. Serdica (articles in different languages): 1975, 1995-2000 г. ;
3. Archive on the development of the Union of Bulgarian Mathematicians: Book with records of board meetings, 1905-1936. In addition to the digital images full texts are also entered. Research on obtaining photographs of the mathematicians whose names appear in the records is being done – this will allow us to offer a complete resource on the dawns of Bulgarian mathematics;

4. Full collection of books and publications of Nikola Obreshkov, a famous Bulgarian mathematician.

The works of the Department are not only limited to mathematical publications. Efforts have been made in order to enlarge the spectrum of digitization activities. As a result additional fields were covered by the Department, thus ensuring more versatility and flexibility:

## 4.5 Digitisation of Historical Musical Periodicals

This work aims to prepare digital copies and descriptions compatible with the Retrospective Index of Musical Periodicals (RIPM) [10]. Currently the following sources has been digitised and described:

- Gusla (1891), printed text, musical fragments, illustrations
- ASO (1934), printed text, illustrations
- Materials from the archive of Stoyan Kenderov (musicologist)

One specific difficulty is the identification of a repository which holds Bulgarian historical musical periodicals. For a variety of reasons, these publications are difficult to find.

## 4.6 Digitisation of Historic Newspapers

This is an effort started recently which aim is to produce collection of digitised old newspapers: Daga, Lampion, Mir, Dnes (selected issues from the period 1890-1930). This activity is a joint project with the Central Library of the Bulgarian Academy of Sciences. The idea is to offer digital images of newspaper pages as a whole, and access to the texts of the separate articles. Photographs which appear in the newspapers will also be included into a collection of images.

## 4.7 Electronic Records of Manuscripts

The department prepared descriptions of Old Bulgarian manuscripts in TEI conformant XML (the total number of these descriptions is 806). Manuscripts are not digitised because what to be done with them is a matter of library policies. However, the detailed description is an important preliminary activity.

## 4.8 Towards The Creation of a National Digitization Network

The “Digitization of Scientific Heritage department” is the initiator of the creation of a nation-wide network of institutions – museums, libraries, archives and research centers, which are intending to start mass digitization, thus avoiding the implementation of scattered non-effective small-scale projects.

There are a number of good examples of joint work with such institutions. We present here briefly only two case studies, which illustrate well the trend of creating synergies in the digitization field.

# 5 SWOT Analysis

The SWOT analysis is used to evaluate the ***S****trengths,* ***W****eaknesses,* ***O****pportunities, and* ***T****hreats* involved in a process or more specifically in a project or in any other situation of an organization or individual requiring a decision in pursuit of an objective. It involves monitoring and analyzing the environment internal and external to the process in question. In order to provide a precise picture of the current situation we have tried to summarize all relevant conditions to the digitization of scientific and cultural heritage in Bulgaria.

| **Strengths** | **Weaknesses** |
|---|---|
| • **Experience already available.**<br>The positive influence is that some institutions already have the feeling what efforts are needed.<br>• **Good contacts with colleagues from the region and other EC countries.**<br>This is important for being in line with the current practices.<br>• **Trainings/specialists meetings done on a regular basis.**<br>The circle of specialists from the community of practice grows although this is quite slow process.<br>• **Established professional bodies.**<br>The existence of departments such as the Digitisation of scientific heritage in IMI is important since it is in contact with many institutions which will play the role of future content providers. | • **Lack of established and working national strategies in the field of digitization, online accessibility and preservation.**<br>This leaves all decisions on specific actions to the institutions which in fact would play the role of content providers. In most cases their ideas and vision on digitisation are quite simplified.<br>• **Strong dependence on external funding.**<br>This is in controversy with the need to set up national priorities. External funding is not reflecting the national vision on importance.<br>• **Scattered experience.**<br>The experience which exists is for small initiatives, not for large projects/programmes. |
| **Opportunities** | **Threats** |
| • **Great amount of work to be done, space for creativity.**<br>Since digitisation and online accessibility are in the beginning here, this gives space for creative approaches and innovative solutions.<br>• **Local specifics may provide interesting cases.**<br>For organisations which seek extension of their activities to Bulgaria, the local cultural materials might be very interesting and enrich their vision on the work which they are doing. | • **Copyright issues.**<br>The unclearness on copyright issues and how they should be approached and solved may create tension for those who do digitisation work.<br>• **Various levels of relevant experience**<br>The vision of museums, libraries and archives differ. These institutions still do not have designated digitisation units.<br>• **Small projects, scattered efforts.**<br>The danger is that small project repeat similar efforts and choose solutions where operability is not guaranteed.<br>• **Lack of crosswalks.**<br>There is no responsible body which would collect data on the standards used in different institutions, respectively there are no crosswalks which could help to build a big shared resource.<br>• **Work in conditions where neither governmental nor institutional policies are well established.**<br>In most cases this will mean that institutions will reinvent the wheel. |

**Table 1: SWOT analysis: digitisation in Bulgaria**

## 6 Conclusion

The analysis reveals the basic problems that put obstacles on the way to the mass digitization in Bulgaria. At the same time it contains as well the potential possibilities to improve the situation. On this basis we have formulated following basic tasks, that in our opinion, will boost the digitization activities in the country, and will put the overall process on European level:

- Creating a joint infrastructure for the key cultural and scientific heritage institutions work;
- Establishing a common methodological network for institutions which take care for different types of heritage;

- Finding common standards for encoding and data interchange for the locally-specific features and workflows assuring quality;
- Overcoming the practice of small scale isolated initiatives and promoting a trend to structured complementary activities;
- Introducing areas such as data protection and integrity and digital curation which are currently not used in the cultural heritage sector in Bulgaria;
- Affecting the training and educational gap in the digital preservation and access field, specialists learn from their own pitfalls, not from structured programs;
- Drawing a “map” of existing resources and expertise – this will facilitate the participation in further EU initiatives.

The cultural heritage which we have inherited from the past is quite rich – over 5 million objects in Bulgaria, comparable to its 7,2 million inhabitants. In the present this heritage is underrepresented in the digital space.
To change this, serious efforts are needed in the future. We wrote this paper with the intention to mark the common lines along which the digitisation of cultural and scientific heritage is developing in Bulgaria, and to contribute to future cooperation and exchange of experiences.
We should not forget that what happens in our country is part of the general development worldwide, which currently seems well manifested as follows:

> *Information technology is now so pervasive and so necessary in our society that we must find ways to effectively manage its costs and its impacts across multiple organizations. The best way to do this is to forge partnerships based on a set of common requirements that individual organizations can refine to meet specific business needs and mission priorities. In terms of implementation, this can take the form of a distributed network where organizations can draw from shared knowledge and leverage a technical infrastructure while operating independently.* [11]

What is said relates completely to the digitisation of cultural and scientific heritage. And yet, there is another issue which we should not forget. The production of digital objects nowadays is a gigantic industry:

> *IDC estimates that the world had 185 exabytes of storage available last year and will have 601 exabytes in 2010. But the amount of stuff generated is expected to jump from 161 exabytes last year to 988 exabytes (closing in on one zettabyte) in 2010.*
> *Chuck Hollis, vice-president of technology alliances at EMC Corp., the data-management company that sponsored the IDC research… said the new report made him wonder whether enough is being done to save the digital data for posterity.*
>
> *"Someone has to make a decision about what to store and what not," Hollis said. "How do we preserve our heritage? Who's responsible for keeping all of this stuff around so our kids can look at it, so historians can look at it? It's not clear."* [12]

Such questions are raised in the countries which are already ahead in the digitisation work and whose heritage is much better exposed in the electronic space, compared to the Bulgarian one. Yet, we still should find better ways to digitise it and make it visible.

## Acknowledgements

We would like to acknowledge the support of the project DIGMAP (Discovering our Past World with Digitised Maps, Programme eContentplus – Project: ECP-2005-CULT-038042)

## Notes and References

[1] The website of i2010 is http://ec.europa.eu/information_society/eeurope/i2010/index_en.htm; see also REDING V. *The role of libraries in the information society*, speech at the CENL Conference, Luxembourg, 29 September 2005. http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/05/566&format=HTML&aged=1&language=EN&guiLanguage=en.

[2] *EC Recommendation of 24 August 2006 on the digitization and online accessibility of cultural material and digital preservation* http://europa.eu.int/information_society/activities/digital_libraries/doc/recommendation/recommendation/en.pdf

[3] Council Conclusions on the Digitisation and Online Accessibility of Cultural Material, and Digital Preservation*, (*2006/C 297/01, Official Journal of the European Union, 7.12.2006, 5 pp. http://eur-lex.europa.eu/LexUriServ/site/en/oj/2006/c_297/c_29720061207en00010005.pdf

[4] REDING V., *Scientific Information In The Digital Age: How Accessible Should Publicly Funded Research Be?*, Closing speech, Conference on Scientific Publishing in the European Research Area Access, Dissemination and Preservation in the Digital Age, Brussels, 16 February 2007. http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/07/90&format=HTML&aged=0&language=EN&guiLanguage=en

[5] ROSS, S; DONNELLY, M.; DOBREVA, M., *Core Technologies for the Cultural and Scientific Heritage Sector* (Technology Watch Report 3), European Commission, ISBN 92-894-5276-5, 2005.

[6] Website of SEEDI, South-Eastern European Digitisation Initiative, http://www.ncd.matf.bg.ac.yu/seedi/

[7] DEVRENI–KOUTSOUKI, A., *Electronic Presentation of Bulgarian Educational Archives: an Ontology-Based Approach*. International Journal Information Theories and Knowledge 2007, 8 pp, (to appear).

[8] DIGMAP project website http://digmap.eu/

[9] The European Library website http://www.theeuropeanlibrary.org/

[10] The Répertoire international de la presse musicale (RIPM) http://www.ripm.org/

[11] PARDO, T. et al., *Building State Government Digital Preservation Partnerships: A Capability Assessment and Planning Toolkit*. Center for Technology in Government, University at Albany, SUNY. 2005, p. 16. http://www.ctg.albany.edu/publications/guides/digital_preservation_partnerships/digital_preservation_partnerships.pdf

[12] http://blogs.warwick.ac.uk/hsirhan/entry/how_much_data/

# Digitisation and Access to Archival Collections: A Case Study of the Sofia Municipal Government (1878-1879)

*Maria Nisheva-Pavlova[1]; Pavel Pavlov[1]; Nikolay Markov[2]; Maya Nedeva[2]*

[1] Faculty of Mathematics and Informatics, ”St. Kliment Ohridski” University of Sofia
and Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
e-mail: marian@fmi.uni-sofia.bg; pavlovp@fmi.uni-sofia.bg
[2] General Department of Archives at the Council of Ministers of Republic of Bulgaria
5, Moskovska Str., 1000 Sofia, Bulgaria
e-mail: sofia@archives.government.bg

## Abstract

The paper presents in brief a project aimed at the development of a methodology and corresponding software tools intended for building of proper environments giving up means for semantics oriented, web-based access to heterogeneous multilingual archival collections. Some widespread international encoding standards for archival description and for representation of structured electronic versions of various kinds of documents have been used. An analysis is made on the applicability of appropriate Semantic web methods and technologies in order to provide versatile, user-friendly access to archival collections based on the semantics of their contents. Some practical results concerning the digitisation of a collection of archival documents from the period of the organization of the Sofia Municipal Government (1878 – 1879) and the development of a website presenting this collection are described in the paper.

**Keywords:** digitisation; metadata encoding; ontology; semantic annotation

## 1 Introduction

Recently Computer Science and information technologies play an important role in numerous successful projects directed to digital preservation of collections of handwritten, typewritten and printed archival documents, photographs etc. which are considered as significant scientific or cultural heritage.

In particular, there is an increasing number of electronic publications of archival collections which are of interest to narrow domain specialists (archivists, historians, linguists etc.) and to the general citizen [1, 2]. However, all these electronic publications give the user access tools oriented to the “standard” archivist’s point of view: it is only possible to browse the full archival structure traditional for the particular country, so the search of documents is very difficult and the given search means are too limited.

The paper presents an ongoing project aimed at the development of a methodology and corresponding software tools intended for building of proper environments giving up means for semantics oriented, web-based access to distributed digitised archival collections. Moreover, we suppose that these collections are heterogeneous, i.e. they may include diverse types of materials (official handwritten, typewritten or printed documents, letters, photographs, newspapers, maps etc.) and the texts of the documents within them may be written in different languages. The practical experiments have been performed on a collection of archival documents from the period of the organization of the Sofia Municipal Government (1878 – 1879).

International encoding standards as well as Semantic web methods and technologies have been used. The main difference with other similar projects is in the exploration of the idea that the usage of proper general-purpose and domain-specific ontologies can minimize the resources necessary for the development of tools for adequate, semantics oriented access to heterogeneous (including distributed) multilingual archival collections. More precisely, the project has the following main objectives:

- To define suitable metadata to accompany digitised documents from archival collections in accordance with the international standards, the Bulgarian traditional experience and the needs of the target groups of users;

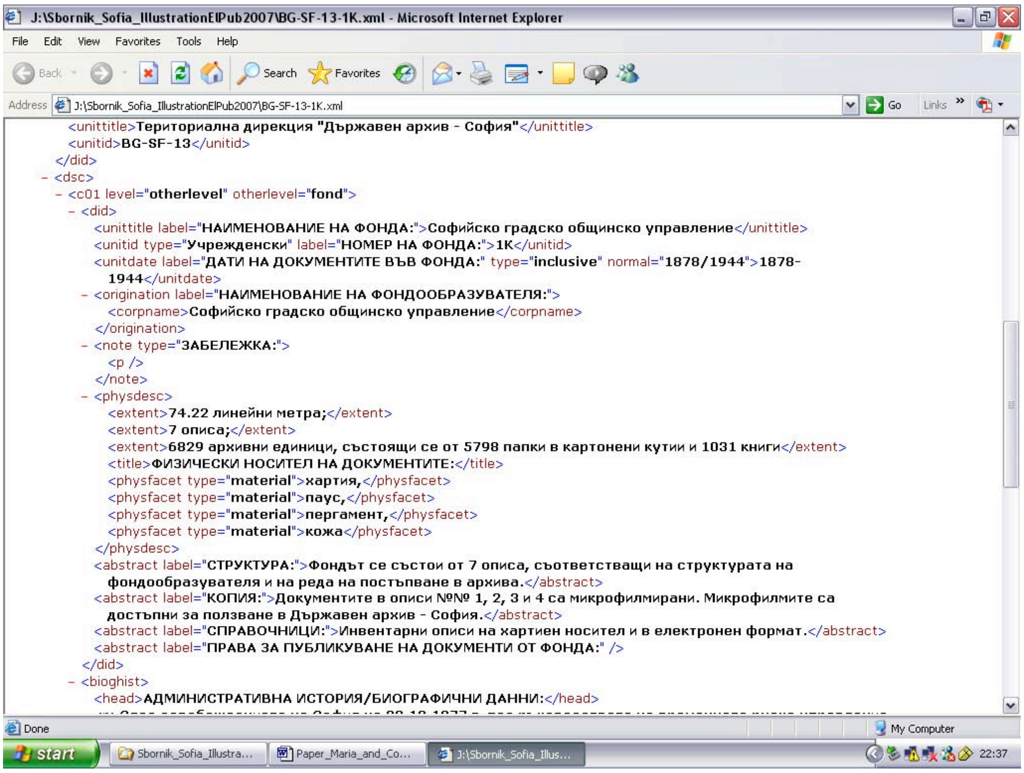
- To study the various aspects of creation of an appropriate ontology for the mentioned collection (e.g. the scope of the ontology, the corresponding linguistic problems etc.);
- To explore the necessities of the typical users of the discussed archival collection (experts in various domains and general public) in order to give proper kinds of access to this collection. In particular, providing versatile, user-friendly access to the collection based on the semantics of its content;
- To develop a framework (that will be intended for users who are professional archivists) for application of Semantic Web methods and technologies to digitised collections of archival documents.

## 2 Representation of the Archival Documents

In this paper we present an ongoing effort aimed at creating an electronic version of an archival collection which consists of approximately 980 original handwritten documents from the period of the establishment of civic authorities of Sofia, building the administrative system, the order and law authorities, communal health services and educational system etc. around and after the end of the Russo-Turkish war (1877 – 1878). This is the period when the building of the fundamentals of the Bulgarian state and municipal institutions has been initiated and the basic rules of the contemporary Bulgarian language have yet to be drawn up. Thus the documents within the collection are of great scientific, historical and social value and are of interest to archivists, historians, linguists etc. Because of these reasons we consider it expedient to include in the electronic version of our collection not only digital images of the chosen archival documents but also structured electronic transcriptions of their full texts and proper descriptions of the collection as a whole as well as descriptions of its parts (known as archival units) and all particular documents in it.

### 2.1 Description of the Structural Parts of the Archival Collection

The discussed descriptions have been prepared in conformity with the traditional practice of Bulgarian archivists. The structure of Bulgarian archives consists of four levels of hierarchy: archival funds, inventory lists, archival units and individual documents. The descriptions at all levels have been structured and accompanied with proper sets of metadata according to the requirements of the EAD encoding scheme [3].



J:\Sbornik_Sofia_IllustrationElPub2007\BG-SF-13-1K.xml - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
J:\Sbornik_Sofia_IllustrationElPub2007\BG-SF-13-1K.xml
Go
Links
Done
My Computer
start
Sbornik_Sofia_Illustra...
Paper_Maria_and_Co...
J:\Sbornik_Sofia_Illus...
22:37

```
      <unittitle>Териториална дирекция "Държавен архив - София"</unittitle>
      <unitid>BG-SF-13</unitid>
    </did>
  - <dsc>
    - <c01 level="otherlevel" otherlevel="fond">
      - <did>
          <unittitle label="НАИМЕНОВАНИЕ НА ФОНДА:">Софийско градско общинско управление</unittitle>
          <unitid type="Учрежденски" label="НОМЕР НА ФОНДА:">1K</unitid>
          <unitdate label="ДАТИ НА ДОКУМЕНТИТЕ ВЪВ ФОНДА:" type="inclusive" normal="1878/1944">1878-
            1944</unitdate>
        - <origination label="НАИМЕНОВАНИЕ НА ФОНДООБРАЗУВАТЕЛЯ:">
            <corpname>Софийско градско общинско управление</corpname>
          </origination>
        - <note type="ЗАБЕЛЕЖКА:">
            <p />
          </note>
        - <physdesc>
            <extent>74.22 линейни метра;</extent>
            <extent>7 описа;</extent>
            <extent>6829 архивни единици, състоящи се от 5798 папки в картонени кутии и 1031 книги</extent>
            <title>ФИЗИЧЕСКИ НОСИТЕЛ НА ДОКУМЕНТИТЕ:</title>
            <physfacet type="material">хартия,</physfacet>
            <physfacet type="material">паус,</physfacet>
            <physfacet type="material">пергамент,</physfacet>
            <physfacet type="material">кожа</physfacet>
          </physdesc>
          <abstract label="СТРУКТУРА:">Фондът се състои от 7 описа, съответстващи на структурата на
            фондообразувателя и на реда на постъпване в архива.</abstract>
          <abstract label="КОПИЯ:">Документите в описи №№ 1, 2, 3 и 4 са микрофилмирани. Микрофилмите са
            достъпни за ползване в Държавен архив - София.</abstract>
          <abstract label="СПРАВОЧНИЦИ:">Инвентарни описи на хартиен носител и в електронен формат.</abstract>
          <abstract label="ПРАВА ЗА ПУБЛИКУВАНЕ НА ДОКУМЕНТИ ОТ ФОНДА:" />
        </did>
      - <bioghist>
          <head>АДМИНИСТРАТИВНА ИСТОРИЯ/БИОГРАФИЧНИ ДАННИ:</head>
```

**Figure 1: Part of the description of archival fund 1K according to EAD standard**

EAD (Encoded Archival Description) is an encoding standard for archival description created and used by archivists to structure and exchange electronic records containing metadata about archival collections. EAD provides a proper framework for seeing particular archive documents in relation to the whole archive collection. Through the use of multiple levels of description the collection can first be described as a whole and then as smaller parts, which get more specific at each level, until at the lowest level the individual archive documents are described.

For example, the EAD – compliant description of an archival fund contains data about the type of the fund, the dates (starting and final years) of creation of its documents, its logical structure and physical extent, the genre(s) and language(s) of its documents, the substances, technologies and methods of creation of documents and other materials in it as well as some short information about the administrative history of the corresponding corporate body, the history of the fund etc. Fig. 1 shows a part of the description of archival fund 1K (part of which is the discussed collection) according to EAD standard.

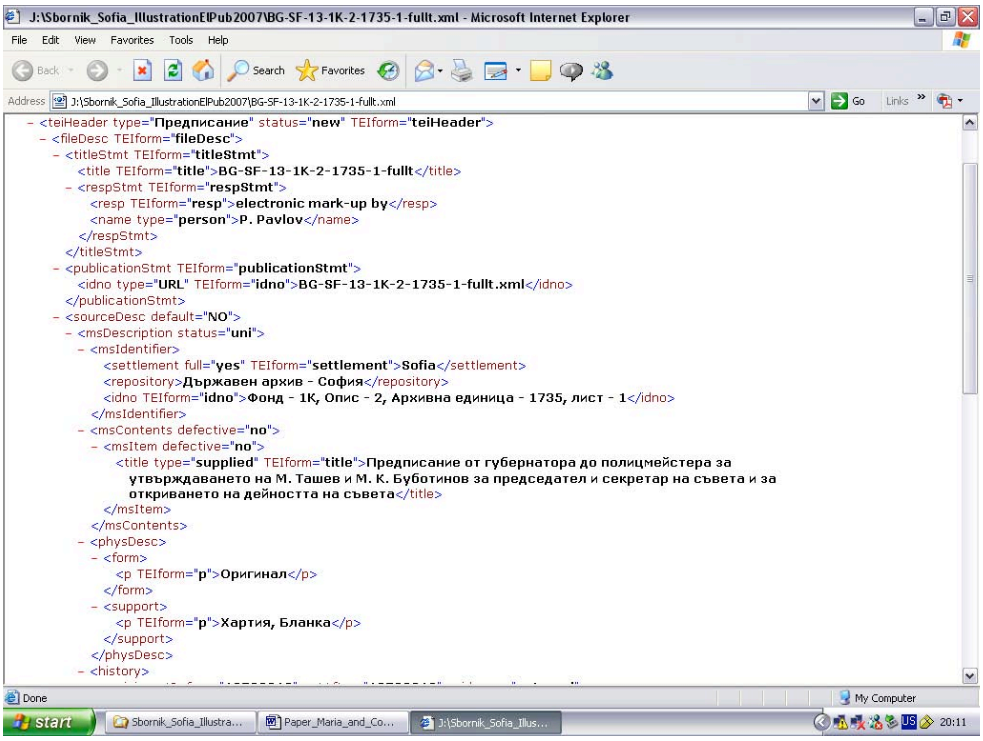
## 2.2 Representation of Electronic Transcriptions of Full Texts of Archival Documents

As it was already mentioned, we maintain two different digital forms of each original archive document: its digital image (in PDF format since this is the most convenient way to have exactly one file containing the image of each particular document independently from the number of the pages of the document) and an electronic transcription of its full text (in XML format). The digital images of the original documents are intended mainly for visualization purposes while the electronic transcripts of the documents and their EAD encoded descriptions will be used to support various types of search and document retrieval activities. For the representation of the structured electronic transcriptions of the full texts of archival documents we use the TEI standard [4].

The Text Encoding Initiative (TEI) may be considered as an established standard for encoding of structured electronic versions of various kinds of documents. The TEI is a flexible encoding framework for electronic documents, allowing the content of the documents to be presented to users in a variety of ways.

We explored the structure and the contents of the various kinds of documents within the collection (instructions, orders, reports, records of sessions, letters, requests, petitions etc.) and created a generalized model of these documents. A proper set of elements and attributes from the TEI document type definition was adopted to describe this model.

The *type* attribute of <teiHeader> defines the class of the corresponding document: instruction or order (“предписание” in the “official” Bulgarian language typical for that historical period, as shown in fig. 2), record of session, request, petition etc.


J:\Sbornik_Sofia_IllustrationElPub2007\BG-SF-13-1K-2-1735-1-fullt.xml - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
J:\Sbornik_Sofia_IllustrationElPub2007\BG-SF-13-1K-2-1735-1-fullt.xml
Go
Links
Done
My Computer
start
Sbornik_Sofia_Illustra...
Paper_Maria_and_Co...
J:\Sbornik_Sofia_Illus...
US
20:11

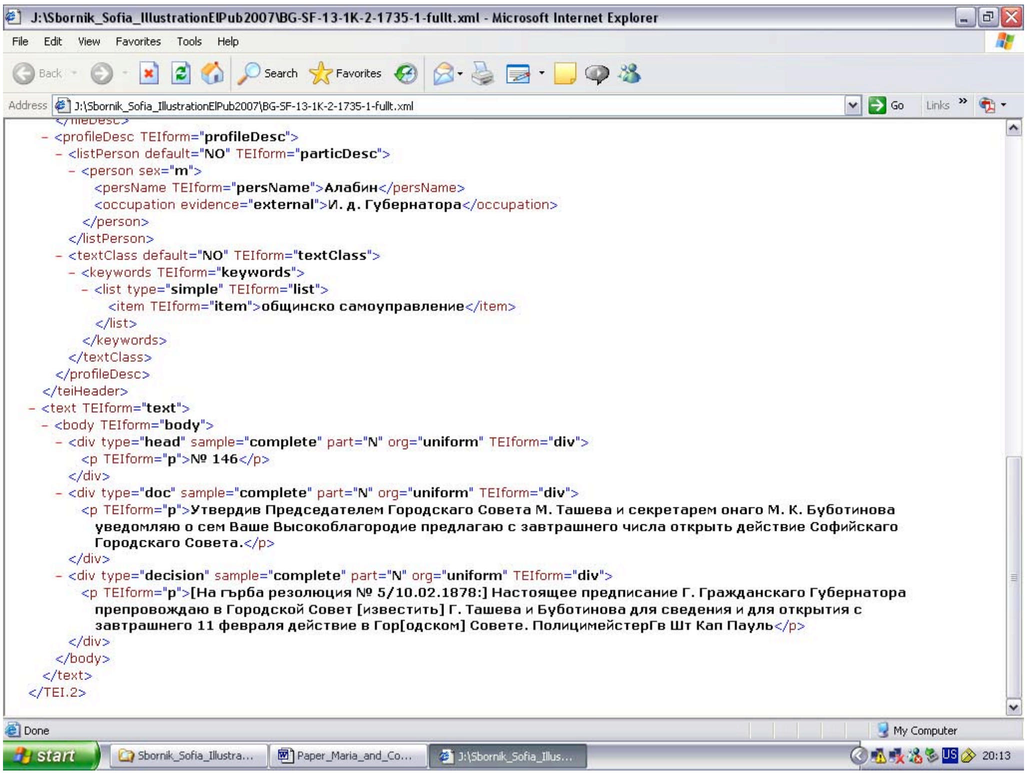
```
- <teiHeader type="Предписание" status="new" TEIform="teiHeader">
  - <fileDesc TEIform="fileDesc">
    - <titleStmt TEIform="titleStmt">
        <title TEIform="title">BG-SF-13-1K-2-1735-1-fullt</title>
      - <respStmt TEIform="respStmt">
          <resp TEIform="resp">electronic mark-up by</resp>
          <name type="person">P. Pavlov</name>
        </respStmt>
      </titleStmt>
    - <publicationStmt TEIform="publicationStmt">
        <idno type="URL" TEIform="idno">BG-SF-13-1K-2-1735-1-fullt.xml</idno>
      </publicationStmt>
    - <sourceDesc default="NO">
      - <msDescription status="uni">
        - <msIdentifier>
            <settlement full="yes" TEIform="settlement">Sofia</settlement>
            <repository>Държавен архив - София</repository>
            <idno TEIform="idno">Фонд - 1К, Опис - 2, Архивна единица - 1735, лист - 1</idno>
          </msIdentifier>
        - <msContents defective="no">
          - <msItem defective="no">
              <title type="supplied" TEIform="title">Предписание от губернатора до полицмейстера за
                утвърждаването на М. Ташев и М. К. Буботинов за председател и секретар на съвета и за
                откриването на дейността на съвета</title>
            </msItem>
          </msContents>
        - <physDesc>
          - <form>
              <p TEIform="p">Оригинал</p>
            </form>
          - <support>
              <p TEIform="p">Хартия, Бланка</p>
            </support>
          </physDesc>
        - <history>
```

**Figure 2: Part of the <teiHeader> element of the electronic transcription of an archival document**

The <fileDesc> element (fig. 2) contains a bibliographic description of the computer file which represents the structured electronic transcription of the document as well as some corresponding information about the document in itself. In particular, the nested elements of <fileDesc> provide information concerning the publication place of the electronic version of the document, the name(s) of the person(s) responsible for the creation of the electronic transcript, the original or supplied name of the document, some physical description of the document etc.

The <profileDesc> element provides a description of the non-bibliographic aspects of the document, e.g. the situation in which it was produced and the persons who contributed to its creation (with their names, positions and ranks). Here we should especially mention the nested element <keywords> which is intended to play the role of a semantic annotation of the corresponding document (fig. 3). This semantic annotation has been used for document retrieval purposes as shown in Section 3.

The <body> element contains the main text of the document divided into separate divisions of different types. Each division consists of a set of paragraphs formatted in the same way as the corresponding paragraphs in the original document. A division of type "*doc*" includes a part of the text of the document in itself. Divisions of type "*decision*" (fig. 3) contain the resolutions made on the document by the corresponding official (e.g. the mayor or the president of the city council). Two other types of divisions ("*execution*" and "*note*") contain some notes concerning the state of the accomplishment of the decisions or orders formulated in the document, some additional instructions to other persons or officials responsible for the execution of the resolutions etc.


J:\Sbornik_Sofia_IllustrationElPub2007\BG-SF-13-1K-2-1735-1-fullt.xml - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
J:\Sbornik_Sofia_IllustrationElPub2007\BG-SF-13-1K-2-1735-1-fullt.xml
Go
Links
Done
My Computer
start
Sbornik_Sofia_Illustra...
Paper_Maria_and_Co...
J:\Sbornik_Sofia_Illus...
US
20:13

```
      </fileDesc>
    - <profileDesc TEIform="profileDesc">
      - <listPerson default="NO" TEIform="particDesc">
        - <person sex="m">
            <persName TEIform="persName">Алабин</persName>
            <occupation evidence="external">И. д. Губернатора</occupation>
          </person>
        </listPerson>
      - <textClass default="NO" TEIform="textClass">
        - <keywords TEIform="keywords">
          - <list type="simple" TEIform="list">
              <item TEIform="item">общинско самоуправление</item>
            </list>
          </keywords>
        </textClass>
      </profileDesc>
    </teiHeader>
  - <text TEIform="text">
    - <body TEIform="body">
      - <div type="head" sample="complete" part="N" org="uniform" TEIform="div">
          <p TEIform="p">№ 146</p>
        </div>
      - <div type="doc" sample="complete" part="N" org="uniform" TEIform="div">
          <p TEIform="p">Утвердив Председателем Городскаго Совета М. Ташева и секретарем онаго М. К. Буботинова
            уведомляю о сем Ваше Высокоблагородие предлагаю с завтрашнего числа открыть действие Софийскаго
            Городскаго Совета.</p>
        </div>
      - <div type="decision" sample="complete" part="N" org="uniform" TEIform="div">
          <p TEIform="p">[На гърба резолюция № 5/10.02.1878:] Настоящее предписание Г. Гражданскаго Губернатора
            препровождаю в Городской Совет [известить] Г. Ташева и Буботинова для сведения и для открытия с
            завтрашнего 11 февраля действие в Гор[одском] Совете. ПолицимейстерГв Шт Кап Пауль</p>
        </div>
      </body>
    </text>
  </TEI.2>
```

**Figure 3: TEI – conformant representation of the text of an archival document**

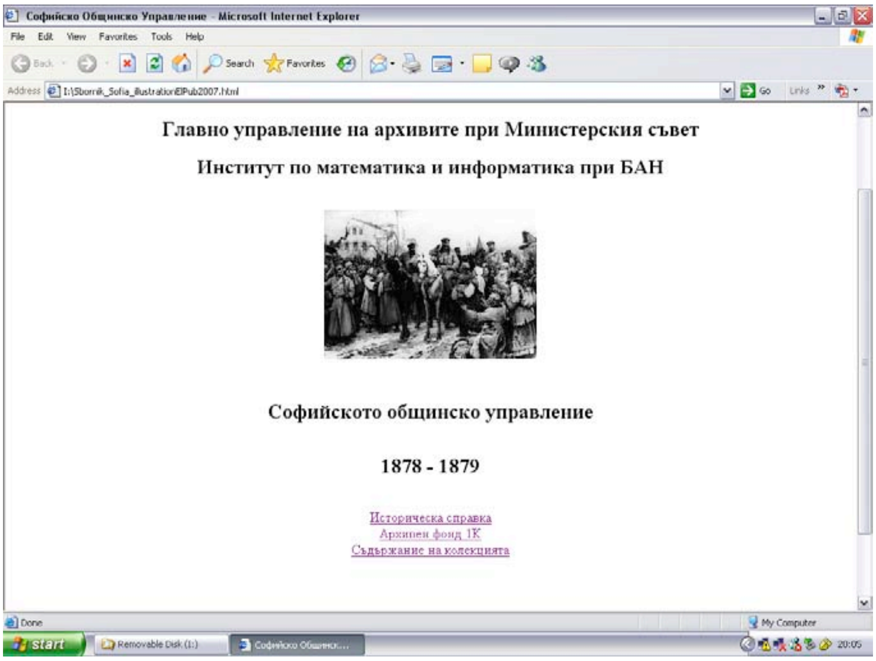
There is a minimal overlap between the metadata held in the EAD encoded descriptions and the TEI encoded document transcriptions. Our experience in the implementation of the project indicates that this overlap causes no serious problems.

## 3 Access to the Collection

The final version of the discussed project will give the user the opportunity to switch between two types of interface to the chosen collection. The first one is based on the principles of the "standard" archivist's view to an archival collection. The second type of provided on-line access to the collection may be described as the semantics oriented one. Fig. 4 shows a screenshot of the current version of the homepage carrying into effect the indicated types of access to the discussed collection.

The interface to the archival collection oriented to the standard archivist's point of view allows the user to browse the hierarchical structure of the collection as a whole (fig. 5). At the archival fund and inventory list levels the user has an access to the EAD encoded description of the corresponding unit (in XML format) and to a properly visualized form of the same metadata (in PDF format).

The user interface at archival unit level allows one to browse five different forms of each particular document in the corresponding archival unit (fig. 6): the EAD encoded description of the document (in XML format), a proper visualization of this description (in PDF format), the TEI encoded electronic transcription of the full text of the document (in XML format), a proper visualization of the electronic transcription of the document (in PDF format) and a digital image of the original document (again in PDF format). Short historical data accompany this type of interface to the collection.


Софийско Общинско Управление - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address I:\Sbornik_Sofia_illustration\ElPub2007.html
Go Links
Done
My Computer
start
Removable Disk (I:)
Софийско Общинск...
20:05

Главно управление на архивите при Министерския съвет

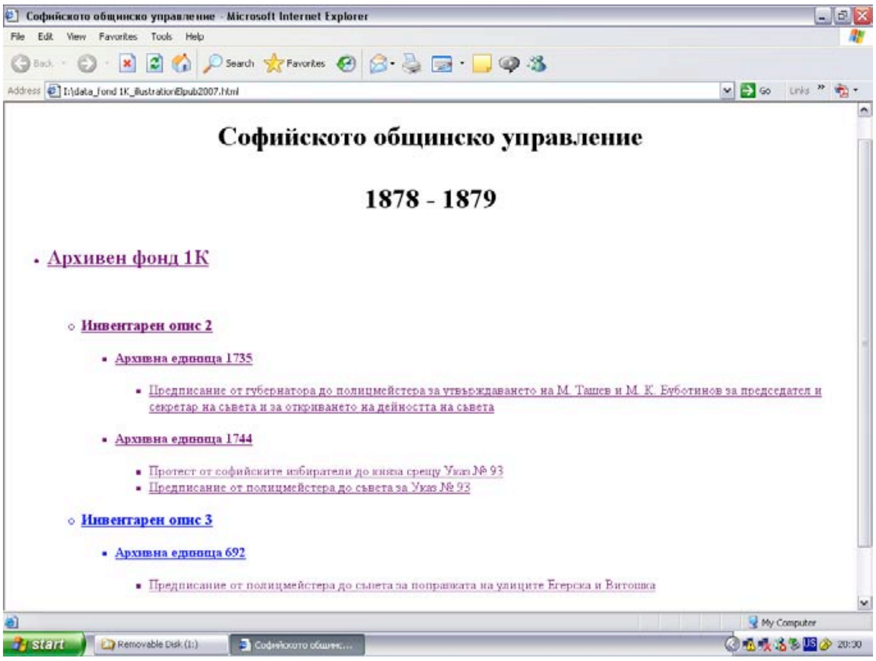
Институт по математика и информатика при БАН

Софийското общинско управление

1878 - 1879

Историческа справка
Архивен фонд 1К
Съдържание на колекцията

**Figure 4: The homepage providing various types of access to the collection**

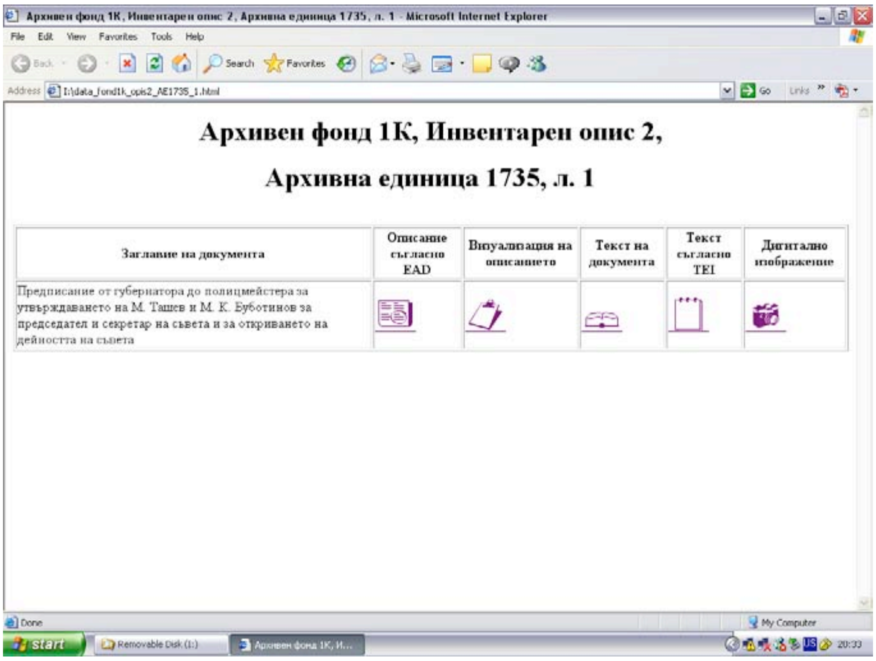

Софийското общинско управление - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address I:\data_fond 1K_illustration\Elpub2007.html
Go Links
My Computer
start
Removable Disk (I:)
Софийското общинс...
20:30

Софийското общинско управление

1878 - 1879

- Архивен фонд 1К
  - ○ Инвентарен опис 2
    - Архивна единица 1735
      - Предписание от губернатора до полицмейстера за утвърждаването на М. Ташев и М. К. Буботинов за председател и секретар на съвета и за откриването на дейността на съвета
    - Архивна единица 1744
      - Протест от софийските избиратели до княза срещу Указ № 93
      - Предписание от полицмейстера до съвета за Указ № 93
  - ○ Инвентарен опис 3
    - Архивна единица 692
      - Предписание от полицмейстера до съвета за поправката на улиците Егерска и Витошка

**Figure 5: Interface to the collection supporting the standard archivist's view (at archival fund level)**


Архивен фонд 1К, Инвентарен опис 2, Архивна единица 1735, л. 1 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address I:\data_fond1k_opis2_AE1735_1.html
Go
Links
Done
My Computer
start
Removable Disk (I:)
Архивен фонд 1К, И...
20:33

Архивен фонд 1К, Инвентарен опис 2,

Архивна единица 1735, л. 1

| Заглавие на документа | Описание съгласно EAD | Визуализация на описанието | Текст на документа | Текст съгласно TEI | Дигитално изображение |
|---|---|---|---|---|---|
| Предписание от губернатора до полицмейстера за утвърждаването на М. Ташев и М. К. Буботинов за председател и секретар на съвета и за откриването на дейността на съвета | | | | | |

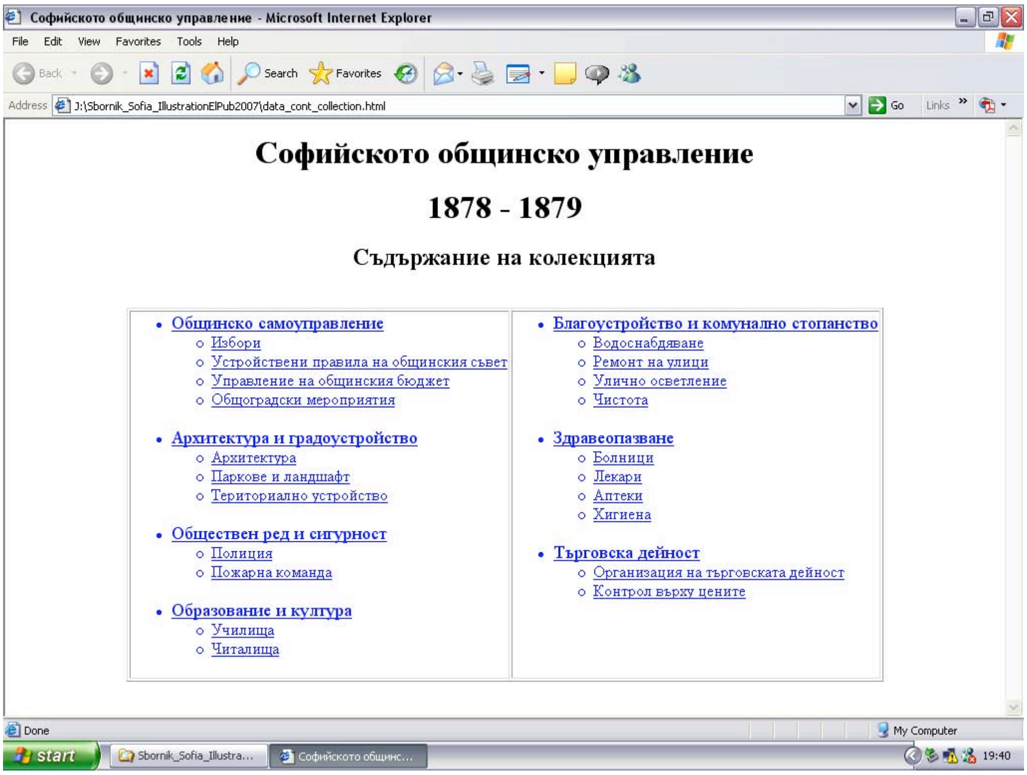
**Figure 6: Interface to the collection supporting the standard archivist's view (at archival unit level)**

The other type of provided access to the discussed archival collection is based on the use of explicitly represented knowledge describing different aspects of the semantics of the collection as a whole and its structural parts. A set of access tools (often called "finding aids") realizing various types of document search and retrieval (chronological, oriented to the kinds of documents within the collection, subject oriented etc.) has been under development for the purpose. The search engines of most of these tools use the values of the corresponding elements of the TEI encoded versions of archival documents. In particular, the subject oriented document retrieval is based on the use of the semantic annotation of the documents. The semantic annotation consists of appropriate words and phrases (chosen from a subject ontology) which describe the content of the document.

Recent Artificial Intelligence textbooks define an ontology as "a shared and common understanding of some domain that can be communicated across people and computers" [5-7]. According to [5], "an ontology can be defined as a formal, explicit specification of a shared conceptualization". Ontologies can therefore be shared and reused among different applications. Moreover, there are at least five serious reasons to create ontologies [8]:

- to share common understanding of the structure of information among people and software agents;
- to enable reuse of domain knowledge;
- to make domain assumptions explicit;
- to separate domain knowledge from the operational knowledge;
- to analyze domain knowledge.

The development of ontologies is still a difficult task, because so far there are no common platforms and verified methods which would prescribe what procedures should be followed in the process of creating an ontology. Nevertheless, there are some reasons to expect that the situation may change in the near future. First, one can find some well-defined principles for design and implementation of ontologies [5]. Second, there is a number of libraries containing already created ontologies (see e.g. [9]) and some of them could be used in the development of new domain-oriented ontologies as examples of good practice. In any case the existence of proper subject ontologies may significantly increase the effectiveness of the implementation of semantics oriented access tools.


Софийското общинско управление - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address J:\Sbornik_Sofia_IllustrationElPub2007\data_cont_collection.html
Go Links
Done
My Computer
start
Sbornik_Sofia_Illustra...
Софийското общинс...
19:40

Софийското общинско управление

1878 - 1879

Съдържание на колекцията

- Общинско самоуправление
  - Избори
  - Устройствени правила на общинския съвет
  - Управление на общинския бюджет
  - Общоградски мероприятия
- Архитектура и градоустройство
  - Архитектура
  - Паркове и ландшафт
  - Териториално устройство
- Обществен ред и сигурност
  - Полиция
  - Пожарна команда
- Образование и култура
  - Училища
  - Читалища
- Благоустройство и комунално стопанство
  - Водоснабдяване
  - Ремонт на улици
  - Улично осветление
  - Чистота
- Здравеопазване
  - Болници
  - Лекари
  - Аптеки
  - Хигиена
- Търговска дейност
  - Организация на търговската дейност
  - Контрол върху цените

**Figure 7: Interface to the collection supporting the subject oriented document retrieval**

In the discussed project we use a subject ontology (covering the main types of municipal activities) especially developed for the purpose. This ontology is prepared using Protégé/OWL [2, 10].

Fig. 7 shows a screenshot presenting the web page which supports the subject oriented document retrieval in the current version of our project. The topics viewed on the screen belong to a subset of the concepts at the highest two levels of the mentioned ontology based on the assumption for typical requests for information according to the characteristics of the discussed historical period. Our future plans include some ideas to generate automatically the list of searchable topics using the results of the preliminary examination of the professional needs of the main groups of potential users.

On the other hand, the semantic annotations of the documents within the collection contain proper terms from all levels of the same ontology. When the user chooses a topic from the list shown on fig. 7, the corresponding access tool finds all documents which contain in their semantic annotations terms matching the user query (i.e. identical to the term chosen by the user or semantically related with it).

A tool for search in the full texts of the document transcriptions is provided as well. We intend to use in its implementation some our former results concerning the development of tools for knowledge based (ontology driven) search in collections of digitised manuscripts [11]. The main idea here is similar to but more sophisticated than the one discussed above. When the user defines his query, the search engine augments it by words and phrases semantically related to these used in the original query (according to a set of available proper ontologies). Then the obtained new query is augmented once more using some synonyms of the main terms and the corresponding terms in Russian or French language (depending on the language in which each particular document is written) from a set of appropriate dictionaries. The final form of the query is processed in a standard way. As a result of the user query processing, the texts of all documents in the collection containing words or phrases semantically related to the one given by the user are properly visualized. The discovered parts of the text matching the concept(s) given as a user query are highlighted.

More complex user queries in the form of conjunctions or disjunctions of “atomic” ones will be processed as well. Some ideas already implemented in our former work [11] will be used for the purpose.

# 4 Conclusion

In this paper we presented a work in progress directed to the exploration of some open questions concerning the development of proper mechanisms and tools providing adequate web-based access to digitised archival collections. The most valuable expected results of our project could be formulated as follows:

- A methodology for application of international standards, ontological knowledge and Semantic web technologies for the development of software tools providing semantics oriented access to heterogeneous multilingual collections of archival documents;
- A model and a prototype of a website which gives the users an interface supporting various types of access to a chosen archival collection.

The analysis of the current results of the implementation of the project gives us a reason to believe that its final version will be compatible with and even more sophisticated in certain aspects than some popular projects like the BAMCO site [1], the LEADERS project [12] etc. The main advantage of our approach is the proper use of ontological knowledge describing the semantics of the individual documents in the archival collection as well as the semantics of the collection as a whole and the semantics of its structural parts. It allows users with different profiles to study and analyze the documents within the corresponding collection from multiple points of view using a single environment for the purpose.

# Acknowledgements

This work has been funded by the EC FP6 Project "Knowledge Transfer for Digitisation of Cultural and Scientific Heritage in Bulgaria" (KT-DigiCULT-BG) coordinated by the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences. The authors are thankful to Dr. Matthew Driscoll from Copenhagen University, Denmark, for the useful advices concerning the TEI encoding of the electronic copies of archival documents.

# References

[1] Brown Archival & Manuscript Collections Online (BAMCO) site. http://dl.lib.brown.edu/bamco/ introcontent.html, last accessed on April 4, 2007.

[2] KNUBLAUCH, H.; FERGERSON, R.; NOY, N; MUSEN, M. The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. Third International Semantic Web Conference, Hiroshima, Japan, 2004.

[3] The Encoded Archival Description (EAD). http://www.loc.gov/ead/, last accessed on April 4, 2007.

[4] The Text Encoding Initiative (TEI). http://www.tei-c.org/, last accessed on April 4, 2007.

[5] Gruber, T. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human-Computer Studies, Vol. 43, 1995, pp. 907-928.

[6] GUARINO, N. Formal Ontology, Conceptual Analysis and Knowledge Representation. International Journal of Human-Computer Studies, Vol. 43, 1995, pp. 625-640.

[7] BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. Scientific American, May 2001, pp. 35-43. http://www.w3.org/2001/sw/, last accessed on April 4, 2007.

[8] NOY, N.; MCGUINNESS, D. Ontology Development 101: A Guide to Creating Your First Ontology. http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html, last accessed on April 4, 2007.

[9] Protégé Ontologies Library. http://protege.cim3.net/cgi-bin/wiki.pl?ProtegeOntologiesLibrary, last accessed on April 4, 2007.

[10] OWL Web Ontology Language Overview. W3C Recommendation, 10 February 2004. http://www.w3.org/TR/owl-features/, last accessed on April 4, 2007.

[11] PAVLOV, P.; NISHEVA-PAVLOVA, M. Knowledge-based Search in Collections of Digitized Manuscripts: First Results. Proceedings of the 10th ICCC International Conference on Electronic Publishing (Bansko, 14-16 June 2006), FOI-COMMERCE, Sofia, 2006, pp. 27-35.

[12] LEADERS: Linking EAD to Electronically Retrievable Sources. http://www.ucl.ac.uk/leaders-project/, last accessed on April 4, 2007.

# The Digital Scholar's Workbench

*Ian Barnes*

Department of Computer Science, The Australian National University, Canberra, ACT Australia
Ian.Barnes@anu.edu.au

## Abstract

In this paper I present the reasoning behind the development of a new end-to-end publishing system for academic writers. The story starts with investigating digital preservation of word processing documents. What file formats are suitable for long-term preservation of text? I believe that the answer is a high-quality structured XML format like DocBook XML or TEI. The next question is how do we get word processing documents into that format without incurring a prohibitive cost? Conversion is possible, but it requires human intervention at some point. It would be far too expensive to have archivists editing every document by hand on ingest, so how can we get authors to do the necessary work, particularly as most academics aren't at all interested in digital preservation of their work? The answer I propose is to offer them more than just an archiving solution. Instead of just getting preservation, they get a full end-to-end digital publishing solution, the digital scholar's workbench, tailored to their needs for document interoperability, collaboration and publication in multiple formats… oh, and they get preservation too.

**Keywords:** word processing; scholarly communication; digital preservation; XML; electronic publishing

## 1 Introduction

Word processing documents are a problem for digital repositories. They are not suitable for long-term storage, so they need to be converted into an archival format for preservation. This is not just a technical issue, but a people issue also. Most university repositories struggle to get academics to deposit their work: making the process more difficult will only make this problem worse. On the other hand, filling a repository with word processor documents that nobody will be able to read in a few years time, is a waste of time, effort and money. In this paper I will address the following questions:

1. What file formats are suitable for long-term storage of word processed text documents?
2. How can we convert documents into a suitable archival format?
3. How can we get authors to convert and deposit their work?

The answer I will propose is an end-to-end digital publishing system that allows authors to continue writing in a word processor, much as they do now, but provides an attractive range of file format conversion and publishing services: the Digital Scholar's Workbench.

While the vast majority of material generated by universities is text, most research on digital preservation concentrates on images, sound recordings, video and multimedia. You could be forgiven for thinking that this is because text is simple, but unfortunately that's not so. Even relatively short text documents (like this one) have complex structure consisting of sections (parts, chapters, subsections etc) and also of indented structures like lists and blockquotes. A significant part of the meaning is lost if that structure is ignored (for example by saving as plain text). In Section 2 I will briefly review some of the previous work in this area.

Most text documents created today are created in a word processor. In Section 3 I will discuss file formats, and give a tentative answer to Question 1 above. The file formats generated by word processors are generally not sustainable, so we need to consider converting documents to better formats. Many archives have chosen PDF, but this has serious problems. XML is a better answer, but it's not a complete answer as XML is not a file format, but a meta-format, a framework for creating file formats. We have to choose a suitable XML file format for storing documents.

In Section 4 I will discuss possible methods available for converting word processing documents into a suitable XML format, addressing Question 2 above.

In Section 5 I give a description of my own current work in progress, the Digital Scholar's Workbench, a web application designed to solve some of the problems with preservation and interoperability of word processing documents.

# 2 Previous Work

There is a lot of published research on digital preservation, but not much of it that I found deals in any detail with preservation of *text*. The DiVA people in Uppsala University Library are archiving documents in XML [1]. They use a custom format which is basically DocBook XML for describing the document itself (content and structure), with a wrapper around the outside allowing for collections of related documents and for comprehensive metadata. At ANU we are considering something similar for large or complex documents, using RDF to describe the relationships between the parts. Slats [2] discusses requirements for preservation of text documents, and the relative merits of XML and PDF. Like several other authors with similar publications, she recommends storing documents in XML, but fails to specify *what* XML format to choose. Anderson et al [3] from Stanford recommend ensuring that documents are created in a sustainable format rather than attempting conversion and preservation later, as I will recommend below. This leaves open the question of what to do with existing documents.

# 3 File Formats

## 3.1 Preservation Formats Versus Access Formats

A *preservation format* is one suitable for storing a document in an electronic archive for a long period. An *access format* is one suitable for viewing a document or doing something with it. Note that it may well be the case that no-one ever views the document in its preservation format. Instead, the archive provides on-the-fly conversion into one or more access formats when someone asks for it. For example, the strategy I recommend below is to store DocBook XML or TEI, but convert the document to HTML for online viewing or PDF for printing. Some file formats may be suitable for both purposes. XHTML has been suggested, with CSS for display formatting. As XHTML is XML (and particularly if the markup is made rich with use of the `<div>` element to indicate structure), it may be an adequate preservation format, at least for simple documents. As it can be viewed directly in a web browser, it is eminently suitable as an access format.

## 3.2 Criteria for Sustainability

What features does a good preservation format have? How do we judge? Lesk [4] gives a list of required features for preservation formats (The points in italics are his, the comments that follow are mine.):

1. *Content-level, not presentation-level descriptions.* In other words, structural markup, not formatting.
2. *Ample comment space.* Formats that allow rich metadata satisfy this requirement.
3. *Open availability.* This means that proprietary formats are not acceptable. Remember what happened to GIF images when Unisys claimed that they were owed royalties because they own the file format [5]. What would happen if Adobe decided to do the same with PDF or Microsoft with Word?
4. *Interpretability.* It should be possible for a human to read the data, and also for small errors in storage or transmission to remain localised. This implies a strong preference for plain text rather than binary file formats. A small error in a compressed binary file can render the entire file useless.

Stanescu [6] looks at this topic from a risk management point of view. Slats [2] discusses criteria for choosing file formats, coming to very similar conclusions.

## 3.3 Word Processing Formats

*Microsoft Word*

The vast majority of all text documents created today are created in Microsoft Word using its native `.doc` format. It would be great if we could just deposit Microsoft Word documents into repositories and be done with it, but unfortunately that won't do, for a few important reasons:

- Word format is owned by Microsoft corporation.
  - They could choose to change the format at any time, possibly forcing repositories to convert all their documents.

  - They could change the licensing at any time.
- Word format is a *binary* format. There is no obvious way to extract the content from a Word document. If the document is corrupted even a little, the content can be lost.
- Word is not just one format but many. Storing documents in Word format would force repositories to support not one but several file formats, or alternatively to engage, every few years, in a process of opening *every* stored document in the latest version of the software, and saving it using the most recent incarnation of the format and fixing any problems. Either way, this is an unacceptable cost.

Microsoft's new Office Open XML `.docx` format [7] is an improvement, but is still unsuitable for archiving. A `.docx` file is a compressed Zip archive of XML files. Compressed files are particularly prone to major loss if corrupted. Also some data is stored as strings that need parsing [8], rather than using XML elements or attributes to separate the different data items. This makes automated processing of these files much more difficult. Microsoft have released the Office Open XML specifications publicly, along with assurances that the format is and will always be free [9]. Despite the mistrust of many in the open source community, who remember the GIF/Unisys controversy [5], this appears to be genuine.

*Open Document Format*

Open Document Format (ODF) [10] is the native file format of OpenOffice.org Writer [11], the word processor component of the OpenOffice.org open source office suite. Open Document Format is an OASIS and ISO standard and a European Commission recommendation. It is also supported by KOffice and AbiWord.

An ODF file is a Zip archive containing several XML files, plus images and other objects. The Zip archiving and compression tool is freely available on all major platforms, so there should never be a problem getting at the content of an ODF document. Using a Zip archive means that the files are prone to catastrophic loss of content with even minor data corruption.

If we are going to archive word processing documents, I believe that ODF is a better option than Microsoft Word format in any of its variations. Even Office Open XML is still a proprietary format.

One possible preservation strategy would be to convert all word processing documents to ODF for storage. This can be done easily using OpenOffice.org itself as a converter. The conversion could be set up as part of the repository ingest process so that it would be almost totally painless for users. Conversion to ODF preserves all the formatting of most Word documents, with only minor differences. For complex documents that use lots of floating text boxes, these minor differences can make a mess of the appearance of the document. For documents that use embedded active content (chunks from live spreadsheets etc), the embedding will probably fail. For most "normal" documents, even complex ones, the conversion is good.

The main disadvantage of this strategy is that Open Document Format is still a word processing format, not a structured document format. What does this mean, and why is it a problem?

- Word processing formats are at heart about describing the appearance of the document, not its structure. For serious processing it's the structure we want. In 20, 50 or 100 years, most readers will probably not care about the size of the paper, the margins, the fonts used and so on. Even today, if we're going to serve up a document as a web page, those details are irrelevant. Sometimes these details can even be a disadvantage, for example if the document insists on fonts that are unavailable on your computer. On the other hand, the division of the document into sections will always be relevant, useful and important, and must be preserved.
- Word processing formats are flat. That is, the document is a sequence of paragraphs and headings. What we'd really like is a deep structure with sections, subsections and so on, nested inside each other (as in DocBook or TEI ). We want this deep structure because it makes structured searches and queries possible, and makes conversion with XSLT much easier.

  It is possible to do automated conversion from flat to deep structure [12], but at the moment only with documents that conform to a well-designed template. We are working to extend this to less carefully prepared documents, but the process is likely to require human supervision.

The other disadvantage of Open Document Format is that even for simple documents it is extremely complex. Unzipping a one-page document of about 120 words results in a collection of files totalling 300K in size. The formatting information is stored in a complex, indirect way. This makes it relatively difficult to locate the meaningful content and structure and transform it into other formats for viewing or other uses. Instead of leaving documents in this complex format and having a hard job writing converters (XSLT stylesheets) for all possible future uses, it would be better to store documents in a simple, clear, well-structured format that makes converters easier to write.

*Other word processing formats*

There are several, but none of them has much market share, nor do any of them have any particularly conspicuous advantages. Probably the best strategy with these is to convert them into Word or Open Document Format, then treat them in the same way as the majority of documents. OpenOffice.org will open many file formats, so it can be used as a generic first stage in any process of converting documents into useful formats. Use OpenOffice.org in server mode to open all documents and save them in Open Document Format, then process them into something better.

## 3.4 PDF

Many repositories have adopted PDF as their main format for text documents, both for storage and for access. PDF has some good points:

- It is easy to create, either using Adobe Acrobat software or using the PDF Export feature available in both Microsoft Word and OpenOffice.org Writer.
- It can be viewed on all platforms using the free Adobe Acrobat Reader software.
- It is extremely effective at preserving the formatting of a document. For some applications (for example in legal contexts) this may be of vital importance.

However, there are some serious problems with using PDF as a storage format [13]:

- The format is proprietary, owned by Adobe. While it is currently open, the company could decide to change this at any time.
- There are some compatibility problems between different versions.
- Documents may rely on system fonts. There is an option in PDF to embed all fonts in the document, but not all software uses this, and some PDF viewing software either cannot locate the correct fonts or doesn't know how to substitute suitable alternatives. Failing to embed all fonts can result in a serious degradation of the on-screen appearance of a document, or in a complete failure to display the content.
- PDF includes extra features like encryption, compression, digital rights management and embedding of objects from other software packages. These all present difficulties for archivists.

PDF is an excellent access format for printing to paper. Any good preservation system should be able to generate PDF renditions of documents for this purpose. PDF is not so good for viewing on screen, as it ties document content to a fixed page size. However, for the reasons given above, it is not a good preservation format.

## 3.5 RTF

RTF stands for Rich Text Format. It is a Microsoft specification [14], but they have published it, so one could argue that it is an open standard. It is certainly widely interoperable, with most word processors capable of reading and writing RTF. There are problems with using RTF as a preservation format:

- It is still proprietary, with all the risks that entails.
- There seem to be parts of the specification that are not in the publicly available specification document, and which have changed over the years.
- The specification is not complete and precise, leaving many little quirks.

The National Library of Australia has chosen RTF as its main preservation format [15]. I think a well-chosen XML file format has significant advantages over RTF, but it might well be worth retaining RTF as an access format, since it has good interoperability, at least for relatively simple documents.

## 3.6 XML

XML [16] is widely accepted as a desirable format for document preservation. See for example the assessment of XML on the US Library of Congress digital formats web site [17] and the related conference paper by Arms & Fleischauer [18]. The main reasons are:

- XML is a free, open standard.
- XML uses standard character encodings, including full support for Unicode. This makes it capable of describing almost anything in any language.

- XML is based on plain text. This gives it the best possible chance of being readable far into the future. Even if XML and XSLT are no longer available, the raw document content and markup will still be human-readable. (This will be true even if the *meaning* of the markup has been lost, although formats designed with preservation in mind should try to make the meaning apparent through the choice of element and attribute names).
- XML can easily be transformed into other formats using XSLT [19].

This last point is very important. It means that documents which are stored in XML can be viewed in multiple formats. A minimal solution would generate HTML for on-screen viewing and PDF for printing.
However, just saying "XML is the answer" isn't enough. XML is only really useful when documents conform to a standard DTD or schema. Having an XML-based preservation strategy means choosing one or more (but preferably very few) XML document formats. It also means having a workable method for converting documents into that format.

*DocBook XML*
DocBook [20] is a rich and mature format that has been in use for about 15 years. It was originally an SGML format designed for marking up computer documentation (particularly the O'Reilly books), but its application is wider, although it still seems a bit awkward and sometimes ill-matched to non-technical writing. DocBook is an OASIS [21] standard.

DocBook is huge, with over 300 elements. This makes it hard to learn, and cumbersome to use directly; few people create DocBook documents by hand. However this is of no concern to the ordinary author if the transformation from word processor formats to DocBook is done automatically. It *may* be a concern for the unlucky person who has to write stylesheets for converting documents to and from DocBook. Fortunately though, Norm Walsh (the guiding force behind DocBook) and others have written a comprehensive set of XSLT stylesheets [22] for converting from DocBook XML into numerous formats including XSL-FO (and hence PDF) and XHTML. This is a huge headstart.

For converting word processor files to DocBook XML, the complexity and number of elements doesn't matter, since the conversion process will probably target only a small subset of DocBook. This is the approach I have adopted with the Digital Scholar's Workbench.

*TEI*
TEI stands for the Text Encoding Initiative [23]. It is designed mostly for the preservation of literary and linguistic texts. Like DocBook, TEI is huge. Furthermore, it's not exactly *a format*, but a set of guidelines for building more specialised formats. One such is TEI-Lite, which has proved very popular, and is used by several repositories.

TEI may be better-matched than DocBook to some scholarly work, particularly in the humanities. It does have some serious shortcomings however:

- It uses abbreviated element names like `<p>` for paragraph (where Docbook uses `<para>`). This is presumably to make it easier to key in by hand, but it is a problem for sustainability since it may make it more difficult to recover the meaning of the markup in the distant future.
- While it has a set of customisable XSLT stylesheets [24], the impression I get is that they are less mature and less comprehensive than the DocBook XSL stylesheets [22].

Whether or not the TEI XSLT stylesheets are up to the job, TEI needs to be considered as a serious candidate for a preservation format for some scholarly writing. Ideally a full solution to the preservation problem would support both DocBook and TEI, allowing authors or curators/archivists to choose the most suitable format for preserving each work (or collection of works).

*XHTML+CSS*
Since XHTML [25] is both valid XML and can be displayed by web browsers directly (with the formatting controlled by a CSS [26] stylesheet) this has been suggested as a possible archival format. I don't recommend it, for the following reasons:

- XHTML is essentially a flat format, which means it's harder to do useful conversion into other formats in the future. It's possible to use the `<div>` element creatively to add lots of structure, but if you're going to do that, you're much better off using a well-defined structured format like DocBook or TEI.

(Why? Because in those formats the structural elements are rigorously defined, while in XHTML you can use divs however you like, making it hard for processing applications to know what to do. See the section on Custom schemata below.)

- CSS relies on consistent use of the “class” attribute in the XHTML. There is no standard for doing this. Same problem as above.
- CSS is not XML, so parsing it to convert it into some new format in the future is much harder than with XML formats.

XHTML might be a good solution for low-value documents that archivists cannot afford to convert into DocBook or TEI. In these cases a reasonable strategy might be to store the document in Open Document Format and add an automatically generated, perhaps poor quality, XHTML+CSS version for easy viewing and searching. This could either be stored in the repository alongside the ODF version, or could be generated on the fly by a front-end like Cocoon [27].

*Custom schemata*

One of the biggest traps in the XML world is the idea that you create your own document schemata that perfectly match your particular needs. A university could create document types for lectures, lab exercises, reading lists, research papers, internal memos, minutes of meetings, rules, policies, agendas, monographs and so on. There are serious problems with this approach [28].

The first problem is maintenance. Each format will require converters to turn word processor documents into that format, and XSLT stylesheets for rendering into whatever viewing formats are needed: a reasonable short list would be HTML, PDF and plain text. What happens next is that someone wants to add an element to one of the document types. Every time this happens, you have to modify all the stylesheets for that document type. With multiple formats, there is likely to be demand for conversion between them: converting a stored research paper into a lecture, for example. The number of conversions needed grows fast.

The second problem is loss of interoperability. One of the long-term goals of the whole repository project is that one should be able to retrieve a chunk of something from the repository, and drop it into another document of a different type. The use of custom schemata acts against these goals.

We’d also like people elsewhere to be able to use the documents that we go to so much trouble to preserve in our repository, but we can’t expect them to know all about our special document types. So then we would have to create even more converters for exporting the documents in well-known interchange formats.

## 4 Converting Documents Into Archival Format

Having decided on a suitable archival format, the second issue is how to convert documents into that format. In a nutshell, this is solved by using OpenOffice.org to convert multiple formats (including Microsoft Word) into Open Document Format, then unzipping the result and applying one or more XSLT transformations to the pieces.

For some parts of the processing, particularly creating deep structure from a flat sequence of headings and paragraphs, XSLT can be quite cumbersome. Direct manipulation (for example using one of the various DOM bindings: Java, Python, Perl, PHP...) is an alternative worth considering.

The only drawback of DocBook (and the same applies to TEI) is that most word processing documents do not contain enough structure information to allow for easy automated conversion. In order to convert word processor documents into DocBook (or TEI), some human effort is required:

- The best scenario is that the document was created using a well-designed word processor template, so that every paragraph has a style name attached to it. These styles can then be used as hooks by an automated conversion process in order to deduce structure. The USQ ICE project [29] is an example of this approach, as is the Digital Scholar’s Workbench (see below). One of my current interests is in the possibility of creating a heuristic “structure guesser” that can use formatting information (indents and justification, space above and below, type size, weight and style etc) to make educated guesses about the structural roles of different paragraphs in documents that were not created using a good consistent set of styles. This is unlikely to ever be a perfect hands-off process however, and will probably always require some human supervision.

- For legacy documents or authors who refuse to use a template, the word processing document will have to be edited by an electronic archivist to get it into a state where it can be converted to DocBook (or TEI). This would require trained staff, and costs time and money.
- For documents that are extremely poorly formatted, or that exist only on paper, another alternative is to send them out to be rekeyed. This is expensive, but for high-value documents or for small projects it may be worth it. A few thousand dollars for typing and marking up a book may compare well with the cost of setting up the infrastructure to do automated conversion, training staff to do the technical editing (cleaning up the markup, making it conform to a template) and so on. One important possibility worth investigating here is of having documents re-keyed in Word using a good template, and then converting to DocBook automatically. A first inquiry about this suggests that it costs roughly three times as much to mark up text in DocBook XML as it does to rekey it in Word. That means we can potentially save two-thirds of the cost if we do the conversion to DocBook with an automated process [30].

# 5 Usage

The main problem faced by institutional repositories (electronic archives) is no longer technical but social. Very few academics deposit their work. They're not interested, and it’s too much trouble. After they finish preparing a piece of work for publication—sometimes a very time-consuming and frustrating process that can take days—they want to move on to the next piece of work. The last thing they want to do is start all over again reformatting and preparing their work for archiving, and then having to type lots of metadata for search and indexing.

I believe that the key to solving this problem is seeing archiving as just another form of publishing. Just like a journal, an archive has its technical requirements in terms of format, metadata and so on. Rather than having to go through this time-consuming and frustrating process by hand, more than once, it would be much better to create a system that can do the whole thing automatically or at least semi-automatically. This is what a good, end-to-end electronic publishing system should be able to do. Once we have a document and its associated metadata in a good, structured XML format, all this should be possible, and more including:

- Sending to a journal
- Submitting to a conference
- Depositing in an archive
- Posting to the department web site
- Posting to a personal blog
- Running off preprints to send to colleagues
- Adding the abstract to the department annual report
- Registering with research productivity measures

A system that offered all these services might be able to attract users from among the academic community. This is the challenge we are trying to meet with the Digital Scholar's Workbench. It’s worth taking a few lines to discuss this in software engineering terms. What I’m proposing here looks very like “requirements creep”, the process by which a simple software development project gets hopelessly bogged down by a constantly expanding list of requirements. Usually developers are instructed to get a list of requirements early in the process and then lock it down so that the size and complexity of the project can be contained.

In the case of this project, that would have defeated the purpose of the work entirely. A single-purpose piece of software that converts Word documents into DocBook XML already exists. (It is a commercial product called UpCast [31].) Apart from the expense involved in making it available across the university, the problem with this software is that very few ordinary authors will take the trouble to learn how to use it. This is because there is no incentive to do so. Academics don’t care about preservation, so almost *any* effort is too much.

Some in the archiving community advocate taking an authoritarian stance and *requiring* academics to archive their work. Apart from any philosophical objections, this approach has another problem, that of quality of metadata attached to documents. If academics are forced to fill in lots of metadata forms, there will be a temptation to save time by entering rubbish. This defeats the whole purpose of archiving. If no-one can find your work, what is the point of preserving it?

Instead of forcing people to do something they don’t want to do, the approach I support is to provide a tool that is so useful they will want to use it. Archiving comes for free as part of the package—once someone is using this tool, clicking a button to archive completed work is no trouble at all. Metadata can be scraped from the document, or at worst only has to be entered once when the document is created. Then it can be used and re-used whenever the document is published or archived or submitted to a journal or conference.

The point here is that what looks like requirements creep is actually a deliberate strategy to create something that will actually be used. By surrounding the archiving functionality with features academic writers will actually want, we make it far more likely that work will be deposited in the archive.

## 6 The Digital Scholar's Workbench

The Digital Scholar's Workbench [32, 33] is a prototype application that implements some of the ideas in this report. At the moment the Workbench is a web application that converts suitably structured word processing documents into archival quality DocBook XML, and then from there into XHTML for onscreen viewing and into PDF for printing.

In order to work with the current version of the workbench, documents must be written using the USQ ICE template [29]. This template has a single set of all-purpose styles [34] designed so that it is possible to automate the conversion to DocBook XML. To many people, this seems like a major restriction, and a very common response is that people simply won't do that. Experience shows however [35], that authors *will* work with a template if:

- it is sufficiently rich to capture their documents without restricting them too much; and
- they can see the benefits in terms of time saved wrestling with their documents and trying to convert them into other file formats for publication.

This approach is backed up by Liegmann, who states that, "All of us … have experienced that you need to tolerate and to adhere to a structured framework in order to profit from its advantages." [36] At the time of writing, the Digital Scholar's Workbench has the basic functionality of being able to convert word processing documents written using the USQ ICE template into DocBook XML and from there into XHTML and PDF. Further development will focus on:

- making its support for Word and Writer documents more robust, and supporting more of the features, available in the word processing software;
- improving the links to repository software, so that documents can be deposited, together with associated metadata and any linked images or other resources, with one click;
- adding one-click publishing to blogs, websites and learning management systems;
- reformatting papers ready for submission to journals and conferences, via a plugin mechanism, so that once an export plugin has been created it can be contributed back to the community for use by others;
- support for complex multi-part documents like books and theses;
- articulation with desktop publishing software to produce high-quality typeset PDF output;
- improved metadata entry and storage;
- linking with a version control system, so that authors have access to all previous versions of their documents at all times;
- round-tripping of documents back to Word or Open Document Format to enable seamless collaboration between co-authors using different word processing software;
- platform- and software-independent bibliography management (perhaps outside the limited and difficult-to-use systems built in to Word and Writer);
- adding support for TeX and LaTeX documents;
- support for presentation slides; and
- attempting to remove the requirement that documents be written using only the set of styles in the ICE template.

The prototype workbench will eventually be made available to developers and early adopters as an open source software project, probably through SourceForge [37].
The Digital Scholar's Workbench is built on open-source technology. The current version uses the Apache Cocoon [27] web application framework, which incorporates the Xalan XML parser [38], the Xerces XSLT processor [39] and the FOP XSL-FO processor [40]. It also uses OpenOffice.org in headless mode to transform Word documents into Open Document Format. It relies on the USQ ICE template. (This architecture may change over the coming months.)

# 7 Example/Colophon

This document was begun in OpenOffice.org Writer on Linux, using the template provided by the conference organisers. When I couldn't format the reference list correctly using the bibliographic support built in to Writer, I moved the document to Microsoft Word on my Mac, and used EndNote [41] for the bibliography formatting. At first this didn't work; something Writer had done to the document meant that Word and EndNote couldn't talk to each other. In order to fix this, I had to start again with a fresh copy of the conference template and cut and paste my content across. Within the time constraints, there appeared to be no practical way to extract my bibliographic data from Writer and import it into EndNote, so I had to enter 43 references by hand. I estimate that I spent at least one entire working day simply wrestling with my word processing and bibliographic software, rather than working on the actual content of the paper.

Imagine now that the Digital Scholar's Workbench existed as described above. None of this would have been a problem. Transferring the paper from Writer to Word would have been accomplished via the DocBook interchange format. I'm not sure how the bibliography support will work—this may well be the hardest part of the work planned, or it may be that a service like Zotero [42] can do everything required—but there would certainly be no need to rekey entries from a bibliographic database. With an appropriate output filter, the workbench could have reformatted my paper to conform to the conference template, saving me the trouble. Alternatively, the conference organisers could have accepted the paper in DocBook XML, giving them the flexibility to typeset the proceedings for publication, and turn them into a web site.

# 8 Conclusion

This project was originally about digital preservation only. The Australian Partnership for Sustainable Repositories wanted to know how to preserve word processing documents. Answering that question led to the need to convert documents into structured XML for preservation. From there, the question was not just "How?", but also "How will we get people to actually do this?" The prototype Digital Scholar's Workbench is an attempt to answer both questions. The first is a purely technical question, but the second one is a people question, in the realms of the sociology of knowledge production. There is probably a long way to go with this, and the form of the software will change as we get feedback from focus groups of academics. The principle is that there is little chance of getting our users to do what the university archive wants them to do—convert their work into structured XML and deposit it in the archive—without offering them something they want, namely vastly simplified workflows for their everyday writing and publishing tasks.

# Acknowledgements

This work was funded by the Australian Commonwealth Department of Education, Science & Training, through the Australian Partnership for Sustainable Repositories (APSR), which is part of the government's Systemic Infrastructure Initiative, part of "Backing Australia's Ability—An Innovative Action Plan for the Future". I thank the Australian National Unversity's Digital Resource Services program, part of the Division of Information, for hosting me while I did this work: Program Leader Peter Raftos, APSR Project Leader Adrian Burton, and my colleagues Chris Blackall, Leo Monus, Scott Yeadon and Margaret Henty. I also thank the ANU's Department of Computer Science for releasing me from teaching while I did this work. I also thank Peter Sefton and his team at the University of Southern Queensland for many useful discussions particularly relating to the USQ Integrated Content Management system (ICE) which has a lot in common with the Digital Scholar's Workbench. I also thank them for creating the ICE word processor templates which I use in the current version of the Digital Scholar's Workbench. This paper is based on, and extends, an earlier report written for and published by APSR [43].

## Notes and References

[1] MÜLLER, E.; KLOSA, U.; HANSSON, P.; ANDERSSON, S.; SIIRA, E. Using XML for long-term preservation: Experiences from the DiVA project. Sixth International Symposium on Electronic Theses and Dissertations. Berlin, 2003. URL: http://edoc.hu-berlin.de/conferences/etd2003/hansson-peter/HTML/

[2] SLATS, J. Practical experiences of the digital preservation testbed: Office formats. File formats for preservation. Vienna, 2004. URL: http://www.erpanet.org/events/2004/vienna/presentations/erpaTrainingVienna_Slats.pdf

[3] ANDERSON, R.; FROST, H.; HOEBELHEINRICH, N.; JOHNSON, K. The AIHT at Stanford University: Automated preservation assessment of heterogeneous digital collections. D-Lib Magazine 2005;11. URL: http://dlib.org/dlib/december05/johnson/12johnson.html

[4] LESK, M. Preserving digital objects: Recurrent needs and challenges. 2nd NPO Conference on Multimedia Preservation. Brisbane, 1995. URL: http://www.lesk.com/mlesk/auspres/aus.html

[5] GIF. Wikipedia, 2006. URL: http://en.wikipedia.org/wiki/GIF

[6] STANESCU, A. Assessing the durability of formats in a digital preservation environment. D-Lib Magazine 2004;10. URL: http://dlib.org/dlib/november04/stanescu/11stanescu.html

[7] Microsoft Office Open XML formats overview. Microsoft, 2005-6. URL: http://www.microsoft.com/office/preview/itpro/fileoverview.mspx

[8] D'ARCUS, B. Citations in "Open" XML. 2006. URL: http://netapps.muohio.edu/blogs/darcusb/darcusb/archives/2006/06/08/citations-in-open-xml

[9] Ecma international standardization of OpenXML file formats frequently asked questions. Microsoft, 2006. URL: http://www.microsoft.com/office/preview/itpro/ecmafaq.mspx

[10] OpenDocument. Wikipedia, 2006. URL: http://en.wikipedia.org/wiki/Open_document_format

[11] Writer. OpenOffice.org, 2006. URL: http://www.openoffice.org/product/writer.html

[12] BALL, S. Multi-level non-uniform grouping of very large flat structured documents. AusWeb04, The Tenth Australian World Wide Web Conference, 2004. URL: http://ausweb.scu.edu.au/aw04/papers/refereed/ball/paper.html

[13] ERPA Advisory. ERPANet, 2004. URL: http://www.erpanet.org/advisory/list.php

[14] Rich Text Format (RTF) Specification, Version 1.8. Microsoft, 2004.

[15] Recovering and converting data from manuscripts collection discs. National Library of Australia, 2002. URL: http://www.nla.gov.au/preserve/digipres/recovering.html

[16] Extensible Markup Language (XML) 1.0 (Third Edition). World-Wide Web Consortium, 2004. URL: http://www.w3.org/TR/REC-xml/

[17] Sustainability of digital formats: XML. Library of Congress, 2006. URL: http://www.digitalpreservation.gov/formats/fdd/fdd000075/shtml

[18] ARMS, C.; FLEISCHHAUER, C. Digital formats: Factors for sustainability, functionality and quality. IS&T Archiving Conference. Washington DC, 2005. URL: http://memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pdf

[19] XSL Transformations (XSLT) Version 1.0. World-Wide Web Consortium, 1999. URL: http://www.w3.org/TR/xslt

[20] WALSH, N.; MUELLNER, L. DocBook: The definitive guide. O'Reilly, 1999. URL: http://www.docbook.org/

[21] OASIS. OASIS Consortium, 1993-2006. URL: http://www.oasis-open.org/

[22] STAYTON, B. DocBook XSL: The complete guide. Sagehill, 2005. URL: http://www.sagehill.net/book-description.html

[23] The Text Encoding Initiative: Yesterday's information tomorrow. TEI Consortium, 2006. URL: http://www.tei-c.org/

[24] RAHTZ, S. XSL Stylesheets for TEI XML. 2006. URL: http://www.tei-c.org/Stylesheets/teic/

[25] XHTML 1.0 The Extensible HyperText Markup Language (Second Edition). The World-Wide Web Consortium, 2002. URL: http://www.w3.org/TR/xhtml1/

[26] Cascading Style Sheets, level 2 CSS2 Specification. The World-Wide Web Consortium, 1998. URL: http://www.w3.org/TR/REC-CSS2/

[27] The Apache Cocoon Project. Apache, 2006. URL: http://cocoon.apache.org/

[28] BRAY, T. Don't invent XML languages. 2006. URL: http://www.tbray.org/ongoing/When/200x/2006/01/08/No-New-XML-Languages

[29] Integrated Content Environment. University of Southern Queensland, 2006. URL: http://ice.usq.edu.au/

[30] MONUS, L. Personal communication. 2006.

[31] upCast. Infinity Loop, 2003-2007. URL: http://www.infinity-loop.de/products/upcast/

[32] BARNES, I.; YEADON, S. One-click DSpace ingestion with the Digital Scholar's Workbench. Open Repositories 2006. Sydney, 2006. URL: http://www.apsr.edu.au/Open_Repositories_2006/barnes_yeadon.ppt

[33] BARNES, I. Integrating the repository with academic workflow. Open Repositories 2006. Sydney, 2006. URL: http://www.apsr.edu.au/Open_Repositories_2006/ian_barnes.pdf

[34] SEFTON, P. OpenDocument or not, you still need to Use Styles. 2005. URL: http://ptsefton.com/blog/2005/09/13/opendocument_or_not_you_still_need_to_use_styles

[35] SEFTON, P. Personal communication. 2005.

[36] LIEGMANN, H. Long-term preservation of electronic theses & dissertations. Sixth International Symposium on Electronic Theses and Dissertations. Berlin, 2003. URL: http://edoc.hu-berlin.de/conferences/etd2003/liegmann-hans/HTML/liegmann.html

[37] SouceForge. Open Source Technology Group, 2001-2006. URL: http://sourceforge.net/

[38] Xalan. Apache, 2005. URL: http://xml.apache.org/xalan-j/

[39] Xerces. Apache, 2005. URL: http://xerces.apache.org/xerces-j/

[40] FOP (Formatting Object Processor). Apache, 2006. URL: http://xmlgraphics.apache.org/fop/

[41] EndNote 9. Thomson ResearchSoft, 2005. URL: http://www.endnote.com/

[42] Zotero. Center for History and New Media at George Mason University, 2006. URL: http://www.zotero.org/

[43] BARNES, I. Preservation of word processing documents. Australian Partnership for Sustainable Repositories, 2006. URL: http://www.apsr.edu.au/publications/preservation_of_word_processing_documents.html

# Evaluating Digital Humanities Resources: The LAIRAH Project Checklist and the Internet Shakespeare Editions Project

*Claire Warwick; Melissa Terras; Isabel Galina; Paul Huntington; Nikoleta Pappa*

School of Library, Archive and Information Studies, University College London
Henry Morley Building, Gower Street, London, WC1E 6BT, United Kingdom
e-mail: c.warwick@ucl.ac.uk; m.terras@ucl.ac.uk, i.russell@ucl.ac.uk, p.huntington@ucl.ac.uk

## Abstract

The following paper presents a case study of the way that the research done by the LAIRAH project may be applied in the case of a real digital resource for humanities scholarship. We present an evaluation of the Internet Shakespeare Editions website according to the checklist of recommendations which we produced as a result of our research. The LAIRAH (Log analysis of Internet Resources in the Arts and Humanities) project based at UCL's School of Library Archive and Information Studies, was a fifteen month study to discover what influences the long-term sustainability and use of digital resources in the humanities through the analysis and evaluation of real-time use. Our research objectives were to determine the scale of use and neglect of digital resources in the humanities, and to determine whether resources that are used share any common characteristics. We also aimed to highlight areas of good practice, as well as aspects of project design that might be improved to aid greater use and sustainability. A further aim was to determine whether digital resources that were neglected. In our study we concluded that well-used projects share common features that predispose them to success. The effect of institutional and disciplinary culture in the construction of digital humanities projects was significant. We found that critical mass was vital, as was prestige within a university or the acceptance of digital methods in a subject. The importance of good project staff and the availability of technical support also proved vital. If a project as to be well-used it was also essential that information about it should be disseminated as widely as possible. Even amongst well-used projects, however we found areas that might be improved, these included organised user testing, the provision of and easy access to documentation and the lack of updating and maintenance of many resources. The paper discusses our recommendations, which were presented as a check-list under four headings: content, users, maintenance and dissemination. We show why our findings led us to make such recommendations, and discuss their application to the ISE case study.

**Keywords:** digital humanities; user studies; good practice resource construction

## 1 Introduction

The following paper presents a case study of the way that the research done by the LAIRAH project may be applied in the case of a real digital resource for humanities scholarship. In it, we present an evaluation of the Internet Shakespeare Editions website according to the checklist of recommendations which we produced as a result of our research (http://www.ucl.ac.uk/slais/research/circah/features/).

The LAIRAH (Log analysis of Internet Resources in the Arts and Humanities) project (http://www.ucl.ac.uk/slais/research/circah/lairah) based at UCL's School of Library Archive and Information Studies, was a fifteen month study to discover what influences the long-term sustainability and use of digital resources in the humanities through the analysis and evaluation of real-time use. Our research objectives were to determine the scale of use and neglect of digital resources in the humanities, and to determine whether resources that are used share any common characteristics. We also aimed to highlight areas of good practice, as well as aspects of project design that might be improved to aid greater use and sustainability. A further aim was to determine whether digital resources that were neglected might be re-used. As a result of this research the project created a list of recommendations for features that, if possible, the idea successful digital resource ought to have. We also made recommendations to aid the UK's Arts and Humanities Research Council (AHRC), who funded to project, to develop their strategy for funding usable digital resources for future humanities research.

Numerous studies have been carried out into the information needs and information seeking practices of humanities scholars, over recent years [1-5]. We are not aware, however, of any literature that has used quantitative methods, particularly deep log analysis, (described below) to measure the levels of use of digital humanities resources. Our research also concentrates not on the generality of resources, but on the question of what *kind* of digital resource is most useful for researchers. Although this has been approached by other projects, evidence has been entirely self-reported. Our research is also the first study which has enabled a comparison of the preferences that users report to quantitative evidence of what they actually use.

## 2 Methodology

The research was funded by the AHRC as part of its ICT Strategy Scheme. We therefore studied digital resources for humanities research which were based in the UK and non-commercially funded. In the first phase of the project we used deep log analysis of web servers of three humanities portal sites in the UK to determine whether it was possible to assess levels of use of digital resources accessed through these portals. These were the Arts and Humanities Data Service, (AHDS) Humbul Humanities Hub, and Artefact (the last two have now merged to become Intute Arts and Humanities) We discuss this analysis in more detail elsewhere, however, in essence we used the data that web server logs automatically record to determine how the sites were used, in terms of levels of use, which parts of the site were used, where users came from and where they went after leaving the site [6]. Although absolute levels are difficult to gauge our research suggested that roughly a third of the resources remained unused. As a result of our analysis we then chose a sample of projects to be studied in more depth. In the following paper, therefore, the results are mainly based on our qualitative methods. We show how the resulting recommendations may be applied to the analysis of an actual digital humanities research resource, and how our work is adding to the dissemination of good practice in digital resource construction and sustainability.

Our qualitative methods involved the selection of a sample of twenty one well used projects for further study. These were studied to see whether there were any common elements of good practice amongst resources that were well used. A representative from the team which constructed the resource, ideally the PI, was interviewed and any documentation available though the project website studied.

To determine whether neglected projects could be reused, we ran two user workshops where participants were asked to examine a sample of eleven resources which were both used and neglected and to discuss their opinions of them. To identify the neglected resources we used the results of the log analysis and also contacted representatives of the AHDS, who gave us additional information about which resources they felt to be most commonly used, or entirely neglected. We also wished to know whether there was any reason why neglected resources were not used, in addition to possible lack of knowledge about them. We did not wish to create bias by telling participants which were used and which neglected, and asked whether they could determine which resources were neglected, and why they felt that this might be. We were surprised to find that participants were highly critical of resource quality, and tended to identify well-used projects as neglected, rather than the opposite.

As a result of our qualitative research we made a number of recommendations for good practice in digital resource construction. These are presented in detail in the project report [7], however, for ease of use, we also produced a simple checklist, intended for those who either are, or would like to be, the producers of scholarly digital resources for humanities research which may be found at http://www.ucl.ac.uk/slais/research/circah/lairah/features/. We hope, however that such recommendations may also be more widely applicable, and relate to other sectors of digital resource publication.

In the following paper, we use the checklist that we created and demonstrate how it may be used to evaluate a real digital resource, the *Internet Shakespeare Editions Project* (ISE) (http://ise.uvic.ca/index.html). This also provides a framework for a more detailed discussion of the findings from the qualitative phase of our research. We have used this site because the ISE team approached us after the intial findings of the LAIRAH project had been made public, and asked if we would be willing to use the checklist to evaluate their site. They will be using the results of the evaluation for further development of the ISE, however it also presents us with an ideal opportunity to show how the checklist can be employed in the case of an actual digital research resource. Thus each section begins with the recommendation, we then explain the basis on which we made it, as a result of the findings of our research , and the results of the evaluation are then discussed.

# 3 Results

## 3.1 Content

### 3.1.1 The ideal digital resource should have an unambiguous name that indicates its purpose or content

Our log data initially showed that the names and search terms that are use are significant. For example, resources entitled 'census data' were, not surprisingly, popular. However, a similar resource appeared neglected, perhaps because it was called 'Enumerator returns for county X'. Since keyword searching cannot automatically link synonymous terms, a search on 'census' would not have found the latter resource. Discussion at the workshops also revealed that participants could be confused by misleading titles. Some participants thought that a resource entitled 'The Channel Tunnel Rail Link Archive' would be neglected, since they assumed that it contained digitised records of a railway or engineering company. In fact it is a very well used archive of archaeological documentation for the excavations carried out before the link was built. Conversely a resource called 'The Imperial War Museum Concise Art Collection' was praised, since it was immediately clear to users what it contained, and gave the reassurance of a trusted brand in the museums world, thus participants assumed the contents would be of high quality.

The case study resource is called Internet Shakespeare Editions. This is an excellent description of the site, which also offers numerous additional resources for Shakespeare scholars. However it would be unrealistic to try to describe these in the site's name.

The URL, http://ise.uvic.ca/index.html, however does not reflect the name and may be difficult to remember, as compared, for example, to the Old Bailey Online project, who have acquired a domain name which is easy to recall. (http://www.oldbaileyonline.org/). However, one important facet of the current URL is that it places teh ISE within the domain of a respected univertisty. This type of institutional brand helps users to trust to integrity of the site and the quality of its contents. When a Google search was performed on the keyword 'Shakespeare' ISE came in the 50-60 screen of results, out of a total of 53,000,000 hits. This is a creditable performance, considering the popularity of Shakespeare as a topic. However, it might be improved by encouraging as many users, English departments, and libraries as possible to create links to the ISE page.

### 3.1.2 The ideal digital resource should concern a subject that is either popular in a wide community or essential for a smaller expert one

The log data demonstrated that certain subjects and themes were particularly popular, for example, warfare, census data, witchcraft, Shakespeare and women's suffrage. We do not know exactly what purpose the resources are being put to, whether high level academic study, family history research or a school history project, for example. However, it is clear that digital resources concerning certain popular subjects are likely to be well used. Nevertheless participants at our workshops stressed the importance of resources which might be vital to research in a relatively small community, whose work would be significantly impoverished without them. It would also be unwise to concentrate research funds on a small number of popular subjects, and neglect less popular areas, since we cannot know which topics, perhaps now neglected, may be widely studied in future.

The ISE website evidently concerns Shakespeare, which is both very popular in the wider community as well as being an important research topic for academics. The website offers two types of navigation, by Academic divisions- done through the metaphor of a building- and by Subject area. This is an interesting way to help different user groups to access the content. As part of the LAIRAH research we found that users felt confused by many sites that were only designed for experts in the subject, and assumed a high level of knowledge about resources and how they should be used. This dual path is thus good practice, since it allows users to access materials in a way that is most useful for them. The functions of the different parts of the site, and the different methods of navigating them are also very clearly described in the About section page, 'How to use this Site' http://ise.uvic.ca/Foyer/index2.html. This is a simple and very helpful page title, given that we have found that clear signposting is invaluable, especially for less expert users (discussed below in section 3.2.3)

### 3.1.3. The ideal digital resource should retain its server logs, and make them available to their funding agency and researchers, subject to confidentiality agreements

During our research we found that it could be relatively difficult to obtain log data, even from large publicly funded portals. Humbul was reluctant to allow us to use their data, even if anonymised, because of worries that individual users might be identifiable, because of their personalisation features. We were able to reassure them that this is not possible, and that any individual machine IP addresses would not be made public in any reports. However, the time taken to do the anonymisation held up our research considerably. Artefact were willing to give us the data, but had not had the technical support to be able to keep it, and what was kept was lacking in detail. Thus we were able only to access a small amount of data.

Many individual projects may be even less likely to realise the importance that their web logs may have as a potential research resource. They may not realise that they should be kept, nor the level of detail of logging that should be made possible, they may also lack technical support to do so. We therefore recommended to the AHRC that if resources were publicly funded they should be asked to keep logs and that as a function of the grant such logs should be made available to researchers for the purposed of monitoring and evaluation. This would avoid the kind of delays we experienced while permissions were negotiated. The ISE server logs have been retained and are made available. We hope to be able to analyse them in detail as part of the next phase of the LAIRAH research, if a funding application is successful.

### 3.1.4. The ideal digital resource should keep documentation and make it available from the project website, making clear the extent, provenance and selection methods of materials for the resource

The participants at our workshop were concerned that in many cases they could not find enough information about the content of the resource, how it was selected, and its provenance. They also wished to know more about the team that constructed the project, and the expertise of its members, and some of the more technically expert participants would have liked to have found out more about the technical methods and standards applied. They also felt the lack of the kind of information about sources that are found in the print world in citations and bibliography. All of this meta information helps to increase the trust that users have in the quality of digital resources.

However, in our study of even well used projects we found that levels of documentation were extremely variable. Some projects were extremely well documented, these tended to be in subject areas like archaeology and linguistics, where documentation of resources has always been an essential research practice. However, many projects kept little or no organised documentation. It could also be hard to find. Ideally documentation should be easily located from the project website. However, in many cases it was absent, was accessible only through the AHDS, or not at all. In other cases, although some documentation could be found via the website, it was complex, and difficult to locate or incomprehensible to the non-expert reader. One of the most effectively documented resources had been compelled to do so in the terms of their grant from the New Opportunities Fund. We therefore recommended that the AHRC should consider making documentation a deliverable of any funded project.

The ISE site is relatively unusual in that it is extremely well documented and there is ample information available. Most of the information is available from the ‘Academic divisions-Foyer’, with links from SubjectArea- About ISE’. It is especially important that the ‘About’ link is available from the top menu, making it as easy as possible for users to access the documentation. The ‘About’ page is also kept relatively short, is expressed in simple terms and has links to further material. This is ideal and should be retained, since our research has found that users become confused if documentation is too dense, complex or presented on a long page which requires them to scroll. The documentation includes:

- History of the site (http://ise.uvic.ca/Foyer/ISEoverview.html)
- Editorial Guidelines
- Details of people on Advisory Boards for different sections [Editorial Board, Advisory board Performance materials, Theatre History and Technical Design and Implementation]
- Details of editors of the plays and poems
- Technical information about the design of the site.
- Information about new site

## 3.2 Users

### 3.2.1 The ideal digital resource should have a clear idea of whom the expected users might be; consult them as soon as possible and maintain contact throughout the project via a dedicated email list or website feedback

Very few of the projects that we studied had any contact with their users, nor did they tend to consult them or provide much user interraction on the website, beyond a 'contact us' email link. This is a wasted opportunity, since contact with users should help project teams understand the needs of those who will use the site, and adapt it accordingly. This should in turn increase levels of use of the site.

The ISE website states that their aim "is to inspire love of Shakespeare's work in a world-wide audience". It would seem that they expect to have a global impact. There is no further explicit indication of their expected users. However, their division of SubjectArea and Academic would seem to suggest that academic users are expected, although the site is welcoming for non-experts, since it gives detailed descriptions of the material it contains and how this might be used. Contact with users is achieved through a discussion section and the provision of contact email links.

#### 3.2.1.1 Discussion [http://ise.uvic.ca/Annex/discussion.html]

This page informs users that 'When complete, this section of the site will provide an informal forum for the discussion of Shakespeare, his works, life, and the performance of his plays.' The discussion section will be launched in April 2007, and should provide an excellent opportunity for users to interact with each other and the site's creators. It should also be noted, however, that such forums tend to be labour intensive to keep updated. We have found that users distrust the quality of a site if there is evidence that it is not entirely up to date. Thus, once the discussion section is launched, sufficient resources will need to be allocated to it to ensure that it does not appear outdated.

#### 3.2.1.2 Contact email links

An email contact address is linked to from the following pages:

- About [http://ise.uvic.ca/Foyer/about.html]
- Policy on copyright [http://ise.uvic.ca/Foyer/copyright.html]
- Guidelines for the acquisition and copyright of performance materials [http://ise.uvic.ca/Foyer/PerfGuide.html]
- Shakespeare in performance –FAQ [http://ise.uvic.ca/Theater/sip/faq.html]
- About Shakespeare in performance [http://ise.uvic.ca/Theater/sip/about.html]

Each contact link is placed within very different contexts and is used for different types of information, thus it may be that the addition of an overall 'contact us' link on the top menu would be beneficial for users.

### 3.2.2 The ideal digital resource should carry out formal user surveys and software and interface tests and integrate the results into the project design

Once again few of even the well used resources in our survey had carried out any kind of formal user testing. One project subsequently regretted this, since it had worked very hard to produce a complex search interface, only to find that in practice it was too complex for the majority of their users.
The ISE carried user testing before the launch of the new design of the site in November 2005. The sample was of about 15 participants, and designed to represent the needs of different user groups. It included students at different levels, English faculty of various ages, skilled computer personnel with no great knowledge of Shakespeare, and several general readers. The results were used to aid the design of the navigation, both to encourage initial entry to the website, and to navigate within the site. ISE made a significant number of modifications as a result of the testing. This testing represents good practice in resource design, and makes ISE relatively unusual in the field of digital humanities resources.

Positive feedback from students and teachers who have used ISE has also been placed the website at http://ise.uvic.ca/Foyer/corporate.html. This shows that users have found the contents helpful in their work. The site has also been given various awards for excellence from internet bodies and those concerned with academic study in general and English literature in particular.

### 3.2.3 The ideal digital resource should be designed for a wide variety of users, and include information to help the non-expert to understand the resource and use its contents

At the workshop participants thought that many resources appeared only to be designed for subject experts, and therefore deterred the more general user. They argued that the inclusion of simple instructions would be very helpful for the non-expert, and would not affect the experience of the expert user. This proved to be important, since the participants quickly gave up trying to use a website if they were unable to work out how it should be used. Simple, clear signposting should therefore help to increase levels of use of digital resources, since it encourage non-experts to persist with user of the site.
The ISE website seems to be designed for a wide variety of users. A large section dedicated to Shakespeare's life and times (http://ise.uvic.ca/Library/SLT/intro/introsubj.html) seems to cater for the novice. Whilst Scholarly Articles on Shakespeare and the Internet (http://ise.uvic.ca/Annex/Articles/index.html) appear to be directed at the Shakespeare scholar.

Access to the main site is achieved by clicking on the image of the library which makes up most of the top page. The visual metaphor is appealing in many ways, and undoubtedly attractive, and the ISE team may not wish to spoil it with instructions for use. However, however it may not be evident to all users that they should click this image and our research shows that users are quick to give up on using a resource if they do not find obvious clues about how to use it. In many cases such instructions were not provided by the project team, as the use of the resource may have seemed entirely obvious to its producers. The ISE are planning to address this potential difficulty by creating a more obvious link to their newsletter page- (see discussion below in section 3.4.2)

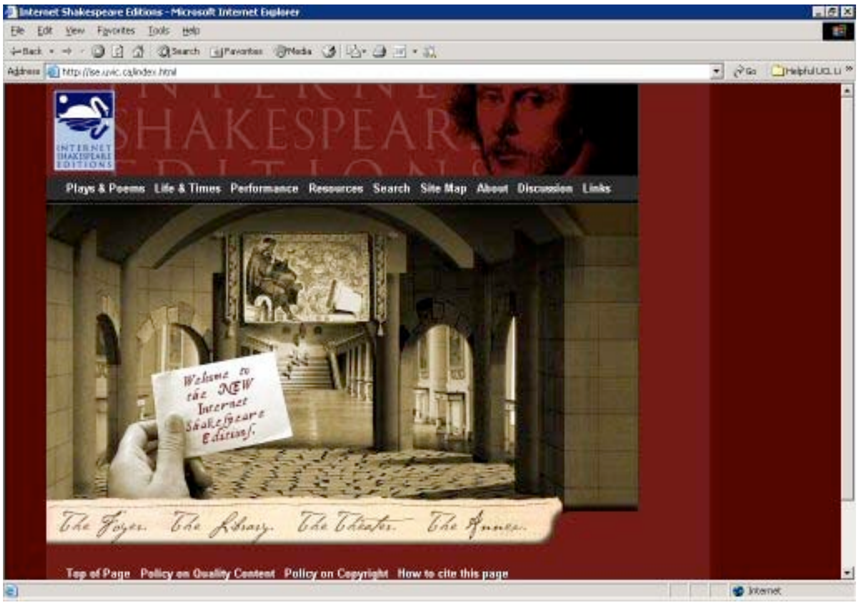

Internet Shakespeare Editions - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://ise.uvic.ca/index.html
Go Helpful UCL Li
INTERNET SHAKESPEARE EDITIONS
Plays & Poems Life & Times Performance Resources Search Site Map About Discussion Links
Welcome to the NEW Internet Shakespeare Editions!
The Foyer. The Library. The Theater. The Annex.
Top of Page Policy on Quality Content Policy on Copyright How to cite this page
Internet

Figure 1: The Top page of the ISE website

Navigation is described in the 'How to use this site' page of the Foyer section. This contains a great deal of very useful information about the contents of the site, who created it, the type of material included, its provenance and extent. This is also easily located by following links from the page. This is vital since this is the kind of information which encourages users to trust the quality of a digital resource. The information provided by ISE should be more than sufficient to reassure users of its high academic standards.

## 3.3 Management

### 3.3.1 The ideal digital resource should have access to good technical support, ideally from a centre of excellence in digital humanities

It was not surprising to find that many of the well used projects were associated with centres of excellence in digital humanities. The Humanities Research Institute at the University of Sheffield for example was the base for several projects whose use was prominent in the log data. This is understandable, since it is difficult for individual humanities researchers to keep up with the latest developments in digital techniques and technical standards. Thus it is vital for researchers to have access at least to a computer support officer, and ideally to such a centre, whose staff will understand not only technical aspects of resource construction, but also the demmands of humanities research.

The ISE is based at the Humanities Computing and Media Centre at the University of Victoria in Canada (UVic) and thus has access to a high level of technical expertise and advice. There is also an advisory board in Technical Design and Implementation (http://ise.uvic.ca/Foyer/techboard.html). The page includes names and details of the four members. Between them they have both technical and digital humanities expertise. This is an ideal arrangement, since it gives the ISE access to the latest information about technical developments and good practice in digital humanities.

### 3.3.2 The ideal digital resource should recruit staff who have both subject expertise and knowledge of digital humanities techniques, then train them in other specialist techniques as necessary

The recruitment and training of staff to work on digital humanities resources was a particular challenge for the leaders of the projects whom we interviewed. It could be especially difficult to find staff who were not only technically adept, but also had sufficient knowledge of humanities research that they understood the material itself, and were thus able to mediate between the needs of researchers and technical functionality.

It was most usual for humanities specialists to be recruited, but they then needed to be trained in various computing techniques. This often proved difficult since the amount and quality of training available through universities was often disappointing. PIs also commented that training new researchers often took longer than expected, which could be a significant problem, when the project was operating on time-limited funding.

The ISE has obviously been able to recruit good project staff. The site lists all research assistants (undergraduate and graduate), from 1992 to the present day. Most have subject expertise and/or digital humanities knowledge, plus other areas of expertise depending on their function. The ISE is also very well provided with academic members, editors and advisors. There is one Editorial Board and three Advisory Boards for Performance Material, Theatre History and Technical Design and Implementation. There is also detailed information about the editors of the online editions. This is important, since detailed information about the academic qualifications and technical expertise of the project team helps to reassure users that the material to be found on the site is of the highest academic standards. It is also important that the ISE list the institutional affiliations of board members, since such affiliations appear to act in a similar way to the trusted brand status of commercial sites, such as the BBC for news resources. Once again they help users trust the quality of the resources to be found.

### 3.3.3 The ideal digital resource should have access to short term funds to allow to retain expert staff between projects

A further problem where staffing was concerned was that in the UK most non-commercial digital humanities projects are made possible by short term grants of public money, usually from the AHRC. This funding is relatively scarce and to it is very difficult for projects to obtain continuous funding, and retain skilled staff. This resulted in wasted resources, since new staff had to be appointed and trained for each new tranche of funding granted, rather than PIs being able to rely on a cadre of expert employees, as is more often the case with scientific funding. We therefore suggested that the AHRC might consider making available small amounts of short term funding, to allow employees to be retained for a few months in the hope of securing further long term funding, as is the practice with some UK science funding councils.

The ISE is relatively fortunate in this regard, in that the Canadian funding system appears to make it easier to access small amounts of grant money on a continuing basis, both from universities themselves, and from public funding. The uninterrupted recruitment of research assistants noted on the site suggests continuous availability of funding. The University of Victoria, The Social Sciences and Humanities Research Council of Canada and the Innovation Development Corporation are listed as ‘supporters’ (http://ise.uvic.ca/Foyer/acknowledge.html). This suggests ongoing access to funding to develop the resource. This is ideal, but rare outside North America, and it is to be hoped that such funding will continue to be available to the ISE.

## 3.4 Dissemination

### 3.4.1 The ideal digital resource should have an attractive usable interface, from which all material for the project may be accessed without the need to download further data or software

The need to have an interface that is attractive and usable may seem too obvious for any need for comment. However, workshop participants found that many of the interfaces to the resources in the sample, even well used ones, were problematic. This may be because very few projects had any contact with experts in HCI, or interface design. It is possible to apply for funding for a professionally designed site, but none of the projects in our sample had done so. As a result participants found most unattractive compared with the professional interfaces of commercially produced resources. This is significant, given that it appears that most users make decisions about websites extremely quickly [8], and so an unwelcoming interface is likely to deter many users before they have even accessed the resource's contents.

We also found that several projects, especially databases for historical research, required users to register to use the data, for which they had to be given a password before they were allowed access. Data would then have to be downloaded and used with specialist software. Such registration is sometimes unavoidable, for example for reasons of copyright. However, it was a serious deterrent, for all but the most determined users, and thus should be avoided if at all possible.

The user studies performed by the ISE have helped the team to develop the interface and navigation to help users find their way around the site. Although some users may find the visual metaphor a little confusing, in general the interface is attractive and easy to use. The site is extensive and complex, with a large amount of material, which by nature will mean that the first time user will need to spend some time exploring to find what they need. The different sections are generally well signposted however, and explanations are provided about what kind of content may be found in each section. The pages are clear and well-written, and should be easily comprehensible by web users. This is important, since users of the web tend to skim pages, and take in less content than if they were reading printed material. Pages therefore need to be concise, divide into easily comprehended sections, and be written in a clear and accessible style [9]. All of the data needed may also be accessed without need for further software or to download the data for local use.

### 3.4.2 The ideal digital resource should maintain and actively update the interface, content and functionality of the resources, and not simply archive it with AHDS

As discussed above, it is important that web sites should be updated regularly. Now that many commercial sites are updated constantly, users have an expectation of currency, and our research has shown that they may therefore distrust the quality of resources that appear not to be actively updated.

Some of the ISE pages contain information about when the site was updated. There is also a detailed description explaining how ISE has been created, updated and what ideas there are for the future. (http://ise.uvic.ca/Foyer/ISEoverview.html). In fact there is an active policy of updating the contents of the ISE site. A full-time student updates content in the database of Shakespeare in Performance: this is usually done daily. She also checks the links regularly. A minor update of the Life and Times section (mainly the bibliographies) is planned for April 2007. The texts are updated as they are completed, but this section of the site is still under construction as the ISE team are looking at various ways of displaying annotations and textual variants.

Our study showed that most schoarly websites can learn from the commercial sector when it comes to providing information about how and when their site is updated. Evidently, for a site such as ISE, there is no need continually to update all pages, however practice in providing such information should be consitent throughout the site. Another way to deal with this is to provide a link to news about the site, and provide information about when new content is added. Some projects use the front page for such updates (http://www.ucl.ac.uk/english-usage/), however another approach is to link from the front page to a news, or what's new page (for example http://ahds.ac.uk/). An RSS feed such as that used by the AHDS is often used commercially to encourage users to revisit sites of interest and could also be used if significant numbers of changes are being made. All of these measures reassure users that updating is happening. The ISE plan to introduce a regular newsletter, the first issue is being prepared for the end of April 2007and the front page will have a hand that invites the user to link to the newsletter rather than simply providing a welcome page. They expect the newsletter to be a regular feature, issued three or four times a year, and plan to send it to a large readership base of libraries and English

Departments, who it is hoped will create links to the site as a response. They also plan to use RSS feeds to automate the provision of information about page updating.

### 3.4.2.1 Maintenance

Maintenance is a problematic issue for non-commercial digital humanities resources, since after funding runs out, there may be no resources to make sure that the resource is maintained. Although a recent funding call from the UK's JISC (Joint Information Systems Committee) requires universities to guarantee that they will maintain funded resources for at least ten years [10] this involves a commitment of server space and personnel to do so, a cost which the institution may not feel able to bear in the long term. There can also be the danger that if a member of staff leaves or retires the university may not feel obliged to maintain the resource, and thus, in the worst case, it could be entirely lost. It is also clear that simply leaving data on a server without active maintenance and updating is not satisfactory. One project in our survey was no longer updated or maintained, and the original researcher who created it was aware that non only was the website seriously outdated, but the functionality of the database itself was gradually deteriorating. No-one was paid to maintain the resource, however, or had time to do so voluntarily, and thus only ten years after its construction it was becoming unusable.

The Humanities Computing and Media Centre at UVic, runs the ISE server. The whole system is, however, backed up automatically, off site, by the University Computing and System Services who also perform basic system-level maintenance. This ensures that the data is safely maintained at present, and the ISE are currently negotiating with UVic for continuing support for infrastructure. The ISE is constituted as an independent, non-profit organisation, therefore, it could if necessary exist independently of the university, and be moved to a different server. However it would be ideal if UVic were able to commit themselves to long term maintenance of the data.

In the UK all research that creates digital output, and is funded by the AHRC, must be offered for archiving with the AHDS, which preserves the data, although it does not, of course update it, or maintain a website. However, this option is not available in Canada, and thus individual projects and their host institutions must negotiate such archiving on a piecemeal basis. For example, ISE archive all artefacts for the database at 600dpi TIFF files. All other files are handled by a version control system (Subversion) which keeps track of all changes. This is good practice individually, however without a central service like the AHDS to offer archiving facilities and advice about good practice, many projects may not be aware of the standards to which they should adhere, and thus their maintenance strategy might not be as rigourous. This piecemeal approach is therefore not ideal for digital resources, and at worst potentially poses a very serious threat to their long-term sustainability.

### 3.4.3 The ideal digital resource should Disseminate information about itself widely, both within its own subject domain and in digital humanities

The strongest possible correlation in our study between a characteristic of a resource and its use was dissemination. All the projects in our study had worked hard to provide information about their work, by giving papers and seminars in both the digital humanities and publishing sectors, and within their own subject areas. This is an important new role for academics, since previously they would have written books, and relied on publishers to market them. In the case of a digital resource, the scholar is now the publisher, and so the responsibility of disseminating information about their work not falls to them.

The ISE has a good level of web visibility. A simple link analysis shows over 5,000 links to any page in http://ise.uvic.ca. Publications and papers given as a result of ISE research are also listed in the annex there on a page entitled: 'Scholarly Articles on Shakespeare and the Internet' (http://ise.uvic.ca/Annex/Articles/index.html). This shows that the ISE members have been active in disseminating information about their research and the project itself. These include both conferences and journals in English literature and conferences on Humanities computing, although papers do not appear to have been published in any humanities computing journals. The website also cites an example of use of ISE:

> *"An idea I discussed at a meeting of the Shakespeare Association of America three years ago, that ambiguous readings and imprecise entrances or exits could be indicated by animation, has been received with a possibly surprising enthusiasm by the editing community in Shakespeare studies: http://ise.uvic.ca/Annex/Articles/SAA2002/rich4.html"*

# 4 Conclusions

This paper has shown how the findings of our research on the LAIRAH project have been used to construct a check-list of recommendations. We have further shown how such recommendations may be applied in the evaluation of an example digital humanities resource.

The Internet Shakespeare Editions project is an example of excellent practice in the construction of digital humanities resources. It maintains consistently high standards both of content, presentation and technical web design. This evaluation shows that the ISE performs very well when judged according to the recommendations made in the LAIRAH checklist. In many aspects it out performs many of the well used resources in the LAIRAH research sample.

It is an attractive, usable resource, with a wealth of useful content, which is comprehensively documented. It is able to maintain levels of funding, which allow it to recruit able research staff, and it is well supported by a humanities computing centre, by expert editors and well qualified advisory boards. All of these factors help to ensure that users will recognise the content as trustworthy and of good quality. It is also clear that information about the resource is widely disseminated, both in digital humanities and English literature. The resource is actively updated and should continue to be maintained by the University of Victoria. Like all digital resources, especially those where no national archiving system exists, it will inevitably face problems of long term sustainability. However, the team is aware of these, and is taking steps to try to ensure the resource's future.

## Acknowledgements

LAIRAH was funded by the AHRC ICT Strategy scheme. We would like to thank all those who took part in workshops, and agreed to be interviewed during the project. We especially thank Michael Best, Roberta Livingstone and the Internet Shakepeare Editions project team for allowing us to use ISE as a case study.

## Notes and References

[1] BARRETT, A. The information seeking habits of graduate student researchers in the humanities. *The Journal of Academic Librarianship*. 2005, vol. 31, no. 4, pp. 324-331.

[2] TALJA, S.; MAULA, H. Reasons for the use and non-use of electronic journals and databases - A domain analytic study in four scholarly disciplines. *Journal of Documentation*. 2003, vol. 59, no. 6, pp. 673-691.

[3] HERMAN, E. End-users in academia: meeting the information needs of university researchers in an electronic age Part 2 Innovative information-accessing opportunities and the researcher: user acceptance of IT-based information resources in academia. *Aslib Proceedings*. 2001, vol. 53, no. 10, pp.431-457.

[4] BRITISH ACADEMY. *E-resources for Research in the Humanities and Social Sciences - A British Academy Policy Review.* 2005. Available from Internet: http://www.britac.ac.uk/reports/eresources/report/sect3.html#part5

[5] DALTON, M. S.; CHARNIGO, L. Historians and their information sources. *College & Research Libraries*. 2004, vol. 65, no. 5, pp. 400-425.

[6] WARWICK, C.; TERRAS, M.; HUNTINGTON, P.; PAPPA, N.; GALINA, I. 'If you build it will they come? The LAIRAH survey of digital resources in the arts and humanities. *Literary and Linguistic Computing*. 2007 (forthcoming).

[7] WARWICK, C.; TERRAS, M.; HUNTINGTON, P.; PAPPA, N.; GALINA, I, The LAIRAH Project:Log Analysis of Digital Resources in the Arts and Humanities. Final Report to the Arts and Humanities Research Council. Arts and Humanities Research Council. 2007 Forthcoming

[8] LINDGAARD, G.; DUDEK, C.; FERNANDES, G.; BROWN, J. Attention web designers: you have 50 milliseconds to make a good first impression. *Behaviour & Information Technology*. 2005, vol. 25, pp. 115-126.

[9] MORKES, J; NIELSEN, J., Concise, SCANNABLE, and Objective:How to Write for the Web. *Useit.com*, 1997. Available from Internet: http://www.useit.com/papers/webwriting/writing.html

[10] JISC. *JISC Circular 03/06: JISC Capital programme*. 2006. Available from Internet: http://www.jisc.ac.uk/fundingopportunities/funding_calls/2006/06/funding_circular03_06.aspx

# Feasibility of Open Access Publishing for Journals Funded by the Social Science and Humanities Research Council of Canada

*Leslie Chan[1]; Frances Groen[2]; Jean-Claude Guédon[3]*

[1] Department of Social Sciences, University of Toronto Scarborough, Toronto, Ontario, Canada
e-mail: chan@utsc.utoronto.ca
[2] Trenholme Libraries, McGill University, Montreal, Quebec, Canada
e-mail: francês.groen@mcgill.ca
[3] Département de Littérature Comparée, Université de Montréal
email: jean.claude.guedon@UMontreal.ca

## Abstract

This paper reports on the results of a feasibility study on open access publishing for humanities and social sciences journals supported by the Social Sciences and Humanities Research Council of Canada's (SSHRC) Aid to Scholarly and Transfer Journals Program. The study is part of a broader effort of the SSHRC to better understand the landscape of Open Access and how best to implement this principle into the current research programs funded by SSHRC. As such, the study was designed to assist SSHRC in making policy and program decisions regarding its Aid to Scholarly Journals Program. In particular, this study focused on the current publishing practices of SSHRC funded journals, with the ultimate goal of understanding the financial implications for these journals if they were to provide open access to the journal content. The more immediate goal of the study was to gain better knowledge of the general level of understanding among journal publishers and editors on the impact of open access and on their scholarly societies' publishing program.

**Keywords**: social sciences and humanities; open access journal; funding policy; research impact; citation analysis

## 1 Introduction

Open Access (OA) is the process by which peer-reviewed research publications resulting from public funding are made freely available through the Internet to all potential users. The purpose is to remove the price barrier and other permission barriers that restrict the dissemination and growth of further research. Though a subject of much debate, OA is now widely seen as a means to improve the accessibility and impact of publicly funded research. Evidence demonstrating that openly accessible publications are more highly cited are emerging [1] and new tools and infrastructure for maximizing the usage and innovative applications of research results are being developed, not only for the natural and medical sciences, but also for the humanities and social sciences [2].

The Social Science and Humanities Research Council (SSHRC) of Canada is the largest funding agency of humanities and social science research in Canada [3], and it is also among a growing number of government funding agencies around the world actively addressing the issue of open access [4]. In 2004, SSHRC's Council adopted OA in principle and instructed SSHRC staff to consult broadly with the research community as to the best way to implement this principle into the current research programs funded by SSHRC. SSHRC has chosen to promote open access for journals because the Council understood that open access improves scholarly communication while ensuring that research is disseminated and useful to all citizens, including the public and private sectors [5].

Between 22 August 2005 and 31 October 2005, SSHRC staff conducted a survey across a significant range of actors, including researchers and scholars, scholarly associations, publishers, editors and librarians to elicit comments and views on the subject of open access. A total of 130 submissions were received (researchers and scholars 84; university presses 2; journal editors 26; librarians 12; scholarly associations 5).

The largest and arguably most significant number of responses came from the scholarly community; and within that group 54 of the 84 expressed their support of open access although many had operational concerns. The second largest group, the journal editors, was more divided with 14 supporting in concept open access and 12 opposed. However, all expressed concerns with the financial issues in the transition to open access [6].

While the findings of this consultation are useful and important for further study, the reality remains that the input from scholars, editors and publishers is very limited in quantity. It was evident that further study would be necessary if SSHRC were to move forward on the open access agenda. It was also clear that the preservation of the integrity of the present system of scholarly communication in the humanities and social sciences had to be guaranteed, and that the transition to a new model of scholarly communication must be judiciously implemented. Over time, a valuable system has been developed for the nurturing of humanities and social sciences journals in Canada and this could be replaced by a new system only if it were of greater value.

SSHRC also recognizes that while financial support for research is crucial, the dissemination and uptake of research is equally important. Research left unread or not built upon has no impact and no financial and social return. It is time to re-examine what returns financial support for SSH journals are bringing to scholarship and to the Canadian public.

The world of publishing in general has been radically altered by the introduction of electronic publishing in the last two decades. New modes of production, of access, of ownership of information, and of financing, have been changing scholarly communication in fundamental ways. In this context and in the light of the initial SSHRC investigations, the authors of this paper were invited to conduct a study of the feasibility of open access publishing for journals currently receiving support under the Aid to Research and Transfer Journals Program of SSHRC.

While the immediate goal of the study is a gain a better understanding of current journal publishing practices and general knowledge of OA amongst SSH journal editors, the longer term goal is to provide evidence on which firm and sustainable policy on OA could be developed and implemented by the Council.

## 2 Materials and Methods

To guide the research process and to keep the scope in check, the following questions were used as guideline:

- To what extent are SSHRC funded journals already available in digital form?
- What are the costs and savings associated with the delivery of these journals in digital form?
- What are the perceived incentives and barriers to moving towards an electronic only version of these journals?
- What are the perceived incentives and barriers to open access publishing of these journals?

### 2.1 Sources of Data

To answer the questions posted above, we drawn data from a number of sources:

#### 2.1.1. Data from Funding Application

Between 2004-2007, 161 journals received funding of varying amounts from *SSHRC's Aid to Research and Transfer Journals Program*. To gain an understanding of the financial health and support resources of these journals, we first conducted a review of the records of the grant applications, which included the operating budgets of the various journals. A preliminary analysis of the journal contents and titles revealed that a broad range of topics with a number of titles are published in the fields of history, literature, law, economics and education, and a considerable number of titles in a broad range of Canadian area studies. The breadth and variety of the titles in both official languages of Canada led us to conclude that the research should not be based along disciplinary lines, but should be carried out within the broadly defined domains of social sciences and humanities.

#### 2.1.2. Online Questionnaire

A web based questionnaire in both official languages of Canada was developed and invitation to participate was sent by email to the journal editors or key contacts for the journals. Respondents' identities were kept anonymous. The questions were intended to elicit responses from editors regarding the journal's delivery medium, funding support from scholarly associations, the use of commercial aggregators, electronic publishing platform, and support and concerns towards open access. The full list of questions is provided in the appendix.

### 2.2.3 Citation analysis

To evaluate the citation impact of SSHRC funded journals on scholarship, an analysis of the journals based upon ISI Journal Citation Report was undertaken. While there are well known concerns with the using is ISI JSR to access the impact of scientific literature in general and particularly with humanities and social sciences (see Discussion), the analysis is intended to serve as a snapshot of the overall visibility of SSH journals published in Canada and how these journals compared with journals in their respective fields.

### 2.2.4 Interviews

To supplement the results from the web questionnaire and to get more in-depth and qualitative information on some of the challenges and opportunities faced by journal publishers, a number of journal editors, publishers and library directors, were selected for interviews, either in person, by phone, or through email. The interviewees were asked to provide their view on the feasibility of open access for the production and distribution of SSHRC funded journals. Their views were integrated into the discussion and recommendations put forth to SSHRC.

# 3. Results

## 3.1 Funding

The 161 journals that were successful in the 2004-7 competition received a total of $6,582,255, with grants ranging from $2,906 to $73,370 over the three year period. Of these titles, 29 journals (18% of the titles) received the maximum grant for a total of $2,127, 730 or 31% of the total funds allocated·

In addition to SSHRC support, some journals also receive support from Heritage Canada and from the Government of Quebec, based on publicly available information on the Internet. In Quebec, the Fonds de Recherche sur la société et la culture du Québec (FRSCQ) in their 2004 competition awarded $2,185,155 to 36 journals, 28 of which also received funding from SSHRC. Heritage Canada reported funding from both Canada Post Corporation and Canadian Heritage as of July 2005 and there are SSHRC titles on this list. However, amounts given to individual SSHRC titles are generally negligible.

## 3.2 Findings from the Questionnaire

A web-based questionnaire in both official languages was sent to editors of journals supported by SSHRC [7]. The survey was opened to respondents between May 1 and July 31, 2006. It received a rate of response of 42 % (67 out of 161). Of the 67 respondents, 56 were English and 11 were French speakers.

More than 80 % of the respondents reported that articles published in SSHRC funded journals are available electronically. For journals that are online, about half came online between 2002 and 2006. A small number of journals were already on-line in the 1990's beginning with 1993. The use of aggregators was highly preferred as a means of providing electronic access, with 84.4 % of English respondents reporting using a variety of aggregators, including Érudit [8], an electronic platform for journal delivery. For the 9 French responses, 100% reported using Érudit.

Of the commercial aggregators listed in the questionnaire, Proquest was the most heavily used, followed by Ebsco and Érudit. Unfortunately, Blackwell was inadvertently omitted from the list of possible aggregators due to a programming glitch, so the number of journals using Blackwell is not clear. Slightly over 40% of the 54 respondents respond that the most recent issues are available online. The rest have no recent issue available on-line. 55% of 53 English respondents reported that they do not receive compensation from an aggregator on a pay-per-use basis, while 10 of the 11 French respondents reported no compensation from aggregators.

For journals published by scholarly association, 54% of the 39 English respondents reported that the journal did not receive financial subsidy from the host association, while 7 of 9 French respondents reported the same.

When asked if they are in favour of open access in principle (leaving economic issues aside for the moment), 78% of the 54 English respondents said yes, while 6 of the 10 French respondents reported yes. With regard to the timing for open access, 74% of the 49 English respondents were in favour of the moving wall solution and 14% were for immediate open access. 91% of the 11 French respondents favoured a moving wall solution and only one respondent favoured immediate open access.

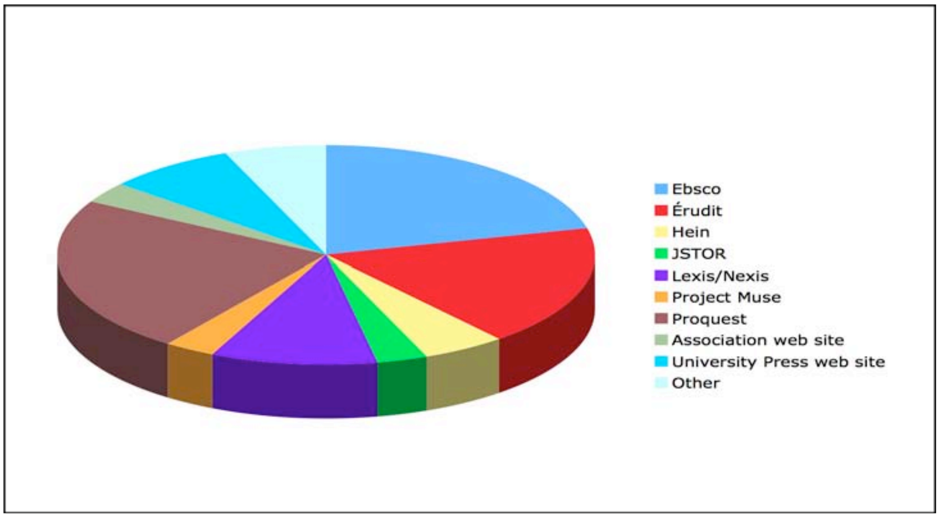


**Figure 1: Chart showing the proportion of aggregators used by SSHRC funded journals**

When asked if SSHRC should make it mandatory that SSHRC supported journals be available for open access, 84% of the 57 English respondents opposed. Similarly, 82% of the 11 French respondents were not in favour of mandatory open access.

72% of the 50 English respondents were in favour of SSHRC providing funding to support institutional repositories designed to support secure and open access to research publications. Only half of the 10 French respondents were in favour of the same.

71% of 52 English respondents supported the idea that SSHRC should provide financial support for journals to become open access. Of the 11 French respondents, 64% said yes.

82% of the 49 English respondents agreed that SSHRC should provide support for open access journals and consider eligibility criteria appropriate for these titles, while 64% of 11 French respondents agreed. Respondents were also asked to provide suggestions on what these criteria might be, and most agreed that the evaluation criteria for funding support for open access journals should be based on the same quality evaluation criteria used for subscription based journals, with the proviso that the requirement for 200 paid subscribers [9] be removed for OA journals and be replaced by other metrics more suitable for the electronic environment.

## 3.3 Findings of Citation Analysis

Impact factors have long been an essential criterion to evaluate journals, particularly journals belonging to the same specialties. It is well known that using impact factors is problematic. They must be handled with caution and they can support comparisons between journals only when they belong to closely related fields. Citation cultures can vary considerably from one discipline to another, making comparison of journals across discipline even more problematic [10]. Furthermore, in the case of the humanities, where monographs remain the dominant currency, and where citations are used in extremely complex ways, impact factors have generally not been used [11].

A preliminary survey of the titles supported by SSHRC shows that they are divided between 71 humanities journals (broadly defined) and 90 social science journals. The results given below apply only to the 90 social science journals and they must be treated prudently, but they nevertheless offer some valid insights, especially when they are used to compare journals covering roughly the same fields of study.

With these caveats in mind, we looked at the rankings of SSHRC-supported journals in the Journal Citation Reports (JCR) published by the Institute for Scientific Information (ISI). In particular, we compared the impact factors of listed journals between 1997 and 2005 and compared them to the leaders in their respective field. 1997 corresponds to the earliest year covered by JCR; 2005 is the most recent year available.

Of the 161 titles searched (90 titles are in the social sciences), and of these, only 21 titles (23 %) had an impact factor assigned by ISI in 1997 or 2005. 2 titles with an impact factor in 1997 had lost it by 2005. Conversely, 3 titles had an impact factor in 2005 but not in 1997. This means that only 19 titles had an impact factor in 2005 (or 21% of the social science titles supported by SSHRC).

| **Journal title** | **Impact factor (IF) 2005** | **Rank in assigned subject area** | **Impact factor (IF) 1997** | **Rank in assigned subject area** | **Highest impact title in subject area for 2005** | **1997 and 2005 IF as % of leader IF** |
|---|---|---|---|---|---|---|
| **Alberta Journal of Educational Research** | NIL | | 0.028[1] | 100 out of 102 | J Learn Sci 2.792 | 0% |
| **Canadian Geographer** | 0.491 | 31 out of 38 | 0.294 | 24 out of 31 | J Econ Geogr 3.222 | 15% |
| **Canadian Journal of Administrative Sciences** | 0.191 | 69 out of 71 | 0.057 | 57 out of 59 | Mis Quarterly 4.978 | 4% |
| **Canadian Journal of Agricultural Economics** | NIL | | 0.129 | 142 out of 161 | Quart J Economics 4.775 | 0% |
| **Canadian Journal of Behavioural Science** | 0.345 | 78 out of 101 | 0.348 | 64 out of 108 | Ann Rev of Psychol 9.784 | 3.5% |
| **Canadian Journal of Criminology & Criminal Justice** | 0.300 | 19 out of 27 | 0.213 | 17 out of 19 | Crime Justice 2.588 | 12% |
| **Canadian Journal of Development Studies** | 0.300 | 32 out of 38 | NIL | | J Rural Studies 2.818 | 11% |
| **Canadian Jrl of Dietetic Practice & Research** | 0.237 | 50 out of 53 | NIL | | Prog Lipid Res 11.372 | 2% |
| **Canadian journal of economics** | 0.635 | 84 out of 175 | 0.153 | 139 out of 161 | Quart J Economics 4.775 | 13% |
| **Canadian Journal of Political Science** | 0.176 | 73 out of 84 | 0.452 | 23 out of 73 | Am Pol Sci Rev 3.233 | 5.5% |
| **Canadian Journal of Sociology** | 0.383 | 63 out of 94 | 0.333 | 58 out of 95 | Am J Sociology 3.262 | 12% |
| **Canadian Journal on Aging** | 0.224 | 22 out of 24 | 0.480 | 12 out of 26 | J Gerontol A – Biol 3.500 | 6.4% |
| **Canadian Modern Language Review** | 0.304 | 37 out of 42 | 0.044 | 36 out of 40 | J Mem Lang 2.815 | 11% |
| **Canadian** | 0.648 | 52 out of | 0.426 | 51 out of 108 | Ann Rev of | 6.6% |

[1] ISI provides impact factors with an unrealistic number of decimals. In calculating the percentages, we have rounded off the results to two decimals.

| | | | | | | |
|---|---|---|---|---|---|---|
| **Psychology** | | 101 | | | Psychol 9.784 | |
| **Canadian Public Administration** | 0.067 | 25 out of 25 | 0.193 | 19 out of 24 | J Pub Admin Res Theor – 1.451 | 4.6% |
| **Canadian Public Policy** | 0.295 | 20 out of 25 | 0.200 | 18 out of 24 | J Pub Admin Res Theor – 1.451 | 20% |
| **Industrial Relations** | 1.657 | 2 out of 16 | 1.148 | 3 out of 17 | Br J Ind Relat 1.689 | 98% |
| **International Journal** | 0.284 | 38 out of 50 | 0.358 | 27 out of 50 | Int Security 2.630 | 10.8 % |
| **Isis** | 0.778 | 4 out of 29 | 0.344 | 15 out of 26 | Biol Philos 1.055 | 74% |
| **Pacific Affairs** | 0.353 | 20 out of 33 | 0.344 | 15 out of 35 | China J 1.174 | 30% |

**Table 1: The 21 titles that appear with impact factor in 1997 and/or 2005**

It must be immediately noted that no SSHRC-supported humanities journal appears in the analysis. This is not surprising: the *Journal Citation Reports* includes only two categories, science and social science; sampling the social science list of 1747 titles showed that few humanities journals were present. Only a few history and ethics journals were spotted.

For the social science titles, the following results emerged:

- The majority of SSHRC-supported journals simply do not appear in ISI.
- Those that appear, with very few exceptions, hold a very modest rank. Their impact factor compared to the leading publication in their own field is often minuscule.
- Only two titles are ranked in the top ten of their respective fields

The absence of French-language journals is not surprising given ISI's general bias in favour of English-language publications. In the case of English-language journals, the exclusion from ISI's lists means a very low status: one cannot expect them to be read very much, and, therefore, they cannot be cited very much either. This in turn probably results from a general lack of accessibility: in other words, many Canadian scholarly journals are probably not very widely available (or visible) in foreign libraries, even when they are integrated in aggregators' packages. The relative invisibility of Canadian journals also brings into question the promise that aggregators can significantly enhance the visibility of their journals. Humanities journals, for reasons already mentioned, remain excluded from this particular analysis.

Impact factors of Canadian journals have on the whole increased 31% between 1997 and 2005, but this may be due to a variety of factors, including the growth of the ISI lists across the years. As more journals are scanned by ISI, more citations are collected, which should lead to higher impact factors.

# 4 Discussion

## 4.1 Transition to e-publishing

It appears that for SSHRC-funded journals, the transition to e-publishing is well underway. However, a number of the more established journals are still available in paper only. Moreover, when editors speak about e-publishing, they may have in mind a quick-and-dirty conversion of scanned images into pdf format with (perhaps) some ability to carry out full-text searching. Some digitization operations appear to be left in the hands of aggregators who use such fast solutions in providing articles online. The issue of meta-data is rarely addressed in a lucid way. Neglecting these issues threatens long term preservation or interoperability of formats across time. If libraries have access to inferior digital files, they will not be able to participate in the preservation effort and valuable scholarship could be lost forever.

## 4.2 The Role of Aggregators or Third Party Providers

The survey showed that most on-line access to SSHRC funded journals is provided through a variety of platforms, developed by both profit and non-profit organizations that offer a variety of services and interfaces. In some cases the publisher is also the on-line provider [e.g. Blackwell]; in other cases the vendor is a third party provider who acquires the rights from the journal publisher, usually but not necessarily an association, to provide electronic access to a title via a package of electronic journals that the vendor sells to libraries [e.g. Proquest]. Usually the aggregator will offer for a fee an electronic copy of an individual article in a journal which it controls if the requester does not have access through a subscription.

The qualitative part of the survey also reveals that there is considerable confusion, on both sides of the linguistic divide, regarding electronic publishing and Open Access. This is doubtlessly the result of the ease with which journals may be accessed when an institution subscribes to an electronic package of journals. These journals appear openly accessible only to individuals who are members of a particular university community. In fact, beyond the circle of readers with access privileges, they are “toll gated.”

However it was not clear from the survey whether journals that joined an aggregator resulted in increased usage of the journal, as many reported that they did not receive any additional revenue from aggregators. Nor were aggregators generally open to providing journals with usage statistics, as journals are generally bundled into packages and licensed to libraries in complicated schemes.

## 4.3 Compensation to the Journals from Aggregators

Our survey also indicates that many of the respondents to the survey appeared to be unsure of the exact nature of their contractual relationship to the publishers and/or aggregators, for example with regard to rights ownership as well as financial compensation. In some cases, it even appears that some contractual agreements between aggregators and journals are not being fulfilled.

The use of aggregator services comes with a cost. The economics of the services are difficult to study since many of these arrangements are confidential and aggregators are reluctant to discuss them. Aggregators are important here they are widely used by SSHRC funded journals. In theory they provide value-added services to end users; and in any economic analysis represent part of the cost of scholarly communication. However, given the low citation ranking of many of the journals who also use aggregators, it is not clear whether joining an aggregator and restricting access to the journal contents represent good return on investment.

Many editors of SSHRC journals who completed the grant application forms were open about these arrangements but it is impossible to say if this is the case in all successful applications. It is reasonable to consider that the SSHRC application form should be revised to allow the journal editor to identify specifically the aggregator used and the cost arrangements that have resulted.

## 4.4 Support for a Modified form of Open Access was Strong

The result to this question is interesting in that it emanates from a set of individuals that actually wear two hats: on the one hand, editors are also researchers and they know, from that perspective, what is good for them; on the other hand, as editors responsible for the financial well-being of journals that often need careful nurturing, they are concerned about the economic effects of Open Access on their publication. This probably explains the muted agreement in favour of some modified form of Open Access, in particular the request for a moving wall, the purpose of which is to minimize financial risks for the journal due to perceived lost of subscription.

The example of many journals in the Érudit collection seems to indicate that most journal editors feel fairly confident about not losing revenue with a two-year moving wall. This looks conservative to the authors of this study. There is also the perception that unlike literature in the sciences, papers in the humanities and social sciences have longer “half-life” and therefore a longer moving wall is necessary. Currently, there is no empirical evidence to support or refute this perception.

## 4.5 Mandating Open Access Is Clearly Not Endorsed By Editors

Academics do not like being forced into anything and, even though they may favour Open Access, they are intent on preserving their ability to choose freely. Obviously, using the argument of public funding to force Open Access on journals may generate a revolt. The fear is that a forced march toward Open Access could be destructive given the uncertain financial implications.

A far more compelling case can be made on the basis of the public good that will come to the Canadian people when journals that cover topics such as adoption, mothering, social policy issues, immigration, refugees, the environment, Shakespeare and the theatre, to name just a few, are available readily to all citizens. Adult education and broad learning will advance. Such access can only improve the knowledge and well-being of the Canadian people, but the economic case for such social benefits has yet to be made and remain a important area for future study [12].

## 4.6 Support for Institutional Repositories

The apparently different attitudes of francophones and anglophones with regard to institutional repositories may be the unexpected consequence of the presence of Érudit. Since most francophone journal editors involved with Érudit seem to accept two-year moving walls, it may be that they wonder what the uses of institutional repositories are. It must also be remembered that Érudit itself incorporates a depository which further confuses the issue. On the English side, the distinction between repositories and OA journals may be a good deal clearer precisely because they are handled in very different locations: the repositories are generally in the hands of librarians while the journals are in the hand of a publisher, a scholarly association, or one (or several) aggregators.

Repositories will remain important to ensure the long-term preservation of the national scholarly heritage and librarians are very much needed in this role. It is one of their traditional functions and publishers are certainly not the best placed to take on this role. Publishers appear and disappear, while libraries remain stable. Even Elsevier has agreed to leave the preservation issue in the hands of the Royal Dutch Library. Many a small publisher of Canadian scholarly journals will disappear before Elsevier does.

## 4.7 SSHRC Funding for Open Access

The last response confirms hints and trends already noted above. The researcher part of the editor wants Open Access; the editor is willing to go there if there is no risk. Should SSHRC find the ways to finance Open Access, the probability is that most Canadian editors would follow the Brazilian SciELO model [13] and accept Open Access without any hesitation. In fact, they would welcome it as it would certainly enhance the international visibility of their publications. And once Open Access is guaranteed for the electronic version of the journal, the issue of a paying paper version can become an interesting strategy to bring revenue to associations or similar organizations. In any case, what is urgently needed here are some experiments and data gathering to properly access the economics of OA publishing and the added funding needed.

Concern also exists about SSHRC's potential intent to fund open-access journals and, in particular, the impact of funding open-access titles upon the funding of traditional journals. The fear, it appears, is to see a limited pie divided into a greater number of smaller slices.

Should SSHRC decide to finance open-access journals, maintaining quality was the essential issue from the perspective of editors; on the other hand, editors were silent about relying on the number of subscribers as a criterion of funding. It appears that, in the electronic world, especially with the various packages offered by aggregators to their customers, the evaluation of usage has to be revised and can no longer safely rest on numbers of subscribers.

## 4.8 Moving Beyond ISI Impact Factor

Results of the citation analysis suggest that authors publishing in the SSHRC-supported journals will not be readily cited given their low visibility, at least according to ISI’s JCR. With regard to impact factors, SSHRC-supported journals display characteristics similar to those observed in most journals from the developing nations. They are national journals rather than international journals, in the sense that their visibility abroad is very limited. Like journals from developing countries, SSHRC-supported journals often suffer from a vicious circle: low impact factors induce low submission rates of generally less significant articles that attract little attention and, therefore, few subscriptions. In other words, and, given ISI's claim that they select the best journals in any given field, this survey raises the general issue of perceived quality and most important, usage of SSHRC-supported social science periodicals (including law journals). The survey also raises the question regarding return on SSHRC’s investment as journal articles that are not widely read and cited translate into low research uptake and impact. The question that SSHRC must address is whether it makes sense to implicitly encourage journals to close off access to the content for the sake of a limited number of subscribers, number that are required by SSHRC's funding criteria. Or does it make more sense to trade-off the limited economic return from subscription with a potentially much larger return on readership, which may in turn leads to higher submission, usage and visibility.

Given the fact that most journals supported by SSHRC do not have impact factors, another issue emerges: what alternate evaluation criteria should be applied to these journals, particularly non-subscription based open access journals, applying in the next funding round? Obviously, when titles are available in electronic format on the Internet, new kinds of metrics can be applied, such as hits, downloads, links and, of course, citations. Development and implementation of such new indexes for the evaluation of open access journals is clearly a priority for the scholarly community and for SSHRC. In this regard it is encouraging to see the growing number of studies and projects that aim to provide alternative and better measurement and metrics of usage and research impact, particularly for literature that are openly available [14].

## 5 Conclusions

The results of the study indicate that many of the journal editors understand that providing Open Access will greatly improve the visibility and citation impact of their journals. However, many editors also worry about the financial conditions under which the transition to OA can be managed. As it stands, the return on research investment, at least as measured in citation counts, is poor for most of the SSHRC subsided journals. Providing a special source of funding to offset possible losses of subscription revenue could become a strong incentive to move toward Open Access. Given that many of these journals have small subscription revenues, the needed financing, which could take the form of a kind of insurance policy, ought to be quite limited, but the precise amount is difficult to determine at this stage and a separate study would be required.

There is also considerable consensus that SSHRC should support open access journals and encourage journals that wish to experiment with conversion to Open Access to work collectively in a SSHRC-supported experiment designed to better understand the financial implications, author's uptake, and usage of publication before and after becoming Open Access. The experimentation will provide better data to gauge the financial viability of Open Access publishing. These results will be useful in turn to examine whether scaling up the process to a larger number of journals is desirable [mention the new funding program in a footnote?].

Perhaps the most valuable consequence of this study has been the important recognition that there is no magic way to move into electronic publishing and Open Access. Testing, exploring and experimenting while consulting and dialoguing should be the principles under which any kind of action plan should be undertaken.

With regard to electronic publishing, environmental pressures as well as various forms of inducements on the part of aggregators or some publishers have led to a transition carried out in such a wide variety of ways that "chaos" might well be the best term to describe it. In the process, SSHRC is finding itself subsidizing some very profitable commercial aggregators, while denying support to some innovative Open Access journals that are deserving of help except for the fact that they do not have any paying subscribers.

A good reason for this chaotic transition to e-publishing may well have been the consequence of the inability to create orderly experiments so as to identify best practices and enhance the sharing of new know-how. Only in Québec has there been the semblance of an organized move toward electronic publishing [15], but it may have been done in such a centralized manner that it may not fit the ethos of the rest of the country. Nonetheless, it remains a valuable source of experience. Elsewhere, there are dispersed endeavours to produce electronic journals, most of the time on tiny scales [16]. On the non-commercial front, only a very few university presses have developed in-house capacity in this regard. Whether they are willing to share this know-how is far from obvious.

Finally, it seems clear that SSHRC must take on a leadership position in this regard. As other granting councils in the USA and in Europe (particularly the UK, Germany and France) have amply shown, this is to SSHRC's advantage. More fundamentally, if SSHRC does not show some national leadership, no one else will do so, except perhaps in the form of some bid to become the monopolistic device for SSH publishing in the country. Clearly, no one wants this outcome. No one wants one university press, or, even worse, a large commercial press to become the sole provider of e-publishing services to Canadian SSH. At the same time, it is clear that the emerging digital environment is challenging all publications to globalize in an effective manner.

It is natural that journals in similar disciplines be grouped together so that a particular journal platform tends to become well known for its coverage in, for example, economics or law, and a number of journals from a variety of publishers in many countries might be housed in that disciplinary platform. But is the notion of a national platform be useful to researchers used to work on well-focused issues with information coming from all over the planet? In other words, how does one reconcile the idea of a national strategy for scientific and scholarly publishing with the universal characteristics of validated knowledge? These questions lie beyond the scope of the present study, but they are part of the changing landscape of Canadian scholarly communications as it impacts SSHRC-supported journals and they should not be neglected.

## Afterword

We are happy to report that in late March 2007, SSHRC announced a new one-year experimental program in support of open access journals [17]. The program adopted several recommendations from our initial report submitted to SSHRC in August 2006 [18]. Amongst the key innovations of the program is the adoption of usage based metrics and cost per article as basis for funding. Of course peer review and the expertise of the editorial board remain as primary quality criteria, but the addition of alternative usage and impact metrics should allow innovative open access to gain the funding support they deserve. We eagerly await the outcomes of this experimentation and we hope this program will generate the much needed economic and usage data for better planning and support of a broader range of open access journals in the humanities and social sciences.

## Acknowledgements

We would like to express our thanks to all the journal editors who participated in the online questionnaire. We also like to thank the following individuals who generously gave their time for interviews: Alan Burke, Lynn Copeland, Ann-Marie Corrigan, Carlin Craig, Gunther Eysenbach, Rowley Lorimar, Carol Moore, Brian Owen, and John Willinsky. This study was supported by a Social Science and Humanities Research Council's President's Grant.

## Notes and References

[1] ANTELMAN, K. (2004). Do open-access articles have a greater research impact? *College & Research Libraries*, 65(5): 372-382. (Online). Accessed Jan. 15, 2007, from http://eprints.rclis.org/archive/00002309/01/do_open_access_CRL.pdf.
See also the regularly updated bibliography of citation impact studies maintained by Steve Hitchcock: http://opcit.eprints.org/oacitation-biblio.html , accessed March 1, 2007

[2] Following the NSF funded report on Cyberinfrastructure for the natural sciences, the American Council of Learned Societies and the Mellon Foundation also funded a parallel study supporting the development of cyberinfrastructure for the humanities and social sciences: http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf

[3] http://www.sshrc-crsh.gc.ca/

[4] See the The ROARMAP list of the strongest funder and university policies: http://www.eprints.org/openaccess/policysignup/

[5] SSHRC's original position on OA is no longer available on its web site http://www.sshrc.ca/web/about/council_reports/news_e.asp, first accessed in August 2004. But SSHRC's position on OA is even more clearly stated in the context of its recent announcement (March 29, 2007) on a new "Aid to Open Access Research Journals" funding program: http://www.sshrc.ca/web/apply/program_descriptions/open_access_journals_e.asp , accessed April 1, 2007

[6] The survey result was made available by David Moorman, Senior Policy Advisor at SSHRC, at a meeting on March 9, 2006: http://open.utoronto.ca/index.php?option=com_content&task=view&id=234&Itemid=226

[7] SurveyMonkey, www.surveymonkey.com, was used to administer the questionnaire

[8] Érudit is a digital publishing and dissemination platform that originated at Les Presses de l'Université de Montréal in 1998 and has since evolved into a network that support a large variety of journals, mostly from the province of Quebec. http://www.erudit.org/

[9] The 200 existing subscribers was one of the requirements for journals to qualify for SSHRC funding. However, this clearly exclude journals that are already open access but still require financial support. It was also clear that some journals that were affiliated with a scholarly association were using the association's membership to inflate the number of subscribers, thereby qualifying them for the grant.

[10] ARCHAMBAULT, E.; VIGNOLA-GAGNE, E; COTE, GREGOIRE; LARIVIERE, V; GINGRAS, Y. (2006) Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics* 68(3):329-342

[11] HICKS, D. (2005). The four literatures of social sciences. In Handbook of Quantitative Science and Technology Research. Edited by H. F. Moed, W. Glänzel and U. Schmoch. Page 473-496. Springer Netherlands.

[12] Studies on the enhanced economic benefits of Open Access on research and development have recently been made, for example: HOUGHTON, J.W., STEELE, C. AND SHEEHAN, P.J. (2006) Research Communication Costs in Australia, Emerging Opportunities and Benefit*s,* CSES Working Paper No. 24, Centre for Strategic Economic Studies, Victoria University, Melbourne. Available http://www.cfses.com/documents/wp24.pdf ; HOUGNTON, J.W., SHEEHAN, P.J. (2006) The Economic Impact of Enhanced Access to Research Findings, CSES Working Paper No. 23, Centre for Strategic Economic Studies, Victoria University, Melbourne. Available http://www.cfses.com/documents/wp23.pdf . But the social benefits of OA to scholarly literature have yet to be well studied and documented.

[13] SciELO stands for Scientific Electronic Library Online www.scielo.br . It is a pioneering project in providing open access to scientific journals published in Brazil, and now from other Latin American countries.

[14] BRODY, T.; HARNAD, S.; CARR, L. (2006) Earlier Web Usage Statistics as Predictors of Later Citation Impact. Journal of the American Association for Information Science and Technology (JASIST) 57(8):1060-1072. Available: http://eprints.ecs.soton.ac.uk/10713/

HARNAD, S. (2007) Open Access Scientometrics and the UK Research Assessment Exercise. In Proceedings of 11th Annual Meeting of the International Society for Scientometrics and Informetrics (in press), Madrid, Spain. Available: http://eprints.ecs.soton.ac.uk/13804

BOLLEN, J.; VAN DE SOMPEL, H.; SMITH, J.;LUCE, R. (2005). Toward alternative metrics of journal impact: a comparison of download and citation data. Information Processing and Management, 41(6):1419-1440. Available: http://public.lanl.gov/herbertv/papers/ipm05jb-final.pdf

BOLLEN, J.; VAN DE SOMPEL, H.(2006). Mapping the structure of science through usage. Scientometrics, 69(2):227-258. Available: http://library.lanl.gov/cgi-bin/getfile?LA-UR-05-7070.pdf

[15] We are referring to Érudit, see note number 8.

[16] Just exactly how many self-started open access journals in the humanties and social sciences produced in Canada is not known and a systematic survey of these journals and their funding and editorial practices would be an important future project.

[17] Details of the funding program, eligibility criteria, and adjudication process are available on SSHRC's web site. Accessed April 1, 2007 http://www.sshrc.ca/web/apply/program_descriptions/open_access_journals_e.asp

[18] http://www.sshrc.ca/web/about/publications/journals_report_e.pdf

## Appendix I: Questionnaire Sent to Journal Editors or Key Contacts

A. Transitioning to electronic publishing

1. Are the articles in your journal available electronically on the Internet?
Yes

No

If “no”, please skip to section B.

2. When did your journal become available on-line?

3. Is your journal available electronically through an aggregator or a portal?

Yes

No

4. If “yes” to question 3, please specify which aggregator or portal you are using:

Blackwell

Ebsco

- Érudit
- Hein
- JSTOR
- Lexis/Nexis
- Project Muse
- Proquest
- Association web site
- University web site
- University Press web site
- Other (please specify)

5. Are all issues, including the most recent, available on-line?

Yes

No

6. If you answered “no” to question 5, please specify which years have been digitized.

7. Who owns the digital rights to your journal?

B. Questions regarding your publisher

8. If the publisher of your journal is not your scholarly association, is your publisher financing part of the activities related to the publication of your journal? (for example, editorial stipend, peer review process, etc.)?

Yes

No

9. Does your journal receive compensation from an aggregator on a pay-per-use basis?

Yes

No

C. The issue of Open Access

10. The Open Access movement: Putting peer-reviewed scientific and scholarly literature on the internet. Making it available free of charge and free of most copyright and licensing restrictions. Removing the barriers to serious research. “Open Access News”, http://www.earlham.edu/~peter/fos/fosblog.html.

11. Are you in favour of Open Access in principle (leaving economic issues aside for the moment)?

Yes

No

12. In order to provide Open Access to your journal, you will have to devise a new business plan for your journal. Which business plan would you favour?

13. All issues immediately available, including the latest (true Open Access)?

A “moving wall” with the latest issues available only through subscriptions, and the earlier issues available in Open Access?

A publishing fee for all accepted articles?

A choice between “b” and “c” offered to authors according to their ability/willingness to use funds from various sources to publish?

14. Should SSHRC provide financial support for journals to become Open Access and non-subscription based?

Yes

No

15. Should SSHRC provide support for Open Access journals and consider eligibility criteria appropriate for these titles

Yes

No

16. If you have suggestions or comments on what these criteria might be, please list them here.

# The Research Impact of Open Access Journal Articles

*Yaşar Tonta; Yurdagül Ünal; Umut Al*

Department of Information Management, Hacettepe University
06532 Beytepe, Ankara, Turkey
e-mail: {tonta, yurdagul, umutal}@hacettepe.edu.tr

## Abstract

The availability of scientific and intellectual works freely through scientists' personal web sites, digital university archives or through the electronic print (eprint) archives of major scientific institutions has radically changed the process of scientific communication within the last decade. The "Open Access" (OA) initiative is having a tremendous impact upon the scientific communication process, which is largely based on publishing in scientific periodicals. This exploratory paper investigates the research impact of OA articles across the subject disciplines. The research impact of OA articles as measured by the number of citations varies from discipline to discipline. OA articles in Biology and Economics had the highest research impact. OA articles in hard, urban, and convergent fields such as Physics, Mathematics, and Chemical Engineering did not necessarily get cited most often.

**Keywords:** open access articles; research impact; scholarly communication; citations analysis

## 1 Introduction

There are some 24,000 scientific journals publishing 2.5 million articles each year. Scientific journals are expensive. The economic model of publishing is based on subscription and licensing. Price hikes in the publishing sector within the last 30 years are well beyond the inflation rates. This has been primarily due to lack of competition. Some publishers can easily become monopolies, as no two journals can publish the same article in view of copyright restrictions. Moreover, those who use the scientific journals (scientists) and those who pay for this service (usually libraries) are different, which results in what is called the "price inelasticity" in economics and empowers the scientific journal publishers further [1]. As scientific journal prices increase, some libraries cancel some of their subscriptions because they cannot afford the price hikes. Publishers then increase prices further to make up the lost income. Consequently, some more libraries discontinue their subscriptions. In response, to make up the lost income, publishers increase the prices again. This vicious circle is not only the main cause of the so called "serials crisis," but also it affects the scientific communication process. Interestingly, the lack of competition in scientific journal publishing enables some publishers to increase their market shares by increasing prices. When the price of an already expensive journal is further increased, libraries tend to cut off subscriptions to cheaper but prestigious journals in order to keep the more expensive ones [2].

Scientific research and its outcome (e.g., scientific journal articles) get supported primarily by public money. Articles are given by scientists to commercial publishers free of charge and refereed by scientists free of charge. Yet, the same scientists pay dearly, through their libraries, to subscribe to the very same journals despite the fact that their salaries are paid for by public monies and their libraries are supported by public funds. The triple payment of public money to support research projects, to pay for salaries of scientists, and to fund libraries is emphasized by the following comment: "What other business receives the goods that it sells to its customers from those same customers, a quality control mechanism provided by its customers, and a tremendous fee from those same customers?" [2]. Universities and governments have recently begun to scrutinize the scientific communication process. Web access to research articles created new opportunities and showed that alternative or complementary economic models can be experimented with [3, 4].

One of these models is what is called Open Access (OA). OA is defined as "free (...) access to" scientific publications. "A complete version of the work (...) is deposited (and thus published) in at least one online repository (...) maintained by an academic institution, scholarly society, government agency, or other well-established organization that seeks to enable open access, unrestricted distribution, inter operability, and long-term archiving" [5]. OA increases the research impact by making articles available, free of charge, to all those interested. Two parallel and integrated strategies to create a more effective and equitable scientific communication process are suggested: (1) researchers "self-archiving" their articles that are published in refereed journals in their web sites or institutional repositories and making them available through the Internet; and (2) researchers publishing their articles in OA journals. More than 90% of commercial publishers support self-archiving. There are currently more than 2,500 OA journals published in all subjects.

Several prominent institutions including OECD and UN support OA. Recently, some universities decided to mandate researchers to self-archive their published articles. A bill (Federal Research Public Access Act) mandating OA to publicly-funded scientific publications in the United States is likely to become enacted in the near future. The European Commission (EC) recommends OA to EC-funded research reports [6, p. 87]. Governments allocate billions of dollars of taxpayers' money to research. For instance, the annual budget (28 billion dollars) of the US National Institute of Health alone is higher than the GDP of 142 nations [7]. OA increases the impact of the publicly-funded research and triggers new research projects, thereby increasing the return on investment [8-11].

In this paper we look into the research impact of OA journal articles in sciences, social sciences, and arts and humanities. The term "research impact" in this study is defined as the number of times that each article is cited in the literature. Journal articles representing nine disciplines were selected from the Directory of Open Access Journals (www.doaj.org). Citations to each article were identified through Elsevier's Scopus. The research impact of articles in different disciplines was compared to find out the underlying trends. Findings were discussed in light of why OA is supported in varying degrees in sciences, social sciences, arts and humanities.

## 2 Literature Review

It has for long been observed that scientific communication processes differ in sciences, social sciences, and arts and humanities. While scientists publish their contributions primarily in journals as articles, social scientists and scholars of arts and humanities prefer monographs as the main outlet of their contributions. Whereas journal articles constitute 90% of all publications in sciences, books and monographs in social sciences constitute 40% of all publications [12]. The intensity of production also differs from discipline to discipline. In chemistry it is not uncommon for a researcher to produce several journal articles in a given year whereas a social scientist would publish a single article perhaps every other year or so. Some social scientists and humanities scholars may not even bother to publish journal articles but concentrate on publishing a few monographs instead throughout their academic careers. "Disciplinary cultures" have an impact on scholarly communication processes and the ways by which researchers in each discipline communicate their findings [13].

The emergence of the Internet and electronic publishing in the early 1990s has profoundly changed the scientific communication patterns. While physicists and computer scientists, for instance, reacted very quickly and began to use electronic publishing as a means of disseminating research results over the Internet, social scientists and arts and humanities scholars were somewhat slow to react. For some researchers the acceptance of electronic publishing in support of scientific communication was "not just a matter of time": field differences have to a large extent determined the acceptance levels [14]. Electronic publishing is seen as a transitory period by some researchers, for example. Some do not trust the electronic media while others see electronic journals inferior compared to printed journals. Copyright concerns discourage some researchers. Reasons are too numerous to discuss in detail here. These cultural issues shape the scholarly communication and explain the degree of use of electronic journals across the fields [15].

Field differences and disciplinary cultures also played an important role in OA movement since mid-1990s. Similar concerns shied away some researchers from self-archiving their contributions through their personal web sites or institutional archives. While almost all articles in sciences (e.g., physics and mathematics) have currently been open access, the percentages are much lower in social sciences, arts and humanities (e.g., 60% in economics, 25%-30% in political science, psychology and sociology, and less than 20% in anthropology and geography) [16, p. 88]. Only 5% of social scientists self-archive their papers.

As mentioned earlier, OA makes scientific papers more visible and increase their research impact [8-11]. OA articles get cited more often by other researchers, thereby bringing their authors more recognition and prestige, and providing them incentives to do more research. The *Proceedings of the National Academy of Sciences* (PNAS) is a prestigious journal with an high impact factor (IF) publishing both OA and non-OA articles. OA articles published side by side with non-OA articles at *PNAS* were cited more quickly and twice as many times than non-OA articles [17]. This finding is somewhat contradictory with that of an earlier study [18] that analyzed the impact factors and citation patterns of OA journals in ISI databases and found that OA journals usually have lower IFs than non-OA journals in their subject categories. It appears that OA articles help increase the IF of a prestigious journal even further.

Earlier studies tended to measure the research impact of OA journal articles mainly by using the Web of Science (WoS) database of ISI (now Thomson Scientific). WoS at that time did not index that many OA journal titles. The situation has changed in 2004, however. Elsevier's Scopus and Google's Google Scholar (GS) citation databases were introduced almost at the same time in November 2004. These databases track citations that come from refereed journals as well as those from resources available on the Web. The overlapping citations between WoS and Scopus, and WoS and GS are not as high as one would expect (58% and 31%, respectively, for articles in library and information studies) [19]. Scopus covered the library and information studies (LIS) literature more comprehensively and retrieved 26% unique citations that were not retrieved by WoS. The percentage of unique citations retrieved by GS was somewhat lower (21%). In different studies, WoS retrieved higher citation counts for articles that were published in 1985 in the Journal of American Society for Information Science and for articles in oncology and condensed matter physics in 1993 than Scopus and GS [20, 21]. This is primarily due to the fact that the WoS database goes back to 1900s while the Scopus database cover citations since 1996. (Information is not available for GS.) Jacso [22] reviewed these three citation databases in more detail and compared them in terms of their major features such as database subject coverage and composition, number of records, and search and retrieval characteristics.

# 3 Research Questions

As reviewed earlier, the research impact of both OA and non-OA articles has been addressed in the past. There is a considerable difference between scientific disciplines in terms of both the rates of research impact and the acceptance of OA as a means of dissemination of research results. Antelman [11] found that OA articles in mathematics and electrical and electronics engineering have a greater research impact than that in political science and philosophy. In a different study Antelman [16] identified different degrees of acceptance of self-archiving in six social science disciplines (economics, sociology, geography, political science, anthropology, and psychology). Based on Becher and Trowler's [13] and Whitley's [23] studies, she posited that "differences between disciplines can be characterized in terms of the degree of mutual dependence between researchers and the degree of task uncertainty in defining shared problems, goals, and procedures" [16, p. 92]. The interdependency in social science disciplines is low and common issues and objectives are defined ambiguously. Moreover, the rates of self-archiving practice were found lower in divergent social science disciplines that concentrate on rural issues (e.g., anthropology, geography, sociology and psychology) and higher in convergent ones that concentrate on urban issues and have close relationships with other disciplines (e.g., economics) [16, p. 92].

Antelman's interpretation of her findings seems interesting. If such a relationship between self-archiving rates and different scientific disciplines exists, one would think that a similar relationship may also hold true for varying degrees of research impact of OA articles in different fields. This paper aims to explore the conjecture that OA articles in the interdependent, convergent and urban disciplines would have higher research impact than that of independent, divergent and rural disciplines.

What is meant by hard/soft, urban/rural, and convergent/divergent fields is that "Physics represents hard science, which is convergent and urban in its social aspects; history is a soft discipline, relatively convergent and rural; sociology is a soft, divergent, and rural discipline; whereas biology is both mostly rural science, and also a mixture of soft and hard elements" [24 p. 68].

Nine fields under three groups were identified along this continuum of hard/soft, urban/rural, convergent/divergent and interdependent/independent scientific fields. In the first group, physics, mathematics, and chemical engineering represent hard and applied sciences that are convergent and urban in their social aspects. In the second group, economics, biology, and environmental science represent disciplines that have both hard and soft components. Economics is a more urban discipline than both biology and environmental science in this group. In the last group, sociology, psychology and anthropology represent soft, divergent, and rural disciplines. According to Whitley's [23] dimensions, disciplines in the first group have "high degree of mutual dependence and low degree of task uncertainty" while the ones in the last group have the opposite. The disciplines in the last group lie somewhere in between.

This paper addresses the following research questions:

- Does the research impact of OA articles differ across the fields in sciences, social sciences, and arts and humanities?
- If it does, do OA articles in hard, urban and convergent fields receive more citations (hence higher research impact) than those in soft, rural, and divergent ones?

# 4 Methodology

What follows is a detailed account of the sampling process of articles published in OA journals. The Directory of Open Access Journals (DOAJ, www.doaj.org) lists more than 2,500 OA journal titles. It was used to select OA journals representing nine disciplines (physics, mathematics, chemical engineering, economics, biology, environmental science, sociology, psychology and anthropology). The detail of each journal title (subject, year, language) was recorded (January 2007). Non-English journal titles and titles that did not have enough back issues (since 1999) published were excluded from the sample frame. DOAJ (www.doaj.org) assigns one or more subject headings to each journal title. Journal titles with a single subject heading were preferred.

Journal titles not covered by Elsevier's Scopus were excluded since Scopus was used to identify citations that each selected article received (more below). It was noted in the Scopus web site (info.scopus.com) that Scopus is the largest abstract and citation database of research literature containing 29 million abstracts from about 15,000 peer-reviewed journal titles in all fields along with 265 million citations. Abstracts and citations go back to 1966 and 1996, respectively.

The total number of articles published in OA journals in 1999, 2001 and 2003 were identified for selected nine disciplines. A sample of 30 articles was selected to represent each discipline, thereby making a total of 270 articles for all nine disciplines. Needless to say, sampling intervals were different for each discipline. As the number of OA journals in each discipline varied, articles in the samples for some disciplines came from a few journals (e.g., anthropology). Similarly, the number of articles published in some disciplines were much higher (e.g., physics), thereby making the sampling rates uneven across fields (Table 1).

| Subjects | # of journals in DOAJ | # of journals in the sample | # of total articles in OA journals | # of OA articles taken from the sample journals | sample rate |
|---|---|---|---|---|---|
| Physics | 23 | 6 | 2,543 | 30 | 1.2 |
| Mathematics | 77 | 16 | 1,092 | 30 | 2.7 |
| Chemical Engineering | 6 | 3 | 818 | 30 | 3.7 |
| Economics | 36 | 2 | 113 | 30 | 26.5 |
| Environmental Sciences | 12 | 3 | 247 | 30 | 12.1 |
| Biology | 50 | 7 | 690 | 30 | 4.3 |
| Psychology | 45 | 4 | 271 | 30 | 11.1 |
| Sociology | 33 | 3 | 97 | 30 | 30.9 |
| Anthropology | 22 | 2 | 111 | 30 | 27.0 |
| Total | 304 | 46 | 5,982 | 270 | 4.5 |

**Table 1: Sampling statistics**

All 270 articles were searched on Scopus for citations (March 2007). Retrieval results were entered into SPSS, a statistical analysis software. The number of citations, citing authors and journals along with years, and self-citations were recorded for each article. The citation age of each article was calculated. Various statistical tests were run using SPSS.

# 5 Findings

Table 2 provides descriptive statistics about citations that 30 OA articles in each subject discipline received. All OA articles (N = 270) were cited 761 times ( $\overline{X}$ = 2.8, SD = 4.7). The average number of citations per OA article ranged between 0.8 (Sociology) and 6.4 (Biology), although the distributions of citations for all disciplines were rather skewed (note the standard deviations being always higher than the averages). OA articles in Biology and Economics received almost half of all citations (25.2% and 20.2%, respectively) whereas the ones in Psychology and Sociology did much fewer (3.7% and 3.2%, respectively).

| Subjects | # of OA articles | # of citations | % | $\bar{X}$ | SD | # of OA articles with zero citations | median | max |
|---|---|---|---|---|---|---|---|---|
| Physics | 30 | 95 | 12.5 | 3.2 | 3.7 | 9 | 2 | 16 |
| Mathematics | 30 | 44 | 5.8 | 1.5 | 1.9 | 11 | 1 | 7 |
| Chemical Engineering | 30 | 63 | 8.3 | 2.1 | 3.2 | 12 | 1 | 16 |
| *Subtotal* | *90* | *202* | *26.5* | *2.2* | *3.1* | *32* | *1* | *16* |
| Economics | 30 | 154 | 20.2 | 5.1 | 7.5 | 6 | 2.5 | 39 |
| Environmental Sciences | 30 | 63 | 8.3 | 2.1 | 2.8 | 12 | 1 | 13 |
| Biology | 30 | 192 | 25.2 | 6.4 | 7.4 | 2 | 4.5 | 38 |
| *Subtotal* | *90* | *409* | *53.7* | *4.5* | *6.5* | *20* | *2.5* | *39* |
| Psychology | 30 | 28 | 3.7 | 0.9 | 1.4 | 17 | 0 | 5 |
| Sociology | 30 | 24 | 3.2 | 0.8 | 1.3 | 20 | 0 | 5 |
| Anthropology | 30 | 98 | 12.9 | 3.3 | 5.3 | 6 | 2 | 26 |
| *Subtotal* | *90* | *150* | *19.7* | *1.7* | *3.4* | *43* | *1* | *26* |
| Grand Total | 270 | 761 | 100.1 | 2.8 | 4.7 | 95 | 1 | 39 |

Note: The percentage is not equal to 100% due to rounding.

**Table 2: Citation statistics of open access articles in different fields**

OA articles in the second group of fields received more than half (53.7%) of all citations, followed by the first group (26.5%) and the third group (19.7%). The second group of fields (Economics, Environmental Sciences, and Biology) that have both hard and soft components scored a much higher research impact than either the first group of fields (hard, convergent and urban) and the third group of fields did. The number of citations for each field within groups also differed. For instance, OA articles in Biology and Economics in the second group received much higher citations than that in Environmental Sciences. The difference was even more substantial for OA articles in Anthropology in the third group: they received about four times more citations than that in Sociology and Psychology.

The average self-citation rate for all subjects was 28.4% (216/761). Self-citation rates were much higher in Mathematics (45.5%) and Physics (43.2%) than that in Psychology (7.1%) and Economics (13.6%). More than one third (35%) of OA articles (95/270) were never cited at all. OA articles in Sociology and Psychology had the highest zero citation rates (67% and 57%, respectively) whereas only two out of 30 articles (7%) in Biology went uncited. About 17% (or 45 articles) were cited only once, 15% (40 articles) twice, 7% (20 articles) three times, and a further 26% (70 articles) four or more times. Two OA articles in Economics and Biology received the highest number of citations (39 and 38, respectively). The most-cited 10 OA articles collected 27% (209/761) of all citations (Table 3).

| Ran k | Authors (Publication Year). Article title. *Journal.* | # of times cited in Scopus | Subject |
|---|---|---|---|
| 1 | Berg, A., & Pattillo, C. (1999). Are currency crises predictable? A test. *IMF Staff Papers*. | 39 | economics |
| 2 | Lyubarsky, A.L. et al. (2001). RGS9-1 is required for normal inactivation of mouse cone phototransduction. *Molecular Vision*. | 38 | biology |
| 3 | Nishida, T., Kano, T., et al. (1999). Ethogram and ethnography of Mahale chimpanzees. *Anthropological Science*. | 26 | anthropology |
| 4T | Plascak, J.A. et al. (1999). Phenomenological Renormalization Group Methods. *Brazilian Journal of Physics*. | 16 | physics |
| 4T | Ishida, H. et al. (1999). New hominoid genus from the Middle Miocene of Nachola, Kenya. *Anthropological Science*. | 16 | anthropology |
| 4T | Miura, M. (1999). Detection of chromatin-bound PCNA in mammalian cells and its use to study DNA excision repair. *Journal of Radiation Research*. | 16 | biology |
| 4T | Yu, Q. et al. (2001). Retinal uptake of intravitreally injected Hsc/Hsp70 and its effect on susceptibility to light damage. *Molecular Vision*. | 16 | biology |
| 4T | S.P. Asprey & Naka, Y. (1999). Mathematical Problems in Fitting Kinetic Models—Some New Perspectives. *Journal of Chemical Engineering of Japan*. | 16 | chemical engineering |
| 9T | Blanchard, O. & Shleifer, A. (2001). Federalism with and without political centralization: China versus Russia. *IMF Staff Papers.* | 13 | economics |
| 9T | Casey, T.G. et al. (1999). Metabolic behaviour of heterotrophic facultative aerobic organisms under aerated/unaeratedconditions. *Water SA*. | 13 | Environmen-tal sciences |

**Table 3: The 10 most-cited open access articles**

Articles in the sample came from 46 different OA journals across the fields. Fifteen articles that appeared in 7 OA journals in different fields (Environmental Sciences, Mathematics, Physics, and Psychology) received no citations while 7 articles appeared in 7 OA journals (6 in Mathematics, 1 in Physics) received only one citation each (see Appendix). In addition to the Scopus database, half (23) of those OA journal titles were also listed in Thomson Scientific's Web of Science (WoS) citation database. There was no difference, however, between the articles listed in the Scopus database only and that listed in both Scopus and WoS databases in terms of the number of citations they received ($\chi^2_{(21)}$=.382, p = .396).

More than 60% of all citations to OA articles were received within the first three years after their publication (Figure 1). OA articles got cited in the literature less often after three years. The "half-life" (the time it takes to receive half of all citations) was 2 years for OA articles in Physics, Mathematics, Biology, and Psychology, and 3 years in Chemical Engineering, Economics, Environmental Sciences, Sociology and Anthropology.

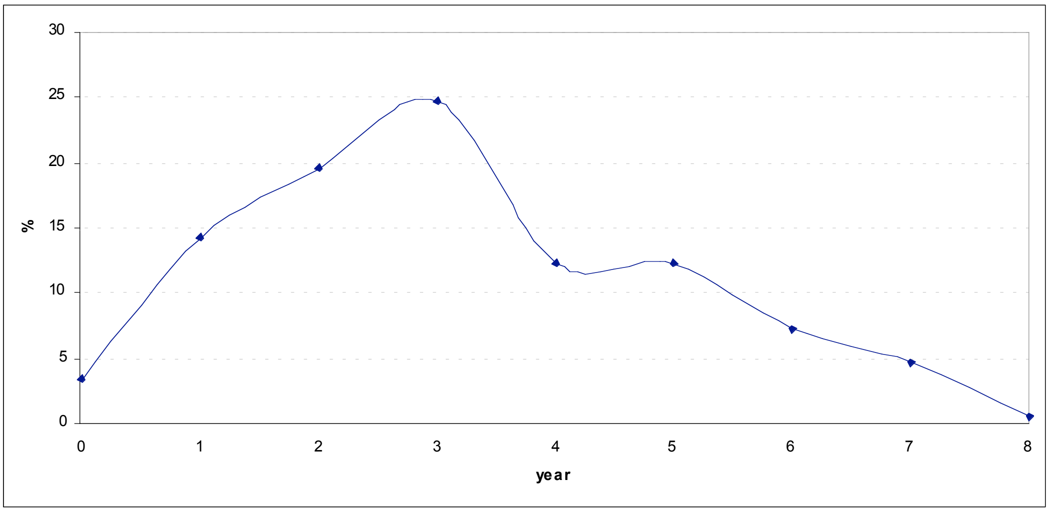


**Figure 1: Temporal distribution of citations to open access articles after publication (in years)**

# 6 Discussion

This study confirmed the findings of earlier ones in that the research impact of OA articles differ across the fields. Some subtle differences were observed, however, in terms of the research impact of certain disciplines (e.g., mathematics and anthropology). Antelman [11] found that mathematics had a greater research impact than some social science disciplines (e.g., political science). Yet, OA articles in Mathematics received much fewer citations in the present study and almost half of them were self-citations. Usually, articles in social sciences and humanities get cited much less often. OA articles in Economics and Anthropology were among the most heavily cited ones (after those in Biology).

Such variations in research impact across the fields may be susceptible to the small sizes of samples (30 articles) for each subject discipline and the uneven distribution of sampled articles to journals in respective fields. For instance, OA articles in Mathematics came from 16 different journals, more than half of which received either zero or one citation only (average being 1.5 citations). On the other hand, those in Economics and Anthropology came from two journals in each subject and they collected relatively higher number of citations per article (averages being 5.1 for Economics and 3.3 for Anthropology). This may perhaps be explained by the research impact of articles that appeared in prestigious OA journals in Economics (IMF Staff Papers, Asian Development Review in Economics) and Anthropology (Anthropological Science, and Journal of Physiological Anthropology and Applied Human Science).

The main objective of this paper was to explore if there is a relationship between the research impact of OA articles and the characteristics of the subject fields (e.g., hard/soft, urban/rural, and convergent/divergent). Findings do not seem to indicate any discernible pattern between these two variables. In other words, OA articles in hard, urban and convergent fields such as Physics, Mathematics, and Chemical Engineering did not necessarily have higher research impact than those that have both hard/soft and urban/rural components such as Biology and Economics. In fact, it was just the opposite: OA articles in the second group (Economics, Environmental Sciences, and Biology) received twice as many citations than those in the first group did. OA articles in soft and divergent fields concentrating on rural issues (e.g., Sociology and Psychology) had lower research impact as expected. Although in the same group with Sociology and Psychology, OA articles in Anthropology had higher research impact than all the subjects in the first group (Physics, Mathematics, and Chemical Engineering) and Environmental Sciences in the second group.

Recall that the research question in this study emerged from Antelman's [16] findings on self-archiving rates in different social science disciplines (higher in convergent and urban fields such as Economics, and lower in divergent and rural fields such as Anthropology, Geography, Sociology and Psychology). We hypothesized implicitly that OA articles in hard, urban and convergent fields receive more citations (hence higher research impact) than those in soft, rural, and divergent ones. It appears that the research impact of OA articles in Economics, Sociology and Psychology resembles the behavior of self-archiving. The research impact of OA articles in Anthropology is quite different, however. Moreover, the research impact of hard, urban and convergent fields (Physics, Mathematics, and Chemical Engineering) have no resemblance whatsoever to self-archiving practices. It may well be that self-archiving and research impact measured by the number of citations are two completely different things. It is also highly likely that, as we indicated earlier, the small sample sizes of OA articles in each subject did not allow any trends to emerge. The hypothesis needs to be tested using much larger samples with carefully designed studies.

# 7 Conclusion

We investigated the research impact of OA articles across the subject disciplines in this exploratory paper and found that it varies from discipline to discipline. OA articles in hard, urban and convergent fields do not seem to have higher research impact as measured by the number of citations than mixed (hard/soft, urban/rural, and convergent/divergent) ones. OA articles in Biology and Economics behaved like hard sciences in terms of research impact. Findings are inconclusive, however. Explanatory studies need to be replicated in order to test the hypothesis that OA articles in hard, urban and convergent fields receive more citations (hence higher research impact) than those in soft, rural, and divergent ones.

## References

[1] MEYER, R.W. (1997). Monopoly power and electronic journals. *Library Quarterly,* 67(4): 325-349.

[2] HOUSE OF COMMONS. (2004). Select Committee on Science & Technology Tenth Report. (Online). Retrieved, 13 April 2007, from http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39902.htm.

[3] PROSSER, D. (2004). The next information revolution-How open access repositories and journals will transform scholarly communications. *LIBER Quarterly*, 14 (1), (Online). Retrieved, 13 April 2007, from http://webdoc.gwdg.de/edoc/aw/liber/lq-1-04/prosser.pdf.

[4] WILLINSKY, J. (2003). Scholarly associations and the economic viability of open access publishing. *Journal of Digital Information*, 4(2), Article No. 177. (Online). Retrieved, 13 April 2007, from http://jodi.ecs.soton.ac.uk/Articles/v04/i02/Willinsky/.

[5] BETHESDA STATEMENT ON OPEN ACCESS PUBLISHING. (2003). (Online). Retrieved, 13 April 2007, from http://www.earlham.edu/~peters/fos/bethesda.htm.

[6] EUROPEAN COMMISSION. (2006). Study on the economic and technical evolution of scientific publication markets in Europe. (Online). Retrieved, 13 April 2007, from http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf.

[7] SUBER, P. (2006). Predictions for 2007. *SPARC Open Access Newsletter*, No. 104. (Online). Retrieved, 13 April 2007, from http://www.earlham.edu/~peters/fos/newsletter/12-02-06.htm.

[8] LAWRENCE, S. (2001) Free online availability substantially increases a paper's impact. *Nature,* 411 (6837): 521. (Online). Retrieved, 13 April 2007, from http://www.copernicus.org/EGU/acp/Nature_ad_1.pdf.

[9] HARNAD, S.; BRODY, T. (2004 June). Comparing the impact of open access (OA) vs. non-OA articles in the same journals. *D-Lib Magazine*, 10(6). (Online). Retrieved, 13 April 2007, from http://www.dlib.org/dlib/june04/harnad/06harnad.html.

[10] HARNAD, S.; BRODY, T.; VALLIERES, F.; CARR, L.; HITCHCOCK, S.; GINGRAS, Y; OPPENHEIM, C.; STAMERJOHANNS, H.; HILF, E. (2004). The access/impact problem and the green and gold roads to open access. *Serials Review*, 30(4): 310-314. (Online). Retrieved, 13 April 2007, from http://dx.doi.org/10.1016/j.serrev.2004.09.013.

[11] ANTELMAN, K. (2004). Do open-access articles have a greater research impact? *College & Research Libraries*, 65(5): 372-382. (Online). Retrieved, 13 April 2007, from http://eprints.rclis.org/archive/00002309/01/do_open_access_CRL.pdf.

[12] SUBER, P. (2004). Promoting open access in the humanities. (Working paper). (Online). Retrieved, 13 April 2007, from http://www.earlham.edu/~peters/writing/apa.htm.

[13] BECHER, T.; TROWLER, P.R. (2001). *Academic tribes and territories: intellectual enquiry and the culture of disciplines.* 2d ed. Buckingham: SRHE and Open University Press.

[14] KLING, R.; McKIM, G. (2000). Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. *Journal of the American Society for Information Science*, *51*(14), 1306-1320.

[15] FRY, J.; TALJA, S. (2004). The cultural shaping of scholarly communication: explaining e-journal use within and across academic fields. In *ASIST 2004: Proceedings of the 67th ASIST Annual Meeting*, Vol. 41, p. 20-30. Medford, NJ: Information Today. (Online) Retrieved, 10 April 2007, from http://people.oii.ox.ac.uk/fry/wp-content/uploads/2006/03/FryTalja_asistfinalsubmission17May.pdf.

[16] ANTELMAN, K. (2006). Self-archiving practice and the influence of publisher policies in the social sciences, *Learned Publishing*, 19, 85-95. (Online). Retrieved, 13 April 2007, from http://eprints.rclis.org/archive/00006023/01/antelman_self-archiving.pdf.

[17] EYSENBACH, G. (2006). Citation advantage of open access articles. *PLoS Biology*, 4(5): e157. DOI: 10.1371/journal.pbio.0040157.

[18] McVEIGH, M.E. (2004). Open access journals in the ISI citation databases: Analysis of impact factors and citation patterns. (Online). Retrieved, 10 April 2007, from http://scientific.thomson.com/media/presentrep/essayspdf/openaccesscitations2.pdf.

[19] MEHO, L.I.; YANG, K. (in press). A new era in citation and bibliometric analyses: Web of Science, Scopus, and Google Scholar. *Journal of the American Society for Information Science and Technology*.

[20] BAUER, K.; BAKKALBASI, N. (2005). An examination of citation counts in a new scholarly communication environment. *D-Lib Magazine*, 11, 9. (Online) Retrieved, 10 April 2007, from http://www.dlib.org//dlib/september05/bauer/09bauer.html.

[21] BAKKALBASI, N.; BAUER, K.; GLOVER, J.; WANG, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, 7. (Online) Retrieved, 10 April 2007, from http://www.bio-diglib.com/content/pdf/1742-5581-3-7.pdf.

[22] JACSO, P. (2005, November 10). As we may search – Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*, 89(9): 1537-1547. (Online). Retrieved, 10 April 2006, from http://www.ias.ac.in/currsci/nov102005/1537.pdf.

[23] WHITLEY, R. (2000). *The intellectual and social organization of the sciences.* New York: Oxford University Press.

[24] KEKÄLE, J. (2002). Conceptions of quality in four disciplines. *Tertiary Education and Management*, 8(1): 65-80.

## Appendix I: Number of Articles and Citations in 46 Open Access Journal Titles

| Journal | Subject | # of times cited | # of articles in the sample | $\overline{X}$ | Indexed in WoS |
|---|---|---|---|---|---|
| Acta Physica Polonica B | physics | 56 | 14 | 4.0 | Yes |
| Brazilian Journal of Physics | physics | 21 | 4 | 5.3 | Yes |
| Entropy: international and interdisciplinary journal of entropy and information studies | physics | 3 | 2 | 1.5 | No |
| New Journal of Physics | physics | 0 | 2 | 0.0 | Yes |
| Pramana: Journal of Physics | physics | 14 | 6 | 2.3 | Yes |
| Turkish Journal of Physics | physics | 1 | 2 | 0.5 | No |
| Balkan Journal of Geometry and Its Applications | mathematics | 0 | 1 | 0.0 | No |
| Bulletin (new series) of the American Mathematical Society | mathematics | 1 | 1 | 1.0 | Yes |
| Electronic Journal of Differential Equations | mathematics | 9 | 8 | 1.1 | No |
| Electronic Journal of Linear Algebra | mathematics | 0 | 1 | 0.0 | Yes |
| Electronic Journal of Qualitative Theory of Differential Equations | mathematics | 1 | 1 | 1.0 | No |
| Electronic Research Announcements of the American Mathematical Society | mathematics | 0 | 1 | 0.0 | Yes |
| Electronic Transactions on Numerical Analysis | mathematics | 6 | 1 | 6.0 | Yes |
| Homology, Homotopy and Applications(HHA) | mathematics | 1 | 1 | 1.0 | Yes |
| Journal of Graph Algorithms and Applications | mathematics | 3 | 1 | 3.0 | No |
| Journal of Inequalities and Applications | mathematics | 1 | 1 | 1.0 | Yes |
| Journal of Integer Sequences | mathematics | 4 | 2 | 2.0 | No |
| Lobachevskii Journal of Mathematics | mathematics | 2 | 2 | 1.0 | No |
| Missouri Journal of Mathematical Sciences | mathematics | 1 | 2 | 0.5 | No |
| The Electronic Journal of Combinatorics | mathematics | 6 | 4 | 1.5 | Yes |
| The New York Journal of Mathematics | mathematics | 1 | 1 | 1.0 | No |
| Theory and Applications of Categories | mathematics | 8 | 2 | 4.0 | No |
| Brazilian Journal of Chemical Engineering | chemical engineering | 5 | 5 | 1.0 | Yes |
| Iranian Polymer Journal | chemical engineering | 4 | 5 | 0.8 | Yes |
| Journal of Chemical Engineering of Japan | chemical engineering | 54 | 20 | 2.7 | Yes |
| Asian Development Review | economics | 14 | 5 | 2.8 | No |
| IMF Staff Papers | economics | 140 | 25 | 5.6 | Yes |
| Electronic Green Journal | environmental sciences | 2 | 3 | 0.7 | No |
| Park Science | environmental sciences | 0 | 5 | 0.0 | No |
| Water SA | environmental sciences | 61 | 22 | 2.8 | Yes |
| Biological Procedures Online | biology | 12 | 2 | 6.0 | Yes |
| Cell Structure and Function | biology | 14 | 2 | 7.0 | Yes |
| Experimental and molecular medicine EMM | biology | 27 | 5 | 5.4 | Yes |
| In Silico Biology | biology | 3 | 1 | 3.0 | No |
| Journal of Biosciences | biology | 21 | 7 | 3.0 | Yes |
| Journal of Radiation Research | biology | 22 | 3 | 7.3 | Yes |
| Molecular Vision | biology | 93 | 10 | 9.3 | Yes |
| Current Research in Social Psychology | psychology | 6 | 5 | 1.2 | No |
| Dynamical Psychology: an international, interdisciplinary journal of complex mental processes | psychology | 0 | 3 | 0.0 | No |
| Journal of Technology in Counseling | psychology | 0 | 2 | 0.0 | No |
| PSYCHE: An Interdisciplinary Journal of Research on Consciousness | psychology | 22 | 20 | 1.1 | No |
| Journal of Criminal Justice and Popular Culture | sociology | 15 | 12 | 1.3 | No |
| IDEA: a Journal of Social Issues | sociology | 0 | 4 | 0.0 | No |
| Journal of Memetics - Evolutionary Models of Information Transmission | sociology | 9 | 14 | 0.6 | No |
| Anthropological Science | anthropology | 58 | 14 | 4.1 | Yes |
| Journal of Physiological Anthropology and Applied Human Science | anthropology | 40 | 16 | 2.5 | No |

# Sharing the Know-how of a Latin American Open Access only e-journal: The Case of the Electronic Journal of Biotechnology

*Graciela Muñoz[1]; Atilio Bustos-González[2]; Alejandra Muñoz-Cornejo[2]*

[1] Instituto de Biología, Facultad de Ciencias Básicas y Matemáticas, Pontificia Universidad Católica de Valparaíso, Av. Brasil 2950, Valparaíso, Chile
e-mail: gmunoz@ucv.cl
[2] Sistema de Biblioteca, Pontificia Universidad Católica de Valparaíso
Av. Brasil 2950, Valparaíso, Chile
e-mail: {abustos; biotec}@ucv.cl

## Abstract

Scientific communication is essential for the advancement of science and in generating benefits for the general society. Also it is fundamental in strengthening the knowledge society with a positive effect on innovation and economic growth. The Open Access journals have demonstrated to be important in providing a reliable and a more accessible mean in communicating science. An example as such, is that they are evaluated by the ISI Thomson Scientific –recognized as an authority for evaluating journals- following the same rigorous selection process as journals in print media. The measurement of the impact factors in the electronic publications demonstrates that these receive a smaller citation level than print journals, ranking in general in the lower half of journals in their subject category. Moreover, the low usage of the electronic media demonstrates a lack of confidence of authors in this new mean of communication. In this sense, editors have to provide answers to some unsolved issues regarding e-publications in order to make these journals more reliable and confident to the scholarly community. The journals edited in Latin America with international visibility represent 0.63% of the total number of publications covered by the ISI Web of Science. In the year 2005, there were 44 Latin American journals covered by the Science Citation Index, of which 89% of these are considered Open Access publications. In that same year, these publications reached an impact factor average of 0.447 in comparison with the impact factor average of 1.588 for all the journals of Open Access at world-wide level. The Electronic Journal of Biotechnology is the only Open Access Latin American journal edited exclusively in the electronic format which is covered by the ISI Science Citation Index. The experience of this journal shows that with commitment to international diversity, quality, academic rigor of the peer review process, transparency, responsibility to scientists, innovation and international cooperation, a high level of visibility and accessibility can be obtained, as demonstrated by an average of more than 83,000 readers during year 2006 and an impact factor of 0.725, which is over the mean value of Latin American journals, offering an unique opportunity to fulfill the ever increasing public demand for science information.

**Keywords:** ISI Web of Science; electronic journal; scientific communication; impact factor; open access

# 1 Introduction

Science and its communication are essential in science advancement and in building the knowledge society with a positive effect on innovation and economic growth. Its dissemination, accessibility and understanding play key roles in its impact on research funding policies and in the benefits for a knowledge-based economy. The success of modern science depends on social acceptance of new scientific results and requires a permanent dialogue with an informed civil society where an open communication system, accessible and visible, is of primary importance for the benefit of society [1].

Although the introduction of the web based technology has raised a continuous debate on science communication, the journal system has remained stable and scientists value journal articles as a recognized mean of communicating original, peer-reviewed and edited information. The key problems regarding the use of this journal system continue to be the high costs of subscriptions, technical barriers, and the specialized language used in the scientific articles which leads to non-equity and discrimination across the international science community [2, 3].

The increasingly pervasive impact of science and technology is reaching every aspect of human welfare and is therefore urgent to make information more accessible and more usable by offering electronic journals a unique opportunity to reach these goals through the instant access to and dissemination of scientific information [4]. An important recent trend has been the development of the Open Access movement, which promotes free online access to full text research articles in every academic field [5, 6].

## 2 Are Scientists and Society Living in the Plato's "Myth of the Cave"?

The changes in information technology and communication that made possible the rise of the knowledge society generated a new type of illiterates. They are citizens and scientists, as authors and publishers, have not sufficiently incorporated the new expertise and possibilities of communication that are based on the use of information technology. This technology not only changed the form of communicating, but also reinvented the strategies used to recover, to analyze, and to diffuse scientific information. In recent years, societal changes have rapidly progressed and caused the previous guidelines of scientific communication to be insufficient for today's society.

An important decision made by the members of the international scientific community was to adopt new guidelines of communication for their results. Also, as evidenced in this article, a large part of the world's researchers continue to resist the new forms of communicating science based in web technology. The consolidation of new communication channels as they are: exclusively electronic journals, the repositories of preprints and postprints, and the institutional repositories. The enhancement of procedures of quality control of science, such as public and open peer review. Furthermore, the opportunities created by multimedia and hypertext that can exist exclusively in an electronic format. New alerting services and new indicators of scientific production, such as the index of Hirsh or the citation tracker, constitute sufficiently forceful changes that have moved for always the guidelines of scientific communication.

All these changes that characterize the knowledge society can lead one to recall Plato's "myth of the cave". It is possible to imagine a parallel link between the scientists who do not incorporate these new guidelines of scientific communication and the inhabitants of the mythical Greek cave. The cave dwellers were convinced that the shadow projected inside the cave was the reality. They were wrong. The reality is that of scientists that live outside of the cave and dominate the useful tools of communication that the world of today offers them. However, the one that remains inside the cave sees a only piece of the present reality and thus runs the serious risk of thinking that what he or she sees is the whole truth. Also, publishers who maintain printed journals and those that impose high subscription costs, have the most part of society living within the cave, thus making difficult the equitable access to high quality scientific data and the possibilities for science to benefit all of society.

## 3 The Traditional System of Academic Journals and the Challenges of the Digital Environment

Key changes must occur in order to make scientific knowledge more accessible, visible, and usable. More editors and publishers should commit themselves to the requirements of the overall society, which claims for innovations that depend in the scientific information.

Moreover, governments should assume a more proactive and strategic role in addressing key international issues regarding the importance of science for society, supporting an efficient communication system. Although the digital era offers a unique opportunity to cope with these goals, the number of Open Access journals indexed in the ISI Web of Science is still low, representing less than 3% of the total number of journals published by this database.

In spite of the well known and unique advantages provided by the electronic journal format in comparison with the print version, such as increased visibility, accessibility to all issues, lower costs of edition, use of hypermedia, the adoption of only electronic journals still poses a challenge to the editor [7, 8].

Some unsolved issues regarding e-publications, for example electronic archiving and uncertainty about future access, generate significant concerns, skepticism, and distrust in the scholarly community. It has taken some time for only e-journals to become integrated into scientific information systems, indexed by major services, appear in library catalogs, and cited by other researchers in main stream journals.

As a result only e-journals covered by the ISI Thomson Scientific database have low impact factors affecting the prestige of these journals. Also it is worth noting that authors tend to stick to traditional formats and do not make use of the advantages offered by electronic media when writing manuscripts for e-journals. The publication of videos, audio, and three-dimensional images between others are all examples of such advantages. In summary, the distrust of the scientific community in this new media directly affects not only the prestige of these journals but also the possibility to have an accessible communication system that satisfies the needs of the scientific community.

## 4 The ISI Thomson Scientific and Impact Factors

The Institute for Scientific Information, ISI, was founded by Eugene Garfield in 1960. Then, in 1992, it was acquired by the Thomson Scientific & Healthcare, thus changing the name to Thomson ISI. It is now a sector of the Thomson Corporation referred to as Thomson Scientific [9].

Recognized by the widespread scientific community as an authority for evaluating journals, it covers the world's leading journals of science and technology. Thomson Scientific, or ISI, offers bibliographic database services, covering thousands of academic journals in all scientific disciplines, social sciences, and arts and humanities that consistently achieve and maintain high quality standards in their editorial processes. The ISI Web of Science includes the Science Citation Index (SCI) with 6,623 journals [10], the Social Sciences Citation Index (SSCI) with 1,962 journals [11], and the Arts and Humanities Citation Index (AHCI) with 1,158 journals [12], all of which are available online through the Web of Science database, a part of the Web of Knowledge database collection.

While the evaluation process is independent of the journal's business model, it depends exclusively on quality standards that are independent of the journal's format, whether it be print or electronic [13].

The ISI Thomson Scientific writes: "E-Journals undergo the same rigorous selection as journals in print media. Publishing Standards, Editorial Content, International Diversity, and Citation Analysis are all considered". This gives clear evidence that both paper and electronic formats are equally reliable and genuinely able to communicate science.

Thomson Scientific also publishes an annual Journal Citation Reports, which lists an impact factor for each of the journals of the SCI and SSCI. This is a quantitative tool, which measures the frequency of citation of an "average article" from a journal in other publications covered by a citation index within a two year period previous to its publication. The impact factor is calculated based on a three-year period, and can be considered to be the average number of times published papers are cited up to two years after publication [14]. For example, the 2007 impact factor for a journal A, which is known in the following year, is calculated as follows:

**X**: 2007 cites in ISI journals to articles published in 2006-2005 by journal A
**Y**: total number of articles published in 2006-2005 by journal A

**Impact Factor 2007 =** $\frac{\mathbf{X}}{\mathbf{Y}}$

Although traditional journals have attained high impact factors, 49.794 being the highest record in 2006, electronic journals in general rank in the lower half of journals in their subject category. Table 1 shows the highest impact factors of e-only journals ranking among the top 12%.

| Journal | Impact factor | Open Acess | ISI subject category | Highest IF of the category | Lowest IF of the category | No. journals of the category |
|---|---|---|---|---|---|---|
| PLos Biology | 14.672 | yes | - Biochemistry & Molecular Biology | 33.456 | 0.097 | 261 |
| PLos Medicine | 8.389 | yes | - Medicine General & Nternal | 44.106 | 0.067 | 105 |
| Genome Biology | 9.712 | no | - Biotechnology & Applied Microbiology | 22.738 | 0.024 | 139 |
| BMC Developmental Biology | 5.41 | yes | - Developmental Biology | 23.69 | 0.66 | 33 |
| BMC Structural Biology | 5.00 | yes | - Biophysics | 16.175 | 0.169 | 65 |

| | | | | | | |
|---|---|---|---|---|---|---|
| BMC Bioinformatics | 4.96 | yes | - Biochemical Research Methods<br>- Biotechnology & Applied Microbiology | 9.876<br>22.738 | 0.404<br>0.024 | 53<br>139 |
| Physiological Genomics | 4.636 | no | - Biochemistry & Molecular Biology<br>- Cell Biology<br>- Physiology | 33.456<br>29.852<br>28.721 | 0.097<br>0.207<br>0.082 | 261<br>153<br>75 |
| BMC Molecular Biology | 4.49 | yes | - Biochemistry & Molecular Biology | 33.456 | 0.097 | 261 |
| BMC Evolutionary Biology | 4.45 | yes | - Evolutionary Biology<br>- Genetics & Heredity | 14.864<br>25.797 | 0.675<br>0.08 | 33<br>124 |
| Pediatrics2 | 4.272 | no | - Pediatrics | 4.272 | 0.208 | 73 |

**Table 1: Ranking of impact factors of the top 12% only e-journals [15, 16]**

The information provided in the table is self explanatory. An effort must be done by editors and publishers of only e-journals in order to make this media more reliable and useful to communicate science and therefore to achieve higher impact factors and to locate journals in the upper half of their subject category.

## 5 The Latin American Context

The journals of Latin America have a low representation in the international databases as in the ISI Web of Science, where 44 journals are covered by the Science Citation Index. These represent a 0.43% of the total of journals on a worldwide basis included in this database. The ranking the impact factors of the Latin American journals is between 0.078 and 3.234, with an average value of 0.442.

It is worth to mention as Table 2 shows, that a high percentage of these publications are Open Access.

| | **Number** | **Percentage %** |
|---|---|---|
| Open Access | 39 | 89 |
| Non Open Access | 5 | 11 |
| Total | 44 | 100 |

**Table 2: Comparison between Open Access and non Open Access Latin American journals covered by the Science Citation Index**

The ISI criterion to identify Open Access journals is that they are available in full text for the data bases DOAJ , J-Stage and SciELO.

The SciELO (Scientific Electronic Library Online) project is an initiative by FASESP (Foundation of Support to the Research in the State of Sao Paulo) and by BIREME (Latin American and Caribbean Centre with Information in Health Sciences) who is headquartered in Brazil. It includes a selected collection of scientific articles in full-text from Latin American scientific publications. Thanks to this project, the selected Latin-American journals are publishing their articles in the electronic format, remaining freely available in the SciELO website and thus acquiring the character of Open Access.

## 6 The Case of the Electronic Journal of Biotechnology

The Electronic Journal of Biotechnology is an Open Access, scientific international peer-reviewed journal which has gained a position in the international scene as the only Latin American journal edited exclusively in the electronic format that belongs to the 1% core of only e-journals covered by the ISI Web of Science. It has an impact factor of 0.725, over the average of the impact factors of journals in Latin America and is positioned number 6 in ranking of impact factors in the 44 Latin American journals covered by the ISI Science Citation Index.

It was created in 1998 by the Pontificia Universidad Católica de Valparaíso, Chile with the declared purpose of servicing the international scientific community to make information more accessible, searchable, relevant, and usable. It supports the principles of equal opportunities and freedom of access to scientific information, making the full contents of all articles permanently accessible and searchable for anyone. Therefore, it satisfies the demands of Open Access initiatives. Also, no charge is required for publication and articles are published under the Creative Commons Public License, where no restrictions apply on subsequent redistribution, allowing unlimited use, distribution, and reproduction in any medium, provided the original work is properly cited. Moreover, the provision of CD-ROMs with the Electronic Journal of Biotechnology website to UNESCO and the subsequent distribution to least developing countries allows for a shortening of the digital divide between countries with and without internet facilities, as the CDs also contains the browser internet explorer.

We have an outstanding international academic editorial board, composed of 72 members from 21 countries with Dr James D. Watson (Nobel Prize Laureate) as the Honorary Member of the board.

The journal covers a broad scope of topics in biotechnology, from molecular biology and the chemistry of biological processes, to policy, educational, and ethical issues related directly to this topic. It publishes review and research original articles, short communications and technical notes after submission to full and strict peer review, engaging a geographically broad group of well-recognized scientists as evaluators. Manuscripts are handled electronically, which drastically reduces the time of publication and accepted articles are published in HTML and PDF formats.

In order to maximize its visibility, the journal is located on two servers, one in the Northern hemisphere (http://www.ejbiotechnology.info) and the other in the Southern hemisphere (http://www.ejbiotechnology.cl) receiving in March 2007 more than 110,000 visits and over 1 million hits per month. Also, the use of CrossRef, a citation-linking network, allows the connection of cited references with full text papers while enhancing visibility and accessibility. The knowledge and skills developed during our 10 years of publication can be summarized in seven commitments:

## 6.1 Commitment to Internationality

The editorial board is international, conformed by 72 members, 34% from North America, 33% from Latin America, 28% from Western Europe, 3% from Near East, 1% from Pacific, and 1% from Asia [17] (see Fig. 1).

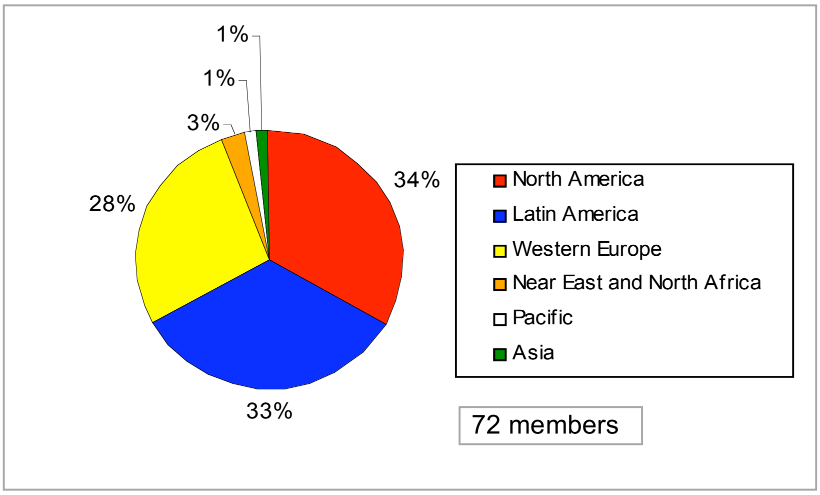


**Figure 1: Editorial board internationality**

Also, the internationality applies to authors (see Fig. 2) and reviewers, as they come from nearly each region in the world.

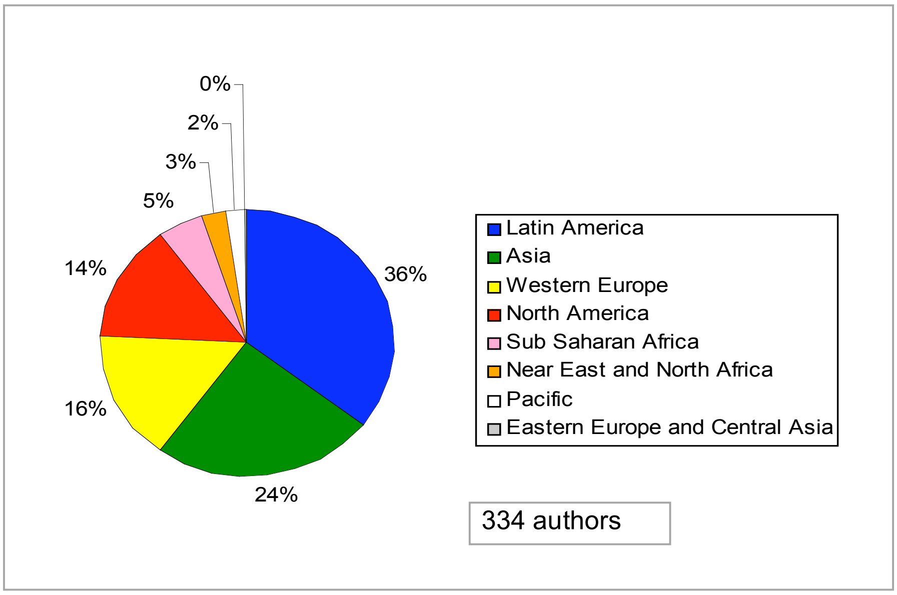


**Figure 2: Corresponding authors internationality**

A statistics software included in the server shows that readers also hail from different regions, with the USA ranking first as the most active country with visitors on both servers. India, UK, Malaysia, Singapore, Canada, Italy, Germany, Australia and Chile follow for the server located in the Northern hemisphere. The activity of the website located in the Southern hemisphere shows that visitors are mainly from Mexico, Chile, Colombia, Spain, Argentina, Brazil, Peru, Germany and India.

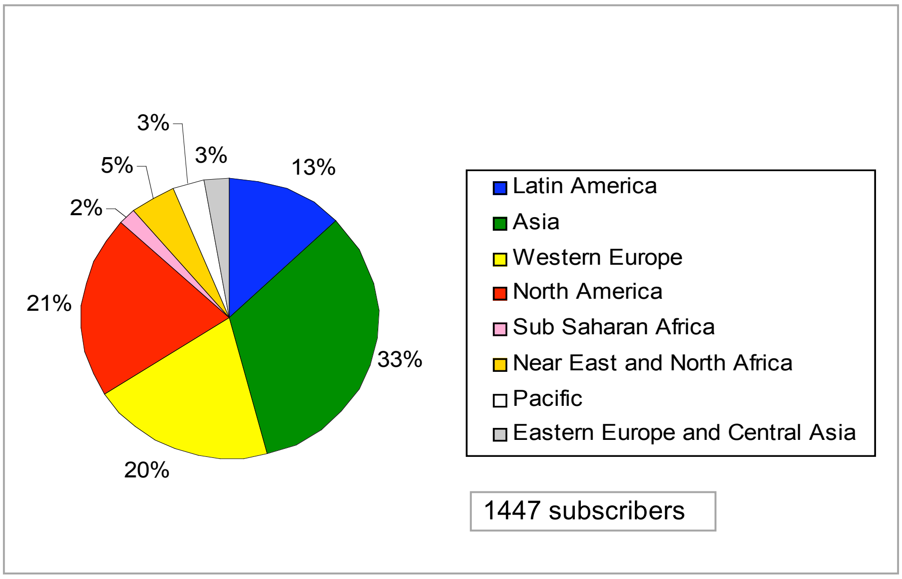


**Figure 3: Subscribers internationality**

Figure 3 shows the international diversity of the subscribers to an email alerting service of the Electronic Journal of Biotechnology.

## 6.2 Commitment to Quality

The journal follows the high standards of scientific publications recommended by the ISI Thomson Scientific. Editorial board members are selected by their publishing records taking into account where their articles have been published and if the manuscripts have been cited. A prominent Honorary Member and a well-recognized editorial board are among the best guarantees for the scientific quality of published articles and are indicative of a reliable media of communication.

The Electronic Journal of Biotechnology follows international editorial conventions, for example, the informative journal title, abstracts, full address information for every author and keywords between others. Also the journal is strictly published according to its stated frequency, 4 times a year, in order to comply with guidelines of publication, which is an important standard criteria for quality.

Complete bibliographic information for all cited references is essential and authors are required that at least 75% of the cited bibliography must be from the last decade while at the same time from ISI indexed journals.

## 6.3 Commitment to Academic Rigor in the Peer Review Process

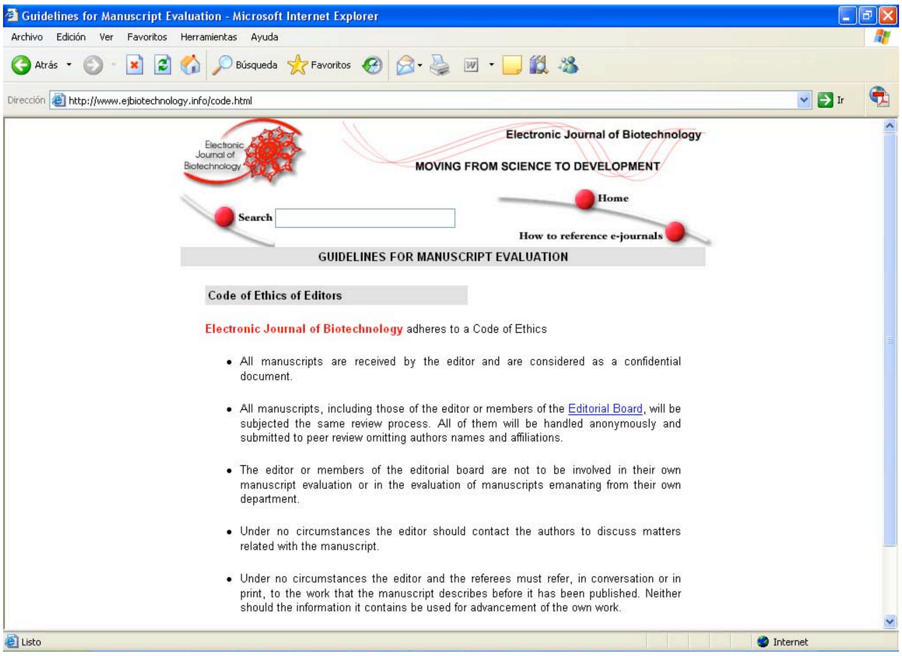
We follow an independent, international and blind peer review process. Evaluators are selected by their expertise from international bibliographical databases and the success of this system is demonstrated not only by the high quality of revision performed on each manuscript, but also because several reviewers have subsequently submitted their manuscripts to the Electronic Journal of Biotechnology in order to be considered for publication. It is worth mentioning, that the refusal of manuscripts has been increasing with time, reaching at present over 70% of rejection (see Table 3).

| **Published articles** | **1998** | **1999** | **2000** | **2001** | **2002** | **2003** | **2004** | **2005** | **2006** | **2007** | **TOTAL** | **Percentage %** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Research | 3 | 7 | 3 | 13 | 18 | 13 | 21 | 26 | 69 | 23 | 196 | 59 |
| Review | 17 | 7 | 11 | 3 | 2 | 5 | 5 | 2 | 3 | 1 | 56 | 17 |
| Short communications | | | 5 | 3 | 5 | 6 | 2 | 4 | 7 | 5 | 37 | 11 |
| Educational Resources | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| Biotechnology issues Developing Countries | | | | 8 | 8 | 5 | 4 | | 3 | 1 | 29 | 9 |
| Issues in Biotechnology Teaching | | | | | 5 | 1 | 2 | 3 | 0 | | 11 | 3 |
| Letter to editor | | | | | 1 | | 2 | | 1 | | 4 | 1 |
| **Total published** | **20** | **14** | **19** | **27** | **39** | **30** | **36** | **35** | **83** | **31** | **334** | **100** |
| | | | | | | | | | | | | |
| **Rejected articles** | | | | | | | | | | | | |
| Research | 4 | 2 | 3 | 9 | 9 | 15 | 53 | 74 | 79 | 80 | 328 | 72 |
| Review | | 1 | 4 | 1 | 1 | 12 | 4 | 7 | 7 | 1 | 38 | 8 |
| Short communications | | | 2 | 1 | 9 | 9 | 3 | 6 | 16 | 33 | 79 | 17 |
| Educational Resources | | | | | | | | | 1 | 1 | 2 | |
| Biotechnology issues Developing Countries | | | | | | 1 | 3 | 3 | 1 | 1 | 9 | 2 |
| Issues in Biotechnology Teaching | | | | | | | | | | 0 | | 0 |
| Minireview | | | | | | | | 1 | 0 | 0 | 1 | 0 |
| **Total rejected** | **4** | **3** | **9** | **11** | **19** | **37** | **63** | **91** | **104** | **116** | **457** | **100** |

**Table 3: Comparison of received, published and rejected articles**

## 6.4 Commitment to Transparency

Instructions to authors, the composition of the editorial board and the statistics of the website are all easily visible and accessible from the homepage of the journal. The items considered in the evaluation process are also transparent to the authors, the originality of the work being of utmost importance. Furthermore, a code of ethics is also visible to the visitors (see Fig. 4).


Guidelines for Manuscript Evaluation - Microsoft Internet Explorer
Archivo
Edición
Ver
Favoritos
Herramientas
Ayuda
Atrás
Búsqueda
Favoritos
Dirección
http://www.ejbiotechnology.info/code.html
Ir
Electronic Journal of Biotechnology
MOVING FROM SCIENCE TO DEVELOPMENT
Home
Search
How to reference e-journals
Listo
Internet

GUIDELINES FOR MANUSCRIPT EVALUATION

Code of Ethics of Editors

Electronic Journal of Biotechnology adheres to a Code of Ethics

- All manuscripts are received by the editor and are considered as a confidential document.
- All manuscripts, including those of the editor or members of the Editorial Board, will be subjected the same review process. All of them will be handled anonymously and submitted to peer review omitting authors names and affiliations.
- The editor or members of the editorial board are not to be involved in their own manuscript evaluation or in the evaluation of manuscripts emanating from their own department.
- Under no circumstances the editor should contact the authors to discuss matters related with the manuscript.
- Under no circumstances the editor and the referees must refer, in conversation or in print, to the work that the manuscript describes before it has been published. Neither should the information it contains be used for advancement of the own work.

**Figure 4: Code of ethics of Electronic Journal of Biotechnology**

## 6.5 Commitment to Scientists

The editor is obliged to scientists, and must acknowledge within three working days the reception of a manuscript. Also, the editor has to respond to the requirements of every author and reader, independent of their academic position and geographic location. The commitment to the scientific community is also demonstrated by the support for an Open Access journal with Open Access licenses that clearly facilitate the retrieval of manuscripts.

## 6.6 Commitment to Innovation

Electronic Journal of Biotechnology provides a good graphical user interface which enhances the usability of the website. This is based on the scientists' requirements of speed and efficiency which are necessary for the identification and retrieval of articles and documents of interest. Also, the use of searchable descriptive metadata greatly increases the accessibility of the journal to search engines. As for example, if the term "journal biotechnology" is searched in Google, one the first documents to be retrieved is the Electronic Journal of Biotechnology.

Also we have adopted the DOI system [18], which provides a persistent and unequivocal identification of each article. It allows the use of CrossRef, a citation linking system that permits a researcher to click on a cited reference and link directly to that reference on the publisher's platform, subject to the publisher's requirements regarding the access to information [19].

## 6.7 Commitment to Cooperation

Electronic Journal of Biotechnology welcomes cooperation with any group interested in communicating scientific results in the area of biotechnology. In this way, we have interacted with UNESCO, Bioline International, REDBIO/FAO Co-operation Network on Plant Biotechnology for Latin America and the Caribbean.

In summary, Open Access electronic journals offers a unique opportunity to fulfil the increasingly public demand for making scientific information more accessible, visible and usable. Scientific knowledge must be made public, as it is a right of education and essential to human development. The problem of the distrust in electronic communication must be overcome by the inclusion of more e-journals in international scientific information systems. The ISI Thomson Scientific database publishing company has ensured that both paper and electronic formats are equally trustworthy and legitimate to communicate science.

## Acknowledgements

The authors wish to acknowledge the contribution of the team of Electronic Journal of Biotechnology. Also the authors are grateful to Mrs. Marcela Aguirre, Departamento de Información, Comisión Nacional de Investigación Científica y Tecnológica (CONICYT) Chile.

## References

[1] Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee on Scientific Information in the Digital Age: Access, Dissemination and Preservation. Brussels, COM(2007) 56 final. {SEC(2007)181}. February 14, 2007. Available from <http://ec.europa.eu/research/science-society/document_library/pdf_06/communication-022007_en.pdf>.[cited March 30 2007].

[2] TENOPIR, C. Lessons for the future of journals. *Nature***,** October 2001, vol. 413, no. 6857, p. 672-674.

[3] TENOPIR, C.; KING, DW. The use and value of scientific journals: past, present and future. *Serials*, 2001, vol. 14, p. 113-120.

[4] ICSU <http://www.icsu.org/index.php> [cited March 30, 2007].

[5] Budapest Open Access Initiative <http://www.soros.org/openaccess/read.shtml> [cited March 30, 2007].

[6] LIESEGANG, TJ.; SCHACHAT, AP.; ALBERT, DM. The Open Access initiative in scientific and biomedical publishing: fourth in the series on editorship. *American Journal of Ophthalmology*, January 2005, vol. 139, no. 1, p. 156-167.

[7] TENOPIR, C.; KING, DW. Reading behaviour and electronic journals. *Learned Publishing*, 2002, vol. 15, p. 259-265.

[8] ROWAN, L. Editorial Electronic paperless scientific communication, are we ready? *Electronic Journal of Biotechnology*, April 2003. [cited March 30 2007]. Available from <http://www.ejbiotechnology.info/content/vol6/issue1/editorial.html>.

[9] Thomson Scientific. <http://scientific.thomson.com/index.html> [cited March 30, 2007].

[10] Science Citation Index Expanded(™) (*Web of Science*) <http://www.thomsonscientific.com/cgi-bin/jrnlst/jloptions.cgi?PC=D> [cited March 30, 2007].

[11] Social Sciences Citation Index® *(Web of Science)* <http://www.thomsonscientific.com/cgi-bin/jrnlst/jloptions.cgi?PC=J> [cited March 30, 2007].

[12] Arts & Humanities Citation Index® (*Web of Science*) <http://www.thomsonscientific.com/cgi-bin/jrnlst/jloptions.cgi?PC=H> [cited March 30, 2007].

[13] The Thomson Scientific Journal Selection Process. <http://scientific.thomson.com/free/essays/selectionofmaterial/journalselection/> [cited March 30, 2007].

[14] The ISI impact factor. <http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor/> [cited March 30, 2007].

[15] Journal Citation Reports 2005 Thomson Scientific.

[16] Directory of Open Access Journals <http://www.doaj.org> [cited March 30, 2007].

[17] Classification of regions according to Science and Engineering indicators 2002. National Science Foundation

[18] The DOI System <http://www.doi.org/> [cited March 30, 2007].

[19] CrossRef <http://www.crossref.org/> [cited March 30, 2007].

# Open Access Journals: A Pathway to Scientific Information in Iran

*Alireza Noruzi*

University of Tehran, Dep. of Library and Information Science, Tehran, Iran
e-mail: nouruzi@gmail.com

## Abstract

This paper reviews the movement of open access (OA) journals in Iran, investigates and compares the influence of Iranian journals in terms of citation ranking, using the Citation Indexes of Thomson-ISI. There has been growth in the number of open access journals in Iran. The advantages of open access for Iranian researchers are: (i) provides access to other research done in their research fields; (ii) speeds up scholarly communication and scientific dialog between researchers; (iii) provides greater visibility and possibly greater impact, although only if open access to the full text is provided. Authors' experiences and motivations have a vital and key role to play in open access. This study indicates that for linguistic reasons, Iranian (Persian-language) journals may not receive and attract the attention that they deserve from the international scientific community. Since there has been little or no discussion in the literature on the impact that the increasing use of OA journals has on scientific production and academic institutions in developing countries, this case study of Iranian experience should be useful for developing countries.

**Keywords**: open access; scientific journals; Persian-language; Iran

## 1 Introduction

The traditional model of scholarly publishing (i.e., publication through peer-reviewed journals) and the new information and communication technologies (i.e., the Internet and the Web) have converged to publish scientific open access (OA) journals, which are freely available to those who want to read, download and print them. Open access has removed many access barriers to the scholarly literature, sharing the knowledge of developed countries with developing countries and vice versa, accelerating research and enriching education. In this new strategy, researchers generally publish the results of their research in scholarly open access journals without payment. Open access can increase the internationality, readership, visibility and Impact Factor (IF) of a journal.

Open access means making the full text of an article available online to all users free of charge, immediately and permanently. It has been defined as "free availability of [scholarly literature] on the public Internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the Internet itself [1]. Stevan Harnad [2] argues that "open access is free, immediate, permanent online access to the full text of research articles for anyone, webwide." However, there remains one constraint linked to copyright: authors control the integrity of their work and must be properly acknowledged and cited. An open access journal is defined here as full text available toll-free on the Web.

The open access movement is a global effort to provide electronic free access to scholarly literature, especially peer-reviewed journals. Open access to the scholarly literature means removing access barriers and limits (e.g., subscription fees, limited access, and so on) to scientific work. There are two parallel "routes" towards open access: open access journals and self-archiving. “Scholarly articles can be made *freely* available to potential readers in one of two main ways – by being published in an open access journal or by being deposited in an electronic repository which is searchable from remote locations without restrictions on access” [3].

OA journals make published articles immediately freely available on their web site, a model mainly funded by charges paid by the author (usually through a research grant). The alternative for a researcher is “self-archiving” (i.e., to publish in a traditional journal, where only subscribers have immediate access, but to make the article available on their personal and/or institutional web sites [including so-called repositories or archives]), which is a practice allowed by many scholarly journals [4].

Open access journals allow, potentially, a greater number of people to access materials compared with subscription-based journals only and in turn helps to solve the research access/impact problem - where restricted access results in the loss of potential research impact [5, 6]. MacCallum and Parthasarathy argue that papers freely available in an open access journal will be more often read and cited than those behind a subscription barrier [7]. A study by Eysenbach [4] reveals that self-archived articles are also cited less often than OA articles from the same journal.

In Iran, open access journals are coming of age, and in a relatively short time have become a mature medium for reporting the results of research. The convenience of open access journals makes them an attractive information resource for Iranian readers and they are increasingly becoming accepted as credible sources of scholarly information.

Iranian open access journals began to proliferate as the Web emerged. For example, 20 journals published in either English or Persian by *Tehran University of Medical Sciences* are open access. Iran is making an electronic version of its English and Persian journals, originally published in print format, freely available to the world. Providing open access to journals is consistent with their strategic goal of enhancing the global visibility of their research output through increasing readership, attracting more contributors, and expanding the influence of Iranian authors in general. The use of open access journals in other fields such as education, science, health, culture, art and development is also maximizing research access through publishing peer-reviewed articles. The Iranian community can now gain access to the results of research by participating in an open access model of research dissemination and individual researchers achieve increased impact typically measured by the number of times a paper is cited and Iranian science gains recognition.

Ensuring that the results of research supported by public funds are made accessible and available for consultation by the research community and others is an integral part of the research process. It involves a partnership between all players involved (universities and other employers of researchers, funders, publishers, libraries, as well as researchers themselves). Ideas and knowledge derived from publicly-funded research are made available and accessible for public use, interrogation, and scrutiny, as widely, rapidly and effectively as practicable [8].

In Iran, open access has already improved the productivity, quality and effectiveness of scientific output, facilitating scholarly communication between Iranian researchers and their foreign colleagues as well as increasing the impact factor (IF) of scientific work based on their citations. Iranian scholarly literature is vital to national productivity and well-being. Publicly-funded research undertaken in Iranian universities and research centers lies at the heart of a productive economy, as well as supporting the social, cultural and physical health of the Iranian nation. Therefore, open access is a valuable vehicle to promote the scientific productivity of Iran. As Brody says: "increased *access* generates the increased *impact*" [9]. The purpose of this paper is to examine the state of open access scholarly journals in Iran and to analyze their visibility through citations to Iranian OA journals in Thomson-ISI Citation Indexes.

## 2 Literature Review

Previous studies demonstrate that open access articles are more immediately recognized and cited than non open access articles, although it depends on the field of science. The effect of OA publishing may be even higher in fields where journals are not widely available on the Web and where articles from the control group remain *toll-access* [4]. Open access increases impact factor [4, 7, 10, 11], that is, authors who make their peer-reviewed articles open access are cited more than those whose full texts are available only on a subscription-basis from the same refereed venue. It is expected that the growth and use of OA will increase as awareness spreads among authors that OA increases visibility, resulting in more citations and therefore leading to greater impact [6].

Eysenbach [4] shows that articles published as an immediate OA article on the journal *(PNAS: Proceedings of the National Academy of Sciences)* site have higher impact than self-archived or otherwise openly accessible OA articles. It is also shown that OA authors are cited more often per paper. He found strong evidence that, even in a journal that is widely available in research libraries, OA articles are more immediately recognized and cited by peers than non-OA articles published in the same journal. He deduces that OA is likely to benefit science by accelerating dissemination and uptake of research findings and suggests that OA journals facilitate knowledge dissemination to a greater degree than self-archiving, presumably because few scientists search on Google for articles if they have encountered an access problem on the journal web site.

Any scientific open access journal's success depends on authors choosing to submit their research to it for publication. Authors publish research in order for the value of their findings to be recognized. The kudos granted by a solid publication record is crucial for a scientific career. If a journal had a reputation for publishing poor science, it would not receive submissions. Thus the system is inherently self-correcting [12]. However, Ghane reports that a large proportion of randomly selected faculty members, as authors, are not familiar with the concept of open access. Thus, the attitudes and experiences of authors, as owners of the copyright of articles, who have published work in open access journals, play an important role in promoting the idea of open access [13].

Antelman studies the impact of freely available articles in different disciplines (philosophy, political science, electrical / electronic engineering and mathematics). The data of the study show a significant difference in the mean citation rates of open access articles and those that are not freely available online in all four disciplines. The relative increase in citations for open access articles ranged from a low of 45 percent in philosophy to 51 percent in electrical and electronic engineering, 86 percent in political science, and 91 percent in mathematics [14].

Thomson-ISI recently conducted a study of the overall performance of OA journals, using a selection of OA journals in the field of natural sciences and focusing on determining whether OA journals perform differently from other journals in their respective fields. The study's initial findings indicate that there was no discernible difference in terms of citation impact or frequency with which the (open access) journal is cited [15]. On the other hand, Lawrence, investigating the impact of free online articles citation rates in the field of computer science, reported that there is a clear correlation between the number of times an article is cited and the probability that the article is online. More highly cited articles, and more recent articles, are significantly more likely to be open access [11].

The impact factor of journals continues to attract a lot of attention, especially from journal editors, publishers, authors and librarians. Librarians may use the ISI impact factor as one element in selection and de-selection procedures; scientists may be interested in journals with high impact factors in order to reach the highest possible visibility for their published results; funding agencies may consider the impact factors of the journals in which researchers given a grant publish funded research; and university research councils may use journal impact factors as indices in local evaluation studies [16].

## 3 Research Questions

This research seeks to answer the following questions:

- What constitutes a successful open access journal and how can we ascertain and measure such success?
- What is the role of the authors?
- How is certification of an open access journal related to success?
- What incentives and assistance are needed?

## 4 Materials and Methods

The approach used in this study includes the following steps:

- First, we conducted a search on Google and Iranian directories of scholarly journals to find open access journals, see *Iranian Directory of Open Access Journals* [17];
- Second, to determine citation rates and Citation Impact[1][18], *Web of Science* (Thomson-ISI citation index) was searched on April 10, 2007, for all Iranian open access journals.

## 5 Results

It is noteworthy that there are 960 Iranian (either Persian or English language) print-based journals and magazines out of which 247 journals (i.e., 28 English and 175 Persian) are accredited by the Iranian *Ministry of Science, Research and Technology* (MSRT, [19]), and 113 journals (i.e., 23 English and 90 Persian) in the fields of medicine, health, nursing, dentistry, pharmacy, podiatry, and biomedicine are accredited by the *Ministry of Health and Medical Education* (MOHME, [20]). Almost all of the English-language journals accredited by *MSRT* and *MOHME* are now open access or *back access* (back-issue or back-volume open access) (see *Iranian*

[1] The Citation Impact is the ratio of the total number of citations received to the total of citable items published in a journal. Citation Impact can be used as a measure of the *impact* an article has had within its particular field [18].

*Directory of Open Access Journals*). It should be noted that *Thompson-ISI* citation indexes index only 15 English-language journals from Iran. The current study includes only OA journals published in English.

Table 1 shows the total number of citations to Iranian English-language OA journals, either the ministries accredited or not. The total number of citations (with or without self-citations) is a reliable indicator of scholarly impact and influence [21].

| Journal title | Total No. of Citations in WoS | No. of Citations since OA began |
|---|---|---|
| Iranian Polymer Journal | 304 | 304 |
| Iranian Journal of Chemistry & Chemical Engineering | 183 | 29 |
| Iranian Journal of Public Health | 152 | 24 |
| Journal of Sciences (Islamic Republic of Iran) | 150 | 31 |
| Iranian Journal of Medical Sciences | 119 | 77 |
| Archives of Iranian Medicine | 80 | 80 |
| Acta Medica Iranica | 76 | 10 |
| Iranian Journal of Pharmaceutical Research | 47 | 42 |
| DARU | 40 | 32 |
| Journal of the Earth and Space Physics | 38 | 0 |
| Journal of the Iranian Chemical Society | 35 | 35 |
| Iranian Biomedical Journal | 24 | 21 |
| Journal of Agricultural Science and Technology | 21 | 21 |
| Iranian International Journal of Science | 14 | 14 |
| Journal of the Iranian Statistical Society | 13 | 0 |
| Iranian Journal of Biotechnology | 11 | 11 |
| International Journal of Endocrinology and Metabolism | 10 | 10 |
| Iranian Heart Journal | 10 | 2 |
| Iranian Journal of Pharmacology and Therapeutics | 9 | 9 |
| Webology | 8 | 8 |
| International Journal of Environment Science and Technology | 7 | 7 |
| Shiraz E-Medical Journal | 6 | 6 |
| Journal of Research in Medical Sciences | 6 | 4 |
| Iranian Journal of Allergy, Asthma and Immunology | 6 | 3 |
| Iranian Journal of Radiation Research | 5 | 5 |
| Iranian Journal of Immunology | 4 | 4 |
| Iranian Journal of Pharmaceutical Sciences | 4 | 4 |
| Iranian Journal of Reproductive Medicine | 4 | 4 |
| Iranian Journal of Veterinary Research | 4 | 0 |
| Journal of Medical Education | 4 | 4 |
| Iranian Journal of Pediatrics | 3 | 1 |
| Iranian Journal of Clinical Infectious Diseases | 2 | 2 |
| Iranian Journal of Radiology | 2 | 2 |
| Iranian Journal of Mathematical Sciences and Informatics | 1 | 1 |
| Journal of Dentistry of Tehran University of Medical Sciences | 1 | 1 |
| Advanced Research Yields across Atherosclerosis | 0 | 0 |
| Caspian Journal of Environmental Sciences | 0 | 0 |
| Hepatitis Monthly | 0 | 0 |
| International Journal of Hematology- Oncology and Bone Marrow Transplantation | 0 | 0 |
| Iranian Journal of Environmental Health Science & Engineering | 0 | 0 |
| Iranian Journal of Parasitology | 0 | 0 |
| Iranian Journal of Pathology | 0 | 0 |
| Iranian Rehabilitation Journal | 0 | 0 |
| Journal of Tehran Heart Center | 0 | 0 |
| Journal of Respiratory Disease, Thoracic Surgery, Intensive Care and Tuberculosis | 0 | 0 |
| Urology Journal | 0 | 0 |

**Table 1: Total Number of Citations to Iranian English-Language OA Journals**

Table 1 is a ranked list of the English-language OA journals included in the study, although ranking by total citations obviously favors older and more famous journals. The last column shows the number of citations since OA began. It should be noted that Iranian English-language journals, published by well-known universities, are still in their infancy and need more time to be recognized by their peers and the international scientific community. It seems that one of the main reasons why Iranian journals are not widely cited is that they are not indexed and circulated by foreign databases, especially American and British databases (e.g., Medline, CAB, EBSCO, Proquest, ERIC, Web of Science, WorldCat, LISA, INSPEC, Agris, COMPENDEX, etc.). Therefore, not only open access but also wide circulation is important for a journal's acceptance and reputation.

Table 2 comprises a sample of Persian-language OA journals (including English-language abstracts), nationally well-known, for comparison with the English-language journals.

| **Journal title** | **Total No. of Citations in WoS** | **No. of Citations since OA began** |
|---|---|---|
| Iranian Journal of Diabetes & Lipid Disorders | 6 | 6 |
| Iranian Journal of Nuclear Medicine | 2 | 1 |
| Audiology | 0 | 0 |
| HAYAT | 0 | 0 |
| Journal of Dental Medicine | 0 | 0 |
| Scientific Journal of School of Public Health and Institute of Public Health Research | 0 | 0 |
| Tehran University Medical Journal | 0 | 0 |

**Table 2: Total Number of Citations to Persian-Language OA Journals**

The comparison between non-English-language and English-language open access journals from Iran shows that English-language journals are more cited. Examination of citations to Persian-language OA journals from English-language journals shows that they are infrequent and only cited by Persian-speaking authors. Therefore, it can be concluded that English-speaking authors do not cite Persian-language journals. It should also be noted that Thomson-ISI citation indexes have a bias towards the English-language, indexing few non-English-language journals.

## 6 Suggestions for Improvement

Iranian institutions should implement a policy:

- to encourage their researchers to publish their research papers in open access journals where a suitable journal exists (and provide the support to enable that to happen);
- to launch new open access journals, where necessary, to serve individual communities, and should support existing journals who want to make the transition to open access;
- to establish open access repositories for English and Persian papers written by Iranians;
- to require Iranian researchers to deposit a copy of all their published papers in a national open access repository; and
- to submit OA journals for inclusion in a large number of indexing and abstracting databases to be widely circulated and widely read.

Creating and developing digital repositories that will assist Iranian academic organizations in the ongoing process of curating –identifying, selecting, acquiring, managing, describing, and providing access to– their scientific collections is vital if the community is to successfully ensure the preservation and continuing access of electronic resources. It is recommended that the Iranian government and authors consider the following suggestions:

- The Iranian government should provide funds for all universities to launch open access institutional repositories, appointing a central body to launch a national repository for preservation and to coordinate the implementation of a network of institutional repositories;
- Iranian universities should launch and support open access journals, and encourage faculty members to take action in support of open access;
- All Iranian universities should establish institutional repositories as an important first step toward more radical change;

- Authors of articles based on government funded research should deposit articles in their institutional repositories, after publication;
- Authors should self-archive (deposit electronic articles into electronic archives);
- Iranian universities should call on all university faculties to self-archive a digital copy of every article accepted by a peer-reviewed journal into the institutional repository.

## 7 Discussion and Conclusion

To sum up, 'open access' to Iranian scholarly literature is the key element for Iran, improving and accelerating the scientific activities. The Internet makes it possible for Iranian research papers to be read more easily and therefore probably get cited more, because of free, unrestricted access to open access journals. Research institutions that support open access will benefit greatly in terms of impact and influence, due to the greater accessibility and visibility of their research. Iranian researchers should absolutely have the right to see the results of the research that their taxes have paid for.

Some Iranian journals (English or Persian language) currently offer delayed free access, or *back access*, making issues of journals free six months or a year after journal publication. It is worth noting that in fast-moving topics, such information may be out of date when the readers gain access, thus providing *back access* rather than open access. The overall costs of providing open access to scholarly journals are far lower than the costs of traditional print journals, therefore we suggest that Iranian journals, especially international English-language journals become OA, because it is not possible for a print journal to be circulated throughout the world.

Ensuring that the main outputs of research –knowledge and ideas- are disseminated widely is vitally important. Iranian universities should support moves by the research community and scholarly journal publishers to develop new publishing models that are based on the principle that research outcomes should be freely accessed and disseminated as widely as possible via the Internet. It should be noted that OA by itself does not guarantee greater impact and influence for an OA journal, except if the journal publicizes and circulates its contents as widely as possible via international discussion groups, listservs and databases.

Briefly, the advantages of open access for Iranian researchers are: (i) provides access to other research done in their research fields; (ii) speeds up scholarly communication and scientific dialog between researchers; (iii) provides greater visibility and possibly greater impact, although only if open access to the full text is provided.

## Acknowledgments

The author wishes to thank Mrs. Marjorie Sweetko for her helpful comments in producing the paper.

## References

[1] CHAN, L., et al. Budapest Open Access Initiative, (2002, February 14). Available at: http://www.soros.org/openaccess/read.shtml

[2] HARNAD, S. *American-Scientist-Open-Access-Forum*. Re Proposed update of BOAI definition of OA Immediate and Permanent, 2005. Available at: http://www.ecs.soton.ac.uk/~harnad/Hypermail/Amsci/4420.html

[3] SWAN, A.; BROWN, S. *JISC/Open Society Institute Journal Authors Survey*, 2004. Available at: http://www.jisc.ac.uk/uploaded_documents/JISCOAreport1.pdf

[4] EYSENBACH, G. Citation Advantage of Open Access Articles. *PLoS Biology*, 4(5) (2006), e157. Available at: http://biology.plosjournals.org

[5] HARNAD, S. *et al.* The green and gold roads to open access. *Nature*, 2004. Available at: http://www.nature.com/nature/focus/accessdebate/21.html

[6] HARDY, R.; OPPENHEIM, C.; BRODY, T.; HITCHCOCK, S. Open Access Citation Information. Final Report – Extended Version, JISC Scholarly Communications Group, September 2005. Available at: http://eprints.ecs.soton.ac.uk/11536/

[7] MACCALLUM, C.J., & Parthasarathy, H. Open Access Increases Citation Rate. *PLoS Biology*, 4(5) (2006), e176. Available at: http://biology.plosjournals.org

[8] Research Councils UK. Rcuk Position Statement on Access to Research Outputs, 2005, June. Available at: http://www.rcuk.ac.uk/access/statement.pdf

[9] BRODY, T. Citation Analysis in the Open Access World. Author eprint, 2004. Available at: http://eprints.ecs.soton.ac.uk/10000/01/tim_oa.pdf

[10] HITCHCOCK, S. The effect of open access and downloads ('hits') on citation impact: A bibliography of studies, 2005. Available at: http://opcit.eprints.org/oacitation-biblio.html

[11] Lawrence, S. Online or Invisible? *Nature*, 411(6837) (2001), p. 521.

[12] WEITZMAN, J.B. (Mis)Leading Open Access Myths, 2006. Available at: http://www.biomedcentral.com/openaccess/inquiry/myths/

[13] GHANE, M. A Survey of Open Access Barriers to Scientific Information: Providing an Appropriate Pattern for Scientific Communication in Iran. *The Grey Journal: An International Journal on Grey Literature*, 2(1), 2006.

[14] ANTELMAN, K. Do open-access articles have a greater research impact? *College & Research Libraries*, 65(5) (2004), 372-382. Available at: http://eprints.rclis.org/archive/00002309/

[15] Thomson-ISI. *The Impact of Open Access Journals: A Citation Study from Thomson ISI*, 2004.Available at: http://www.isinet.com/media/presentrep/acropdf/impact-oa-journals.pdf

[16] ROUSSEAU, R. Impact of African Journals in ISI Databases. *LIBRES: Library and Information Science Research Electronic Journal*, 15(2), (2002). Available at: http://libres.curtin.edu.au/libres15n2/index.htm

[17] Iranian Directory of Open Access Journals, 2007. Available at: http://nouruzi.googlepages.com/IDOAJ.doc

[18] BRODY, T. *et al.* The effect of open access on citation impact. *National Policies on Open Access (OA) Provision for University Research Output: An International meeting*. Southampton University, Southampton, UK, 19 February 2004. Available at: http://opcit.eprints.org/feb19oa/brody-impact.pdf

[19] MSRT. Ministry of Science, Research and Technology, 2006. Available at: http://www.msrt.gov.ir/

[20] MOHME. Ministry of Health and Medical Education, 2006. Available at: http://www.net.hbi.ir/new/dynamic/journals/journal-index.php

[21] CRONIN, B.; MEHO, L. Using the *h*-index to rank influential information scientists. *Journal of the American Society for Information Science and Technology*, 57(9) (2006), 1275–1278.

# Automatic Sentiment Analysis in On-line Text

*Erik Boiy; Pieter Hens; Koen Deschacht; Marie-Francine Moens*

Katholieke Universiteit Leuven, Tiensestraat 41 B-3000 Leuven, Belgium
e-mail: erik.boiy@law.kuleuven.be; pieter.hens@econ.kuleuven.be
koen.deschacht@law.kuleuven.be; marie-france.moens@law.kuleuven.be

## Abstract

The growing stream of content placed on the Web provides a huge collection of textual resources. People share their experiences on-line, ventilate their opinions (and frustrations), or simply talk just about anything. The large amount of available data creates opportunities for automatic mining and analysis. The information we are interested in this paper, is how people feel about certain topics. We consider it as a classification task: their feelings can be positive, negative or neutral. A sentiment isn't always stated in a clear way in the text; it is often represented in subtle, complex ways. Besides direct expression of the user's feelings towards a certain topic, he or she can use a diverse range of other techniques to express his or her emotions. On top of that, authors may mix objective and subjective information about a topic, or write down thoughts about other topics than the one we are investigating. Lastly, the data gathered from the World Wide Web often contains a lot of noise. All of this makes the task of automatic recognition of the sentiment in on-line text more difficult. We will give an overview of various techniques used to tackle the problems in the domain of sentiment analysis, and add some of our own results.

**Keywords:** sentiment analysis; document classification; artificial intelligence

## 1 Introduction

Automatic sentiment analysis is a topic within information extraction that only recently received interest from the academic community. In the previous decade, a handful of articles have been published on this subject. It's only in the last five years that we've seen a small explosion of publications. The idea of automatic sentiment analysis is important for marketing research, where companies wish to find out what the world thinks of their product; for monitoring newsgroups and forums, where fast and automatic detection of flaming is necessary; for analysis of customer feedback; or as informative augmentation for search engines.

The automatic analysis of sentiments on data found on the Web is useful for any company or institution caring about quality control. For the moment, getting user feedback means bothering him or her with surveys on every aspect the company is interested in. The problems with this approach are making a survey for each product or feature; the format, distribution and timing of the survey (asking to send a form right after purchase might not be very informative); and the reliance on the goodwill of people to take the survey. This method can be made obsolete by gathering such information automatically from the World Wide Web, where the large amount of available data creates the opportunity to do so. One of the sources are blogs (short for "web logs"), a medium through which the blog owner makes commentaries about a certain subject or talks about his or her personal experiences, inviting readers to provide their own comments. Another source are the electronic discussion boards, where people can discuss all kinds of topics, or ask for other people's opinions. We define a topic as the subject matter of a conversation or discussion, e.g. an event in the media or a new model of car, towards which the writer can express his or her views.

There are several additional advantages to this approach. First, the people who share their views usually have more pronounced opinions than average, which are additionally influencing others reading them, leading to so-called word-of-mouth marketing. Extracting these opinions is thus extra valuable. Second, opinions are extracted in real-time, allowing for quicker response times to market changes and for detailed time-based statistics that make it possible to plot trends over time.

This paper is organized as follows: In section 2 we will go over the concepts of emotions in written text. Section 3 gives an overview of various methods that can be used to analyse the sentiment of a text, making a distinction between symbolic techniques and machine learning approaches. In section 4 we describe some challenges in the field that need to be overcome. Section 5 provides a comparison of results from the literature using the aforementioned techniques, to which we add some of our own results1. In section 6 we shortly discuss those results, before coming to conclusions in section 7.

# 2 Concepts of Emotions in Written Text

## 2.1 Concept of Emotions

Before attempting to classify sentiments, we must ask the question what sentiments are. In general we can state that sentiments are either emotions, or they are judgements or ideas prompted or coloured by emotions[2]. An emotion consists of a set of stages, namely: appraisal, neural and chemical changes and action readiness. We will give a quick overview of each of these states.

An emotion is usually caused by a person consciously or unconsciously evaluating an event, which is denoted *appraisal* in psychology. Appraisal does not only denote the evaluation whether something is positive or negative, but it also denotes other measurements such as the significance of an event, the personal control or the involvement of the own ego. In general, the same appraisal gives rise to the same emotion. Appraisal causes *mental and bodily changes*, that make up the actual experience of an emotion. Emotions urge for actions and prompt for plans: an emotion gives priority for one or a few kind of *actions* to which it gives a sense of urgency. We use the term "action" to denote all mental or physical actions (that are the result of an emotion). This includes actions such as moving away from a negative event, mental processes, such as worrying about the event, and other effects that are direct result of the emotion, such as crying or going pale.

## 2.2 Emotions in Written Text

The study of emotions in text can be conducted from two points of view. Firstly, one can investigate how emotions influence a writer of a text in choosing certain words and/or other linguistic elements. Secondly, one can investigate how a reader interprets the emotion in a text, and what linguistic clues are used to infer the emotion of the writer. In this text, we'll take the second point of view. We are interested in the way people infer emotions, so we can mimic this process in a computer program. In the remainder of this section we will investigate how linguistic elements describing appraisal and action-readiness are used in texts to convey the emotion of the author, as they comprise the majority of clues to infer emotion from text.

### Appraisal

A lot of linguistic scholars agree on the three dimensions of Osgood and al. [1], who investigated how the meaning of words can be mapped in a semantic space. Factor analysis extracted 3 major dimensions: (1) positive or negative evaluation (2) a power, control or potency dimension and (3) an activity, arousal or intensity dimension. Although these dimensions are originally proposed as the dimensions of a semantic space, they can also be used to organize linguistic categories of emotion or for the automatic detection of emotions. Most research is devoted towards the appraisal component of emotions, and we will look into it a bit deeper by briefly going over Osgood's dimensions, giving some examples along the way.

(1) Evaluation (positive/negative)
The evaluation dimension is fairly straightforward; it contains all choices of words, parts of speech, word organization patterns, conversational techniques, and discourse strategies that express the orientation of the writer to the current topic. Evaluation is often expressed by using adjectives.
e.g. “It was an *amazing* show.”

(2) Potency (powerful/unpowerful)
This dimension contains all elements that general express whether the writer identifies and commits himself towards the meaning of the sentence or whether he dissociates himself. From a psychological standpoint these phenomena are related to approach and avoidance behaviour. This dimension consists of 3 sub-dimensions: proximity, specificity and certainty.

(2.1) Proximity (near/far)

[1] This research is sponsored by the company Attentio (Brussels, Belgium): attentiocontact@attentio.com. We are very grateful to Per Siljubergsåsen and Simon Mc Dermott for the funding and the relevant remarks. We thank Kalina Lipinska for the annotation of one of the corpora we use in our experiments. We thank Roxana Angheluta for her valuable help when integrating our work in Attentio’s system architecture.

[2] Adapted from the Merriam-Webster On-line Search dictionary.

This category contains all linguistic elements that indicate the 'distance' between the writer and the topic. The proximity from the writer to the current topic expresses whether the writer identifies himself with the topic or distance himself from it.
e.g. “I'd like you to meet John.” versus “I'd like you to meet Mr. Adams.” (social proximity)

(2.2) Specificity (clear/vague)
Specificity is the extent to which a conceptualized object is referred to by name in a direct, clear way; or is only implied, suggested, alluded to, generalized, or otherwise hinted at.
e.g. “I left *my* / *a* book in your office.” (particular vs general reference)

(2.3) Certainty (confident/doubtful)
This dimension expresses the certainty of the writer towards the expressed content. A stronger certainty indicates that the writer is entirely convinced about the truth of his writings and possibly indicates a stronger emotion.
e.g. “It *supposedly* is a great movie.” versus “It *definitely* is a great movie.”

(3) Intensifiers (more/less)
When expressing emotions, a lot of the emotional words used do not express an emotion, but modify the strength of the expressed emotion. These words, the intensifiers, can be used to strengthen or weaken both positive and negative emotions.
e.g. “This is *simply* the best movie.” (adverb)
“He had cuts *all* over.” (quantifier)
“Where *the hell* have you been?” (swearing)

### Direct Expressions

The most direct way to express an emotion is of course to express it directly, without making a detour by using appraisal or action readiness. This can be done among others by using verbs and adjectives [2, 3]. A typical way to express an emotion directly seems to be a pattern similar to "I am/feel/seem [adjective describing emotion]"
e.g. I *ache for* a cigarette.
I *am delighted* of the final results.

### Elements of Action

Excellent examples of actions indicating emotions are of course crying and laughing, but more subtle signs that denote emotion in certain circumstances can be considered as well. An example is looking at your watch when watching a movie, which is most probably a result of boredom and a lack of interest.
e.g. I was *grinning* the whole way through it and *laughing out loud* more than once.

### Remarks

There are additional ways of expressing emotions that don't strictly fall into above categories, like the use of figurative language and irony. It must also be noted that most techniques in sentiment classification focus on terms that do actually not really denote emotions, but denote evaluation, appreciation or judgement. Of course this is not surprising, because most techniques focus on reviews of movies, products, cars, etc., and basically in a review the reviewer evaluates the object under discussion. The sentiment of the reviewer is often not discussed, although of course, it is often easy to infer his emotions. Recognizing the fact that classifying a review is in essence classifying it according to appraisal, doesn't only improve understanding but can also lead to the discovery of new techniques.

## 3 Methodology

In the previous section we discussed the indicators of sentiment in text. In this section we will see methods of identifying this information in a written text. There are two main techniques for sentiment classification: symbolic techniques and machine learning techniques. The symbolic approach uses manually crafted rules and lexicons, where the machine learning approach uses unsupervised, weakly supervised or fully supervised learning to construct a model from a large training corpus.

## 3.1 Symbolic Techniques

### 3.1.1 Lexicon Based Techniques

The simplest representation of a text is the bag-of-words approach. Hereby, we simply consider the document as a collection of words without considering any of the relations between the individual words. Next, we determine the sentiment of every word and combine these values with some aggregation function (such as average or sum). We will discuss a selection of methods to determine the sentiment of a single word.

#### 3.1.1.1 Using a Web Search

It was already indicated by Hatzivassiloglou and Wiebe [4] that adjectives are good indicators of subjective, evaluative sentences. Turney [5] recognizes that, although an isolated adjective may indicate subjectivity, there may be insufficient context to determine semantic orientation. For example, the adjective "unpredictable" may have a negative orientation in an automotive review, in a phrase such as "unpredictable steering", but it could have a positive orientation in a movie review, in a phrase such as “unpredictable plot". Therefore he used tuples consisting of adjectives combined with nouns and of adverbs combined with verbs.

The tuples are extracted from the reviews and the semantic orientation of a review is calculated as the average semantic orientation of the tuples taken from that review. To calculate the semantic orientation for a tuple (such as "unpredictable steering"), Turney uses the search engine Altavista. For every combination, he issues two queries: one query that returns the number of documents that contain the tuple close (defined as “within 10 words distance”) to the word "excellent" and one query that returns the number of documents that contain the tuple close to the word "poor". If the combination is found more often in the same context as "excellent" than in the same context as "poor", the combination is considered to indicate a positive orientation, and otherwise to indicate a negative orientation.

#### 3.1.1.2 Using WordNet

Kamps and Marx use WordNet [6] to determine the orientation of a word. In fact, they go beyond the simple positive/negative orientation, and use the dimension of appraisal that gives a more fine-grained description of the emotional content of a word. Kamps and Marx developed an automatic method [7] using the lexical database WordNet to determine the emotional content of a word along Osgood et al.'s dimensions. In essence, the WordNet database consists of nodes (the words) connected by edges (synonym relations). Kamps and Marx define a distance metric between the words in WordNet, called minimum path-length (MPL). This distance metric counts the number of edges of the shortest path between the two nodes that represent the words. For example, the words "good" and "big" have a MPL of 3. The shortest path from the word "good" to the word "big" is the sequence <good, sound, heavy, big>.

To estimate the magnitude of a dimension of appraisal for a particular word, they compare the MPL of that word towards the positive and towards the negative end of that dimension. Both ends of a dimension are represented by prototype-words. The positive end of the evaluative dimension is represented by the word "good" and the negative end is represented by the word "bad". The prototypes for the potency dimension are respectively "strong" and "weak" and for the activity dimension "active" and "passive".

Only a subset of the words in WordNet can be evaluated using this techniques, because not all words are connected to one of the prototype words. After examination, it showed that the subset of words connected to either "good" or "bad" was composed of 5410 words. Interestingly, the subset of words connected to either "strong" or "weak" consisted of exactly the same 5410 words, and so did the subset connected to "active" or "passive". It seems that all important words expressing emotive or affective meaning are included in this one set.

### 3.1.2 Sentiment of Sentences

So far, we've seen different methods that determine the sentiment of a single word and assumed a simple approach to combine the sentiments of words within a single sentence. The bag-of-words approach has some important drawbacks. As already briefly indicated in section 3.1.1.1, it can often be advantageous to consider some relations between the words in a sentence. There are several approaches in this field; we mention here briefly Mulder and al.'s article [8], which discusses the successful use of an affective grammar. They note that simply detecting emotion words can tell whether a sentence is positive or negative oriented, but does not explain towards what topic this sentiment is directed. In other words, what is lacking in the research towards affect is the relation between attitude and object. Mulder and al. have studied how this relation between attitude and object can be formalized. They combined a lexical and grammatical approach: (1) lexical, because they believe that affect is primarily expressed through affect words, and (2) grammatical, because affective meaning is intensified and propagated towards a target through function-words and grammatical constructs.

## 3.2 Machine Learning Techniques

In this section a description and comparison of state-of-the-art machine learning techniques used for sentiment classification are discussed. First a description is given of a selection of different features that are commonly used to represent a document for the classification task, followed by an overview of machine learning algorithms.

### 3.2.1 Feature Selection

The most important decision to make when classifying documents, is the choice of the feature set. Several features are commonly used, like unigrams or part-of-speech (the linguistic category of a word, further shortened to "POS") data. Features and their values are commonly stored in a feature vector.

**Unigrams**
This is the classic approach to feature selection, in which each document is represented as a feature vector, where the elements indicate the presence (or frequency) of a word in the document. In other words, the document is represented by its keywords.

**N-grams**
A word N-gram is a subsequence of N words from a given sequence (e.g. a sentence). This means that the features in the document representation are not single words, but pairs (bigrams), triples (trigrams) or even bigger tuples of words. For example, "easy" followed by "to" becomes "easy to" in a bigram. Other examples of positive oriented bigrams are: "the best", "I love", "the great", ... and negative oriented: "not worth", "back to", "returned it", ... [9]. With the use of N-grams it is possible to capture more context. N-grams are for example effective features for word sense disambiguation [10]. When using N-grams, the feature vector could take on enormous proportions (in turn increasing sparsity the of the feature vectors). Limiting the feature vector size can be done by setting a threshold for the frequency of the N-grams, or by defining rule sets (e.g. only incorporate N-grams that satisfy a certain pattern like *Adjective Noun* or *Adverb Verb*).

**Lemmas**
Instead of using the words as they literally occur in the text, the lemmas of these words can be used as features for the document. This means that for each word its lemma, being its basic dictionary form, is identified. Examples are:

| | | |
|---|---|---|
| *writes -> write* | *was -> be* | *better -> good* |
| *written -> write* | *cars -> car* | *best -> good* |

The advantage with lemmatisation is that the features are generalized and it will be easier to classify new documents, but this is not always true: you still have to look out for overgeneralization. Dave et al. [9] report a decrease in accuracy of sentiment classification when the words in the documents are conflated to their dictionary form. Lemmatisation comes with loss of detail in the language. For example, Dave notes that negative reviews tend to occur more in the past tense, which cannot be detected after lemmatisation.

**Negation**
Another extension of the unigram approach is the use of negation. When you only consider the words in a sentence and someone writes "*I don't like this movie",* a program can think that this person loved the movie, when it looks at the word "like". A solution for this is to tag each word after the negation until the first punctuation (with for example NOT_). The previous sentence will then become: "*I don't NOT_like NOT_this NOT_movie"*. This was done by [11]. In this experiment, the negation tagging gives a slight decrease in performance. Dave et al. [9] note that simple substrings (N-grams) work better at capturing negation phrases.

**Opinion Words**
Opinion words are words that people use to express a positive or negative opinion [12]. Opinion words are obtained from several POS classes: adjectives, adverbs, verbs and nouns [13, 14]. These opinion words can be

incorporated into the feature vector, where they represent the presence or absence of such a word. Two techniques can be used to define opinion words:

- Use a predefined lexicon; Wiebe and Riloff [14] constructed such an opinion word-list. This approach combines the lexicon based method described above with the machine learning methods.
- Identify the words (mostly adjectives; see below) that describe a certain feature of a product in a text [12]. e.g. After nearly 800 pictures I have found that this camera takes *incredible* pictures.

### Adjectives

Wiebe noted in [15] that adjectives are good indicators for subjectivity in a document. According to these findings you can assume that documents only represented by their adjectives should do well in sentiment classification. Experiments where only adjective features are used, were done in [11, 16]. The results showed that you get better results when using all POS data. This doesn't mean that adjectives are bad sentiment classifiers, as adjectives only represent on average 7.5% of the text in a document.

Salvetti used WordNet to enrich the only-adjective feature vectors. He translated the adjectives into synsets of adjectives and used hypernym generalization on them (both synsets and hypernyms can be found using WordNet). Using this procedure he found a decrease in the accuracy of the sentiment classification, which was due to the loss of information produced by the generalization.

## 3.2.2 Machine Learning Techniques

### Supervised Methods

In order to train a classifier for sentiment recognition in text, classic supervised learning techniques (e.g. Support Vector Machines, naive Bayes Multinomial, Maximum Entropy) can be used. A supervised approach entails the use of a labelled training corpus to learn a certain classification function. The method that in the literature often yields the highest accuracy regards a Support Vector Machine classifier [11]. In the following section we discuss a selection of classification algorithms. They are the ones we used in our experiments described below.

(1) Support Vector Machines (SVM)

Support Vector Machines operate by constructing a hyperplane with maximal Euclidean distance to the closest training examples. This can be seen as the distance between the separating hyperplane and two parallel hyperplanes at each side, representing the boundary of the examples of one class in the feature space. It is assumed that the best generalization of the classifier is obtained when this distance is maximal. If the data is not separable, a hyperplane will be chosen that splits the data with the least error possible.

(2) Naive Bayes Multinomial (NBM)

A naive Bayes classifier uses Bayes rule (which states how to update or revise believes in the light of new evidence) as its main equation, under the naive assumption of conditional independence: each individual feature is assumed to be an indication of the assigned class, independent of each other. A multinomial naive Bayes classifier constructs a model by fitting a distribution of the number of occurrences of each feature for all the documents.

(3) Maximum Entropy (Maxent)

The approach tries to preserve as much uncertainty as possible. A number of models are computed, where each feature corresponds to a constraint on the model. The model with the maximum entropy over all models that satisfy these constraints is selected for classification. This way no assumptions are made that are not justified by the empirical evidence available.

### Unsupervised and Weakly-supervised Methods

The above techniques all require a labelled corpus to learn the classifiers. This is not always available, and it takes time to label a corpus of significant size. Unsupervised methods can label a corpus, that is later used for supervised learning (especially semantic orientation is helpful for this [17]). Turney's technique using AltaVista (see section 3.1.1.1) can be viewed as a form of weakly supervised learning, where a set of seed terms is expanded to a collection of words. We mention two more methods for determining the sentiment of single words based on weakly-supervised methods. Hatzivassiloglou and McKeown[18] presented a method for determining the sentiment of adjectives by clustering documents into same-oriented parts, and manually label the clusters positive or negative. OPINE [19] is a system that uses term clustering for determining the semantic orientation of an opinion word in combination with other words in a sentence. The idea behind this approach comes from the fact that the orientation of a word can change with respect to the feature or sentence the word is associated (e.g. The word *hot* in the pair: *hot water* has a positive sentiment, but in the pair *hot room* it has a negative sentiment).

# 4 Challenges

With the techniques described above, pretty good results can be obtained already (see section 5), but nevertheless, there are some challenges that need to be overcome.

## 4.1 Topic-Sentiment Relation

Our goal is to determine sentiments towards a certain topic. It often happens that a person expresses his opinion towards several topics within the same text or sentence. For example, in a movie review he may state he dislikes the special effects and some of the acting, but likes the movie nonetheless. His opinion about these topics is in contradiction with his thoughts about the movie in general. When a sentence contains a lot of negative subjectivity, but all expressed toward a different topic than the one we are investigating, the sentence is still classified as negative. Therefore, it is useful to investigate the relation of the sentiment to the topic. One way of doing this is by looking into the sentence parse tree (i.e. a syntactic analysis of the sentence according to the language's grammar) to derive better features.

Related to this problem is the classification of whole texts. Until now we have only looked at the classification of sentences, in which topic and terms indicative for the sentiment are assumed to appear together. This is however not a realistic assumption. For the detection of the topic-sentiment relation in texts, coreference resolution needs to be applied across sentences. Even when there is only one topic in the text, it is also advantageous to use a more advanced metric to combine the predictions for the sentences than a simple sum of the sentiments found in the individual sentences. Taboada and Grieve [20] state that opinions expressed in a text tend to be found in the middle and the end of that text. Therefore, they weigh the semantical orientation of a sentence based on its position in the text, giving improved results.

## 4.2 Neutral Text

A first question is what to do with neutral text, as not all text is either positively or negatively oriented. It is often useful to determine whether a piece of text expresses subjective or objective content. Subjective sentences are used to communicate the speaker's evaluations, opinions, emotions and speculations, while objective sentences are used to convey objective, factual information [21]. Both kinds often appear in the same text, for instance in movie reviews, where the writer can express his attitude toward the movie (which is the semantic orientation of the document), but can also describe, within the same review, objective statements about the movie itself (e.g. a summary of the plot). Most subjectivity classifiers use machine learning techniques (see [22]) and classify between subjective and objective sentences or between positive, negative and objective sentences. To our best knowledge, there has been only one attempt to use a symbolic technique that classifies subjective sentences, done by Wiebe [23, 24].

## 4.3 Cross-domain Classification

Other research in the sentiment classification field regards cross-domain classification. How can we learn classifiers on one domain and use them on another domain (e.g. *books* and *movies*)? A reason why cross-domain classification might be necessary is because there is not always enough training data available to train a classifier for a specific domain. The classifier should then be trained with data from another domain. Tests are done by Aue et al. [25] and by Finn et al. [26]. Overall, they show that sentiment analysis is a very domain-specific problem, and it is hard to create a domain independent classifier. One possible approach is to train the classifier on a domain-mixed set of data instead of training it on one specific domain.

## 4.4 Text Quality

A last important issue is the quality of the text to be evaluated. When text is automatically gathered from the World Wide Web, one can expect a fair amount of junk to be returned (e.g. adds, web site menus, links, ...). This junk may be mixed with other information we are interested in, making it more difficult to filter it out. Also the language used by the writers may be of poor quality, containing lots of Internet slang and misspellings. Both issues have a negative influence on the classification for both types of methods discussed. Especially the junk may confuse a machine learner by providing it with a lot of irrelevant features. This also means extensive manual filtering of the text in order to acquire a good training corpus, and makes it harder to perform deeper

NLP techniques like parsing. An example of dirty input text (the topic is the movie “A Good Year”) is the following:

*Nothing but a French kiss-off Search Recent Archives Web for (rm) else &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; ONLINE EXTRAS SITE SERVICES Movie Listings Friday Nov 10 2006 Posted on Fri Nov. 10 2006 MOVIE REVIEW A Good Year a flat bouquet Nothing but a French kiss-off Gladiator collaborators seem defeated by light-weight love story.By ROBERT W.*

Needless to say that using a sentence parser (that detects the syntactic structure of the sentence) on this example will have little success.

# 5. Results

## 5.1 Evaluation Measures

As a first evaluation measure we simply take the classification accuracy, meaning the percentage of examples classified correctly. This measure is not sufficient when we classify individual sentences and include the neutral class as a third option next to the positive and negative ones. Neutral examples are a majority in texts and their correct detection largely influences simple accuracy results. With this in mind, it makes more sense for us to use precision for positive and negative examples as the evaluation measure. When generating sentiment statistics a high recall is just as desirable as a high precision. Other evaluation metrics that influence the performance are also considered: the speed of the classification method, the feature vector size and the available resources.

## 5.2 Symbolic Techniques

In the previous section we have seen a selection of two symbolic techniques, for which we give concrete results. Turney reports accuracies ranging between 65.83% on a collection of movie reviews, to 84.0% on a collection of automobile reviews when applying his method using a web search engine. Kamps and Marx achieved an accuracy of 68.19% on the manually constructed list of the General Inquirer for classification along Osgood's evaluative dimension, when applying their approach using WordNet. The accuracy rose to 76.72% when increasing the interval for which words are considered neutral.

An interesting experiment done by Pang et al. [11] shows the difficulty to construct a lexicon (or another knowledge-resource) that has a (close to) complete coverage of the target domain. A lot of information is often not captured in the hand-built model and lost. In the experiment they compared the ability of humans in selecting appropriate words for an emotion lexicon, with automatic methods. Although the lists of words created by humans seemed intuitively valid, they resulted in poorer performance: the best human created list resulted in 64% accuracy (with 39% ties), while a simple automatic method (a count of the frequencies of words in positive and negative reviews) resulted in a list with 69% accuracy and only 16% ties. Interestingly, some words that have no significant emotional orientation were quite good indexes. For example, the word “still", was found to be a good indicator of positive orientation, because it appeared in sentences such as “Still, though, it was worth seeing".

Given the above results, we did not perform any experiments with symbolic techniques, instead we focused on machine learning techniques of which the results are given below.

## 5.3 Machine Learning Techniques

### 5.3.1 Corpora

A corpus is a large, electronically stored set of texts. Corpora are used by machine learning approaches both for training and testing (and just for testing in the case of symbolic approaches). Evaluation will often be performed by using cross-validation. This means that over several iterations, in each iteration part of the corpus will be used for training, and the other part for testing. After all iterations, each example from the corpus will have been used for testing once, resulting in a full evaluation of the corpus. In order to compare results of different approaches, they need to be compared on the same corpus, as some corpora can be considerably easier to work with than others. We performed tests on two corpora to obtain the results presented in this paper:

- Pang and Lee's[3] movie review corpus, consisting of 1000 positive and 1000 negative reviews, is often used to evaluate sentiment analysis approaches in the literature. These movie reviews seem hard to classify. A possible explanation of this phenomenon is the mix of words that describe the storyline and words that describe the evaluation of that particular movie.
- A corpus gathered from blogs, discussion boards and other websites containing 759 positive, 205 negative, 1965 neutral and 1562 junk examples, annotated with a sentiment towards the topic under evaluation. The latter two categories were considered as one for our test purposes. The topics include various movie titles and car brands. The examples are of poor quality, displaying the problems described in section 4.4 (the example given there was taken from this corpus). As the number of examples in each category is very unbalanced, corrective measures were taken by adding additional examples from the Customer Review Datasets corpus by Hu and Liu[4]. In total, 550 negative sentences from the customer reviews were added to the corpus, and 222 extra positive sentences were used for training only.

### 5.3.2 Our Experiments

In Table 1 we show some of our results on the movie review corpus, indicating the features that perform well in the literature (discussed above), optional processing and the machine learning methods used. For both the support vector machine (SVM) and naive Bayes multinomial (NBM) methods the Weka[5] implementation was used, the Maxent[6] package from OpenNLP was used as implementation of the maximum entropy classifier. For our tests using SVM's, an error tolerance of 0.05 was set for training, the other parameters (e.g. linear kernel) were kept default for all methods. We used QTAG as POS tagger for obtaining the adjectives. It achieves a rather low accuracy[7], but it is fast and easy to incorporate into software. “Subjectivity analysis” stands for a simple subjectivity analysis using a NBM classifier, trained on the subjectivity dataset introduced in Pang and Lee [22], which removes all objective sentences from the examples. A cut-off of four was used for the bigram feature, meaning that only bigrams occurring at least four times were included in the feature vector. Frequencies of the features were used in the feature vector for SVM and NBM, while binary feature presence was used for Maxent.

| ***Features*** | ***SVM*** | ***NBM*** | ***Maxent*** |
|---|---|---|---|
| Unigrams | 85.45% | 81.45% | 84.80% |
| Unigrams & subjectivity analysis | 86.35% | 83.95% | 87.40% |
| Bigrams | 85.35% | 83.15% | 85.40% |
| Adjectives | 75.85% | 82.00% | 80.30% |

**Table 1: Results in terms of accuracy on the movie review corpus for different machine learning methods using a selection of features (and processing)**

Table 2 shows our results on the second corpus that realistically represents blogs found on the World Wide Web. The corpus was extended with 550 negative review sentences, which are included in the results. In the first column are the baseline results on the corpus. The baseline uses the approach that gives the best results for the movie corpus (see Table 1), i.e., an approach comparable to the literature and with a low novelty factor. In the second column are our latest results. A total of 84 examples were beyond the reach of our current methods and are excluded from the results. In order to include those examples, we could consider them as neutral; resulting in a slight decrease in the total accuracy and in the recall for positive and negative, compared to the results shown in the second column, while still being much better than the baseline results. The methods, features and processing used to arrive to these results may not be disclosed by us. For more information on the methods used, the reader may contact Attentio, the company that sponsors our research.

[3] Available at http://www.cs.cornell.edu/people/pabo/movie-review-data.

[4] Available at http://www.cs.uic.edu/~liub/FBS/FBS.html.

[5] See http://www.cs.waikato.ac.nz/~ml/weka/.

[6] See http://maxent.sourceforge.net/.

[7] Our own experiments indicate an accuracy of about 86%, while current state of the art POS tagging achieve ca. 96% accuracy.

| | ***Baseline NBM*** | ***Our latest approach*** |
|---|---|---|
| accuracy % | 84.25 | 90.25 |
| precision/recall % for positive | 64.52/49.93 | 74.39/75.62 |
| precision/recall % for negative | 88.48/72.96 | 87.43/82.70 |

**Table 2: Results on the blog corpus, comparing the results of the baseline approach (cf. Table 1) and those of our latest methods**

# 6 Discussion

Although we have not done any experiments using symbolic techniques ourselves, we deemed machine learning approaches more promising after reviewing methods from both categories, and conducted our research in that direction. Judging from the good results we have achieved, this seems like it has been the right choice.

The results in Table 1 show that there is rather little difference in accuracy between the experiments using different features (except for the adjectives). With this in mind, it becomes interesting to look at other factors influencing the choice of which features and processing to use. The advantages of unigrams and bigrams over the other features are that they are faster to extract, and require no extra resources to use, while e.g. adjectives require a POS tagger to be run on the data first, and subjectivity analysis requires an additional classifier to be used. A downside is the feature vector size, which is substantially (over 5 times for unigrams) larger e.g. than when only adjectives are included. For the machine learning method we see a more substantial difference between NBM and both SVM and Maxent. It might however still be advantageous to use NBM, as it is considerably faster. The results of the state of the art techniques for sentiment classification on the movie review corpus shown in Table 1 are comparable with the ones found in the literature that use this corpus.

The results from Table 2 need some more explanation. The blog corpus used in the experiments of Table 2 is considerably more difficult to work with, and is annotated in three classes (including neutral), where the movie review corpus (results in Table 1) only had two. However, compared to the baseline (current state-of-the-art) approach, our latest method performs significantly better. The lower precision and recall for the positive class compared to the negative one, are due to the added negative examples from the easier Customer Review Datasets corpus, and due to the higher correlation of positive examples with neutral ones, making misclassifications between those classes more common. The results we obtained are encouraging, and show that it is possible to overcome the difficulties explained in section 4.

# 7 Conclusion

In this paper we have indicated the usefulness of sentiment classification, and have given an overview of the various methods used for this task. While many of the methods show encouraging results, there are still challenges to be overcome when applying them to data gathered from the World Wide Web, especially from blogs. We have demonstrated that in these circumstances improvements over state of art methods for sentiment recognition in texts are possible.

# References

[1] OSGOOD, C. E.; SUCI, G. J; TANNENBAUM, P. H. *The Measurement of Meaning*. University of Illinois Press, 1971 [1957].

[2] BIBER, D; FINEGAN, E. Styles of stance in english: *Lexical and grammatical marketing of evidentiality and affect*. Text 9, 1989, pp. 93-124.

[3] WALLACE, A. F. C.; CARSON, M. T., *Sharing and diversity in emotion terminology*. Ethos 1 (1), 1973, pp. 1-29.

[4] HATZIVASSILOGLOU, V.; WIEBE, J., *Effects of adjective orientation and gradability on sentence subjectivity*, Proceedings of the 18th International Conference on Computational Linguistics, ACL, New Brunswick, NJ, 2000.

[5] TURNEY, P., *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*, Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 417–424.

[6] FELLBAUM, C. (ed.), *Wordnet: An electronic lexical database*, Language, Speech, and Communication Series, MIT Press, Cambridge, 1998.

[7] KAMPS, J.; MARX, M.; MOKKEN, R. J.; DE RIJKE, M., *Using WordNet to measure semantic orientation of adjectives.* LREC 2004, volume IV, pp. 1115—1118.

[8] MULDER, M.; NIJHOLT, A.; DEN UYL, M.; TERPSTRA, P., *A lexical grammaticaimplementation of affect*, Proceedings of TSD-04, the 7th International Conference Text, Speech and Dialogue, Lecture Notes in Computer Science, vol. 3206, Springer-Verlag, Brno, CZ, 2004, pp. 171–178.

[9] DAVE, K.; LAWRENCE, S.; PENNOCK, D. M. *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews.* In Proceedings of WWW-03, 12th International Conference on the World Wide Web, ACM Press, Budapest, HU, 2003, pp. 519–528.

[10] PEDERSEN, T. *A decision tree of bigrams is an accurate predictor of word sense.* In Proceedings of the Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics, 2001, pp. 79–86.

[11] PANG, B.; LEE, L.; VAITHYANATHAN, S. *Thumbs up? Sentiment classification using machine learning techniques.* In Proceedings of EMNLP-02, the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Philadelphia, US, 2002, pp. 79–86.

[12] HU, M.; LIU, B. *Mining opinion features in customer reviews.* In Proceedings of AAAI-04, the 19th National Conference on Artificial Intellgience, San Jose, US, 2004.

[13] BETHARD, S.; YU, H.; THORNTON, A.; HATZIVASSILOGLOU, V.; JURAFSKY, D. *Automatic extraction of opinion propositions and their holders.* In James G. Shanahan, Janyce Wiebe, and Yan Qu, editors, Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, Stanford, US, 2004.

[14] RILOFF, E.; WIEBE, J.; WILSON, T. *Learning subjective nouns using extraction pattern bootstrapping.* In Walter Daelemans and Miles Osborne, editors, Proceedings of CONLL-03, 7th Conference on Natural Language Learning, Edmonton, CA, 2003, pp. 25–32.

[15] WIEBE, J. *Learning subjective adjectives from corpora.* In Proceedings of AAAI-00, 17th Conference of the American Association for Artificial Intelligence, AAAI Press / The MIT Press, Austin, US, 2000, pp. 735–740.

[16] SALVETTI, F.; LEWIS, S.; REICHENBACH, C. *Impact of lexical filtering on overall opinion polarity identification.* In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, Stanford, US, 2004.

[17] BEINEKE, P.; HASTIE, T.; VAITHYANATHAN, S. *The sentimental factor: Improving review classification via human-provided information.* In Proceedings of ACL-04, the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, ES, 2004, pp. 263–270.

[18] HATZIVASSILOGLOU, V.; MCKEOWN, K. R. *Predicting the semantic orientation of adjectives.* In Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Madrid, ES, 1997, pp. 174–181.

[19] POPESCU, A.; ETZIONI, O. *Extracting product features and opinions from reviews.* In Proceedings of HLT-EMNLP-05, the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing, Vancouver, CA, 2005.

[20] TABOADA, M.; GRIEVE, J. *Analyzing appraisal automatically.* In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, Stanford, US, 2004, pp. 158–161.

[21] WIEBE, J.; BRUCE, R. F.; O'HARA, T. P. *Development and use of a gold-standard data set for subjectivity classifications.* In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, College Park,US, 1999, pp. 246–253.

[22] PANG, B.; LEE, L. *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.* In Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Barcelona, ES, 2004, pp. 271–278.

[23] WIEBE, J. *Recognizing Subjective Sentences: A Computational Investigation of Narrative Text.* Technical report, SUNY Buffalo Dept. Of Computer Science, Buffalo, NY, 1990.

[24] WIEBE, J. *Tracking point of view in narrative.* Computational Linguistics, 20 (2), 1994, pp. 233–287.

[25] AUE, A.; GAMON, M. *Customizing sentiment classifiers to new domains: a case study.* In Submitted to RANLP-05, the International Conference on Recent Advances in Natural Language Processing, Borovets, BG, 2005.

[26] FINN, A.; KUSHMERICK, N. *Learning to classify documents according to genre.* J. American Society for Information Science and Technology, Special issue on Computational Analysis of Style, 57(9), 2006.

# A Comparison of the Blogging Practices of UK and US Bloggers

*Sarah Pedersen*

Department of Communication and Media, The Aberdeen Business School
The Robert Gordon University, Aberdeen, UK
e-mail: s.pedersen@rgu.ac.uk

## Abstract

This paper describes the results of an investigation into the differences and similarities between the blogging techniques of UK and US bloggers undertaken in the winter and spring of 2006-7 and funded by the UK Arts and Humanities Research Council. Blogging started in the US, while British bloggers are relative latecomers to the blogosphere. How has this late arrival impacted on the ways in which Britons blog in comparison to US bloggers? A survey was administered to 60 UK and 60 US bloggers and data was also collected directly from their blogs and by means of online tools. A blog was also set up in order to discuss the findings of the research within the blogosphere. Since blogging started in the US, the majority of research into blogging so far has focused on the US and it is suggested that this focus has resulted in all bloggers being defined through the US experience. The findings of this project suggest that bloggers outside the US may have different approaches to blogging and find different satisfactions. It also suggests a new financial motivation for blogging, which had not previously been identified, and which may be an indication of the way in which the blogosphere is evolving.

**Keywords:** blogs; weblogs; comparitive analysis

# 1 Introduction

This paper reports on the first attempt to compare and contrast the blogging practices of US and UK bloggers. While there is a growing body of academic research into this new form of computer-mediated communication, the vast majority of such research has so far focused on North American bloggers. Blogging started in the US while British bloggers are relative latecomers to the blogosphere. The paper asks how this later arrival has impacted on the ways in which Britons blog in comparison to US bloggers.

Blogging has joined e-mail and home pages as a mass use of the internet. Blogs are usually defined, following Blood as 'frequently updated, reverse-chronological entries on a single webpage'.[1] The original blogs were filter-type web pages, directing the reader to other blogs and websites on the Internet and offering commentary and often the opportunity for readers' discussion. Blogging took off as a publishing platform, at first mainly in the US and predominantly amongst students or recent graduates (see Schiano et al [2]; Herring et al [3]). The expansion of the blogosphere that we see today, when the blog-tracking website Technorati claims to be tracking over 70 million blogs, occurred after the introduction of cheap and easy-to-use build-your-own-blog software such as Blogger, Pitas and Groksoup in 1999. Unlike early bloggers, who needed advanced programming skills to construct their blogs, it is now possible for anyone with access to the Internet to set up their own weblog. Filter blogs have been joined by so-called 'journal' blogs, which tend to have fewer links, fewer readers and are more like online, public diaries. However, both types of blog conform to the pattern of frequently updated posts arranged in a chronological order, the majority offering the opportunity for readers to post comments. The ability to simply publish a weblog online is ever increasing with community websites such as MySpace adding blogging to their services.

Although the first blogs did not appear until 1997, there has been a remarkably swift growth in academic research into this new form of computer-mediated communication. Early research focused on the categorisation of blogs or bloggers. For example, Krishnamurthy proposed the classification of blogs into four basic types along two dimensions: personal versus topical and individual versus community.[4] Another focus for scholarly research has been the on-going debate about the role of blogging as a form of journalism (for examples of the research on this topic see Singer; Matheson; Kahn and Kellner; Wall; Pedersen and Chivers)[5]. At the other end of the blogosphere, research teams lead by Herring, Schiano and Nardi have investigated journal bloggers, and in particular pointed out that, despite the media's focus on blogs written by white, educated, US males, over 50% of journal blogs are actually written by women and young people.[6] Other research has investigated the dynamics of different communities of bloggers; for example, Huffaker and Calvert have surveyed teenage bloggers[7] while

Mortensen and Walker discussed the way in which academics can use blogs as a research tool.[8] As Thelwall points out, ‘It is difficult to summarise the findings of the extremely diverse body of blog research except by pointing to the wide variety of uses of blogs and the fact that blogs do sometimes create genuine online communities’.[9] One thing that does immediately stand out when reading through early research into blogging is the focus on North American bloggers. There has been a very limited amount of research into the second wave of blogging that occurred outside the US, although this is now being addressed, for example, Trammell et al’s recent examination of the state of the Polish blogosphere, Tricas-Garcia & Merelo-Guervos’ work on the Spanish blogosphere and Abold’s discussion of the use of blogs in the 2005 German election campaign.[10] There has been a limited amount of research into the UK blogosphere. Discussion of gender issues within the UK blogosphere has been undertaken by Pedersen & Macafee while Auty has investigated the blogs of UK politicians and Thelwall undertook a descriptive analysis of blog postings around the London bomb attacks of July 2005.[11]

## 2 Methodology

Following a pilot stage of the project, a survey was administered to 60 UK and 60 US bloggers. The bloggers were selected randomly through two blog directories: Britblog and Globe of Blogs. Both directories offered the possibility of selecting blogs by state or county and so it was possible to ensure that all regions of the UK and US was covered by the survey. The survey was distributed to equal amounts of male and female bloggers. The criteria for selection were that the blogger had to have posted on their blog within a month of the start of the selection process, that the blog was written in English, that the blogger was resident in either the UK or the US and was over the age of 18. Teenage blogging is acknowledged by most researchers to be a very different type of computer-mediated communication from that of adult bloggers, associated generally with use of community sites such as Bebo and MySpace, and therefore it was decided to focus only on bloggers over the age of 18, which also avoided many ethical issues. Data was also collected directly from the survey respondents’ blogs and by means of online tools (Technorati, The Truth Laid Bear and SurfWax). Areas investigated included average time spent blogging and when that blogging occurs; the promotion of blogs; attitudes to blogging and issues of privacy and openness related to the use of photographs and other personal material. A measure of success was devised (based on traffic, links and directory rankings). In addition, a blog related to the research was established. This gave the researcher first-hand experience of the challenges of blogging and also offered the opportunity for further data collection since the surveyed bloggers were invited to comment on the research as it was ongoing, an opportunity which they took up with enthusiasm.

## 3 Results

**Demographics**

It has already been stated that efforts were made to send the survey to an equal number of men and women. In the final analysis, the respondents were as follows: 32 UK males, 30 UK females, 32 US males and 28 US females. It should be noted that, during the period of research, two of the UK females actually moved to North America. Out of this random group, one of the UK women identified herself as a lesbian, one of the US males as gay and one of the UK men as a transvestite.

Age ranged from 18 to 73. 40% of the UK respondents were under 30, in comparison to 26% of the US respondents. 4% of UK respondents were over 56, with the US figure being 18%. Thus, for this random sample, the US bloggers were on average older than the UK bloggers. Differences between the two countries were also found in terms of educational attainment, with US respondents having achieved higher educational attainment. 47% of US respondents were educated to bachelors degree level, compared to 32% of the UK respondents, and 35% of US respondents held a postgraduate degree, in comparison to 18% of UK respondents. 28% of the UK respondents reported that their highest level of educational attainment was as a school leaver compared to only 10% of US respondents. To a certain extent, this must be linked to the youth of the UK respondents, although it should be noted that 11 UK respondents were currently undertaking education compared to 13 US respondents.

Previous studies into the blogosphere have characterised bloggers as usually educated to graduate level or beyond. However, these studies were based in the US and investigated the first wave of bloggers. Some of these studies even focused on university bloggers through their selection of survey participants. It may be, therefore, that the second wave of blogging outside the US is attracting a different type of person to the blogosphere. Is blogging in the UK more associated with youth culture?

As far as employment is concerned, 105 of the survey respondents stated that they were employed, with most of those not employed being either retired or looking after dependents in the home. The number of those employed on a part-time basis was evenly split across the two countries. It should be noted that the number of women working part time was much greater than the number of men in part-time employment: 40% (21 out of 52 women who stated that they were employed) in comparison to 11% of all employed men (6 out of 53). As will be seen later, this may be related to the higher number of women who were attempting to gain financial reward from their blogging.

**Blogging practices**

Respondents were asked where they did the majority of their blogging. Only 14% of all respondents indicated that they blogged at a workplace outside the house while 43% blogged at home. Interestingly, another 30% stated that they blogged at home, which was also their place of employment. It is possibly not surprising that few bloggers choose to blog at a workplace outside their home considering the number of high-profile cases of bloggers being sacked or reprimanded for blogging at or about their place of employment. However, it is noteworthy that such a high number of respondents *worked* at home, either fully or partially, and blogged from there. When this was discussed on the blog set up in association with the research, several bloggers suggested that it was only because they worked at home that they had the time to blog:

> *Speaking as a blogger who works from home, I sometimes wonder how a person who doesn't work from home would find significant time to blog. I have the luxury of an extra couple of hours which would otherwise be used as commute time to compose my thoughts, and blog. I also have intervals during the day where I can break from my client projects to respond to reader comments.*

A respondent to the survey agreed that blogging was easier because she worked as a freelance: 'I am self-employed so sometimes it [blogging] will be immediately after something happens in work' although another confessed that easy access to the Internet brought its own problems: 'Being self-employed I can almost do as I please (as long as I get the work done) and there have been times in the past when I've been obsessed with blogging and work would suffer as result.'

Thus respondents were more likely to blog at home than outside the home. This finding is also related to *when* they blogged, with the most popular choices being evenings and breaks in the day. US bloggers were somewhat more inclined to blog in the mornings before they went to work, with 35% of US respondents admitting to this in comparison with 17% of UK bloggers. Again, this was discussed on the blog with one US blogger suggesting: 'Americans are familiar with a fast-paced, over-time-heavy schedule, which means that getting up in the morning to blog is a convenient way to blog each day but still get to work and/or started on the home-based work before the 8am rush.' A report in *The Guardian* of 29 November 2006 of a survey conducted by the European Interactive Advertising Association also supports the finding that Europeans access the Internet later in the day:

> *The survey also looked at when people access the internet. From 6am to 10am the majority of European internet users prefer to listen to the radio or read a paper. But that picture inverts dramatically as the day wears on. From 5.30pm to 9pm, three-quarters of web users are watching TV but almost as many are accessing the internet.*[12]

68% of all respondents admitted blogging for up to 5 hours a week, with another 18% blogging for between 5 and 10 hours a week. One or two respondents were blogging for up to 35 hours a week, although it should be noted that some respondents were what might be called professional bloggers, either setting up blogs for others or using their blogs as part of newspaper columns. A surprising 52 respondents, spread evenly across the countries, admitted to writing more than one blog. The reason usually given for this was to focus on different subjects, although some bloggers kept one blog private, accessible only through the use of a password, while the other was public access. For example, one female blogger kept a blog about her pregnancy private while publicly blogging about food and cooking.

**Content of blogs**

The respondents' blogs were analysed for content. Unfortunately, only 112 of the blogs were able to be analysed in this way because during the eight-month period of research eight of the blogs were abandoned and removed from the Web. On the basis of the last ten postings made on the blog they were placed in one of the following categories: personal, creative work, criticism, politics and opinion, IT, business and work, religion, chance discoveries and food. It should be noted that it was sometimes difficult to distinguish between the IT and business and work categories since those who blogged about IT were usually also working in IT. Therefore the IT category should be seen as a subset of the business and work category.

51 out of the 112 blogs (46%) were categorised as 'personal'. It should be noted that far more female blogs were categorised as personal than male blogs (15 men and 36 women) and, in particular, only four US male blogs were categorised as personal. In contrast, five male blogs were characterised as being about religion, but none of the female blogs. 12 male blogs were characterised as opinion and politics, but only three female blogs. If work and business and IT are seen as one category, 21 blogs are found here: 5 of these were female blogs and the other 16 were male. 3 blogs (all female) were characterised as being about food.

**Blog promotion**

Success in the blogosphere is linked to popularity. The more links to a blog, the greater its success rating. Although it must be conceded that not all bloggers are interested in the type of success that comes with membership of the 'A list' (Technorati's list of the 100 most linked blogs), the survey did investigate how far bloggers promoted their blogs to other bloggers in order to encourage a higher readership and more incoming links. As far as promotion of their blog was concerned, the most popular methods used by respondents were to submit their blog to a blog directory or blog search engine (97 respondents) or to post on other blogs (66 respondents). Male US respondents were the most likely to submit an RSS feed to a blog directory or search engine, rather than merely submitting their URL, which perhaps suggests a higher level of technical ability in blogging.

A form of promotion particularly popular amongst the UK bloggers was blogrings. Blogrings connect a circle of blogs with a common theme or purpose. A link to the blogring is displayed on a blog and clicking on that link takes the reader to the blogrings page, where the other members of the blogring are listed. Alternatively, clicking on the link takes the reader directly to the next blog in the ring. UK respondents were more likely to state that they used blogrings to promote their blog. 26 UK respondents (11 men and 14 women) admitted to using blogrings in comparison to 15 US respondents, only five of whom were male. Analysis of respondents' blogs showed a large selection of blogrings with few being named by more than one or two bloggers. The more popular blogrings were either those which linked bloggers of the same sex, such as 'Blogs by Women' or 'Crazy/Hip Blog Mamas', or those which linked geographically similar bloggers. 24 bloggers linked to blogrings related to location, such as 'Blogging Brits', 'Scots Bloggers' or 'Expat Bloggers'. Male bloggers were more likely to belong to a blogring which promoted an interest or hobby, such as blogrings for birdwatchers, Methodists or transvestites, which reflects the male bloggers' preference for issue-based blogging, while female bloggers were more likely to belong to blogrings that celebrated their femininity (16 female bloggers belonged to female-only blogrings), which again reflects the female proclivity for more journal blogging with a focus on themselves.

Survey respondents were asked their opinion of membership of blogrings. While most acknowledged that they could provide more traffic, in terms of readers, to a blog, the opinion of many was that they were not worth joining any more, having been replaced in usefulness by blog directories. One respondent commented: 'I think blog rings can be a little random. I'd rather have a focused directory that points specifically to my site'. Others were concerned that it would be assumed that they would have identical opinions with others in the blogring. Several respondents explained that they had joined blogrings at the start of their blogging, but would not join any more now. The relative popularity of blogrings amongst the British bloggers – and the high number of blogrings related to the UK or regions of the country – is noteworthy in comparison with the lower interest from US bloggers, in particularly US male bloggers, and may point to a desire to mark themselves out as different, or a need to group together, in the face of the much more numerous US bloggers.

**Concerns about privacy**

56 respondents – just under half – had concerns about privacy. These were divided equally between the two countries. Privacy concerns tended to be about two areas of the bloggers' lives: their family and their work. Respondents reported that they tried not to mention their family on their blog or to make their address identifiable. Those that worried about colleagues or management at work identifying them, which might lead to trouble at work, also mentioned the worry that potential future employers might search for them online in order to assess their suitability for employment.

Are bloggers identifiable through the information they give on their blog? US female respondents were the least likely to state their full name on their blog, with only 14% of surveyed blogs giving this information in comparison to the rest of the surveyed blogs where the figure was around 50%. On average 70% of the blogs did not show an identifiable photograph. However, the US males again seem slightly different to the others. 54% of the US male blogs analysed did show an identifiable photo. From the anecdotal evidence given in the survey responses it seems that bloggers are right to be concerned about being identified through photographs. Several respondents told stories of being identified from the photos on their blog, including one man who was accosted by a complete stranger while walking through the departures lounge of an airport.

**Opinions on blogging**

Respondents to the survey were asked a number of questions about the way in which they perceived blogging in order to ascertain any differences in attitude between the two countries. Firstly, they were asked whether their blogging had replaced any sort of paper documentation. The most frequently reported replacement by a blog was a diary. 28 respondents agreed that they had replaced their diary by blogging. 17 stated that project journals had been replaced by a blog and 15 that a travel diary had been replaced by blogging.

There were some differences between the sexes to be discerned here. 19 of those who had replaced diaries were women, as opposed to 9 men, whereas 10 men had replaced travel diaries in comparison to 5 women. While this might conform to gender stereotyping to a certain extent, it is interesting to note that 12 women stated that blogging had replaced project journals as opposed to only 5 men.

Respondents were also asked whether they blogged mainly for their own records. Interestingly, UK female respondents answered this question very differently to all other respondents, with 50% declaring that they did blog mainly for their own records. Other respondents were far less likely to respond positively, for example only 4 (13%) of US males agreed that they blogged mainly for their own records.

Respondents were asked whether, in general, they considered blogging to be a form of publishing, journalism, creative writing, diary keeping or other. Overall there was uniformity in many responses, with many respondents selecting all four of the named choices. However, it should be noted that US males were particularly unlikely to see blogging as a form of diary keeping, with only 12 selecting diary keeping in comparison to 21 for UK males and females and 23 for US females.

Respondents were asked how they saw their own blogging activity. They were asked to select as many as necessary from a selection of statements. It is obvious from the results that blogging is seen very much as a leisure activity by respondents on both sides of the pond. 74 selected ‘Leisure time activity’ and another 60 selected ‘A welcome distraction’. The small amount of students amongst the respondents was again demonstrated by the low number who saw blogging as a quick break from, or an adjunct to, studying – only 7. Men were slightly more likely to see blogging as a quick break from work (23 to 10), and this is probably related to the higher numbers of male than female respondents in full-time employment.

What was particularly interesting in the response to this question was the number of respondents who indicated another way of looking at blogging: as a form of income generation. The work of teams led by Schiano and Nardi on the motivations of bloggers suggests that there are five main reasons for blogging. These are: documenting the author’s life; providing commentary and opinions; expressing deeply felt emotions; working out ideas through writing; and forming and maintaining communities and forums. They note that such motivations for blogging are not mutually exclusive. Pedersen’s work on the motivations of women bloggers suggests that another motivation may be the women’s need for validation of their thoughts and actions.[13] However, this survey has brought a further motivation to light: that of financial reward. Among many responses to the question of why respondents blogged along the lines of the motivations outlined by Schiano et al was the introduction of a financial motive. Preliminary findings from this research have suggested that the financial motivation is particularly strong amongst women bloggers, who may be looking for ways in which to generate income as an alternative to full-time employment outside the home. Of the 31 respondents who mentioned a financial motivation in their written responses to the survey, 21 were women, and their responses showed very clearly that they were hoping that their blogging would lead to some sort of financial gain. As one female respondent stated: ‘I hope to eventually make enough money from my blog to support my family, I see it as the beginnings of an online business.’

The ways in which bloggers hoped to make money through their blogging differed. Some bloggers used their blog as a marketing tool for themselves or for their businesses. 24 respondents agreed that their blogging brought custom for their business. For example, one UK female blogger stated: ‘I started the blog as a way of promoting my online business, enhancing online word-of-mouth marketing for my business and developing my brand’. Another, who blogs about parenthood, stated that her blogging had started as a leisure activity but was now opening up serious work opportunities. One respondent, who worked as a children’s book illustrator, reported that she showcased her work and sold associated greetings cards through her blog. Another respondent, who described herself as a courtesan, explained that her blog helped attract suitable clients.

Blogging might also offer direct financial reward as a profession – one UK respondent worked as a freelance blogger, setting up blogs for West End shows and individual actors. An American blogger reported that her blogging 'started out as a leisure time activity and has become my work. The postings on my blog are the same as the reviews that now appear in my syndicated column of movie reviews, which appear in various newspapers across the Northeast, thanks to a deal made with a company that saw the work on my blog and hired me to be their critic.' Interestingly enough, her blog was one those ranked as least popular by this research (see below), indicating that unpopularity online by no means translates into lack of success elsewhere.

Blogs can also make money through carrying advertising or requesting subscriptions. One of the most famous bloggers on the web is Heather B Armstrong, the author of the blog Dooce.com, who reportedly supports her entire family through the advertising that her blog carries. While none of the respondents to this survey mentioned such large financial earnings, a UK male respondent's blog carries a section offering the possibility of running a banner advertisement at the top of his blog for a month with the guarantee that no other advertising will be accepted during this time. He charges £200 for this privilege. As well as carrying advertisements on their blogs, bloggers might also earn money through 'pay-per-post' advertising where bloggers write about certain products or services in their blogs in return for payment. Bloggers might even hope for income through the paper publication of their entire blog. Blogs which have been successfully published as books include *Belle de Jour: Intimate Adventures of a London Call Girl* or Tom Reynolds' *Blood, Sweat and Tea: Real Life Adventures in an inner-city ambulance* (taken from his blog 'Random Acts of Reality'). Recent press coverage in the UK has focused on the £70,000 book deal given to ex-*Sunday Times* education correspondent Judith O'Reilly for her blog *Wife in the North*. Several respondents to the survey mentioned hopes that their blogging would attract potential publishers: 'I have aspirations to write a book about the food industry and I believe that writing the blog is a tool to (1) exercise my writing muscles and developing a voice; (2) distinguishing or creating a unique vice; (3) offer me opportunities for credibility and to be viewed as a subject matter expert.' In fact, one male respondent from the UK reported that his blogging had helped clinch a book publishing contract for a book on his subject specialism.

In contrast, a few respondents reported that blogging had actually lost them money. One US male felt that the strongly held views of the government policy he discussed on his blog had lead to loss of work from the defence industry. He also considered that blogging was a threat to his career as a journalist: 'Sadly I feel my work abilities (writer, reporter, photojournalist) are going to go the way of the dinosaur. Note the rise in cheap digital stock imagery, blogging, "citizen" journalist submissions to network and cable TV, websites, publications, etc. (all for free, mind you).'

There were some differences between UK and US bloggers when discussing the gains – financial or otherwise – to be found in blogging. US bloggers were far more willing to acknowledge that they found blogging 'useful'. 31 US bloggers stated that they found blogging useful because it widenened the audience for their intellectual work, in comparison with 14 UK respondents, and 44 US respondents felt that it widened the audience for their creative work, in comparison with 26 respondents. UK respondents were far more likely to respond that blogging had no use. One stated: 'It's not the pretentious thing you seem to think it is. It's sharing. It's putting yourself out there. Not for recognition or to "help" people, although that might happen on occasion. It's not there for me to make people like or respect me. It's just me, warts and all. No other agenda.'

**Blogrolls**

A blogroll is a collection of links to other blogs and is seen as a list of recommended reading. The majority (82) of the blogs surveyed for this project offered their readers access to a blogroll, although interestingly more than 82 survey respondents answered the questions about their blogroll. Respondents were asked what they had in common with the contacts on their blogroll. The most popular choice here was 'Interests' (92 respondents). 59 respondents, just under half, also chose 'A sense of humour'. The least popular choice was 'Economic or domestic circumstances', with only 8 respondents. Interestingly, bearing in mind the popularity of blogrings which linked bloggers located in the same geographic region, 'Part of the world' was also an unpopular choice with only 19 respondents. Bearing this in mind, an analysis of the blogrolls of all respondents was undertaken in order to ascertain how willing bloggers were to link to blogs from outside their own country.

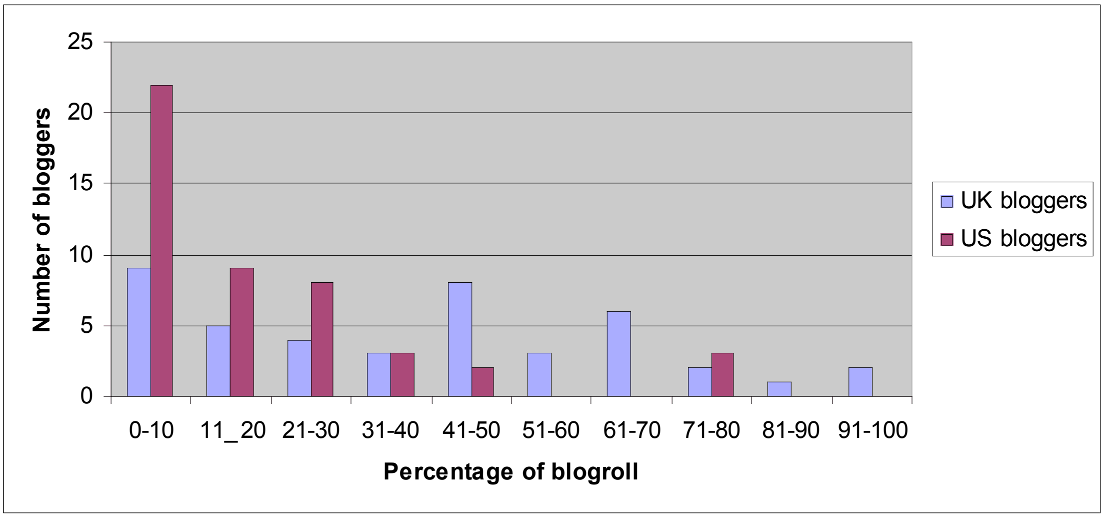


**Figure 1: Percentage of blogrolls containing links to foreign blogs**

Of the 47 US blogs which carried a blogroll, 31 (66%) had less than 20% of their blogroll links to blogs from outside the US. 15 of these bloggers had no links at all to blogs from outside the US. The male blogger with the most links to outside the US was actually a German expat living in the US who wrote a blog on international affairs and culture, primarily German. The female blogger with the most links to blogs located outside the US wrote a blog about the English author Jane Austen. Only three bloggers had more than 50% of links in their blogrolls to blogs outside the US.

Of the UK blogs, 14 out of the 43 which featured blogrolls had less than 20% of their blogroll linked to blogs from outside the UK. Of these, 7 had no links to any blogs located outside the UK. 14 bloggers had more than 50% of their links to blogs outside the UK. It is not that surprising that UK bloggers link more to blogs outside the UK since there ARE more blogs outside the UK. Riley estimated in July 2005 that there were 2.5 million British bloggers, compared to up to 30 million US bloggers, although there are difficulties in enumerating specifically British blogs, because of what Riley calls 'the Anglosphere problem', i.e. the existence of a common body of service providers and readership across the English-speaking internet.[14] However, the limited amount of linking that the average US blogger does to blogs outside the US should be noted.

While bloggers might prefer to link to other bloggers in their own country, there was less evidence that they preferred to link to bloggers within their own state or town. 64% of all bloggers had less than 10% of their blogroll devoted to links to others in their area. Only 8% had more than 50% of the links in their blogroll devoted to local bloggers. However, all four of the bloggers whose links were 100% local were from the United States and only two UK bloggers had more than 50% of their webroll devoted to local links.

## Ranking in terms of popularity

Using data gathered from the blog-monitoring sites Technorati and The Truth Laid Bear and information concerning the number of links made to a blog's front page from Surfwax, the 120 survey respondents' blogs were ranked in terms of popularity. The Truth Laid Bear (http://truthlaidbear.com/) and Technorati (http://www.technorati.com) are websites that use links from other blogs as the measure of the relative worth of a blog. Surfwax is a metasearch engine whose Site Snaps function offers a quick abstract of any web page, including the number of links made to that page. Since popularity, as demonstrated by number of links, is used as the main criteria for success in the blogosphere, the surveyed blogs were ranked using the data collected and in the top and bottom twenty in the listing were analysed to discover common characteristics.

The top 20 blogs were as follows: 12 US respondents (10 males and 2 females) and 8 UK respondents (4 males and 4 females). The bottom 20 blogs were 8 US respondents (1 male and 7 female) and 12 UK respondents (6 male and 6 female). What is suggested by this exercise is that the survey's US male respondents are on average more successful in the blogosphere than the other three groups. This finding corresponds to the general tenor of research findings about gender in blogging. (For references to the extensive online debate, see Pollard; Ratcliff; Garfunkel [15]). Ratcliff has recently produced evidence that men's postings receive more comments than women's.[16] Meanwhile, Henning suggests that women's blogs make up only 15% of all blogrolls [17]. It has also been claimed, in the North American context, that a greater amount of attention is given in the media to male bloggers (Herring et al[18]).

Surfwax data was also used to investigate the bloggers in terms of number of links, number of images used and number of words used in their blogs. In terms of number of links, again the US males dominated with six in the top ten. They included a birding enthusiast, an evangelical Christian, an expert in global current affairs, the expert in German culture and an expert in betting on American football. The two UK males included another Christian, this time a minister, and a blogger with a long blogroll relating to mental illness. The female bloggers were both promoting their businesses through the Internet, one as a children's book illustrator and the other as a sex therapist. If we are therefore seeing a high number of links in the blog as evidence of success, again we have more successful American male bloggers, but it is also obvious that bloggers who focus on one particular subject, which may or may not be a career or source of income for them, are the most active in terms of links. Out of the ten most successful bloggers, five were blogging about some aspect of their career.

In terms of the number of images used on the blogs, the top ten bloggers included seven US males, two UK females and one US female. Four of the top bloggers here were also in the top links list above: the birding expert, the international affairs expert (who writes for a variety of magazines and journals on the subject), the evangelical Christian and another blogger whose blog focuses on funny and strange things to be found on the Internet. One UK female blogger uses her blog as part of her online shop which sells objects for the home and therefore illustrations and photos are very necessary. It appears that US bloggers are happier to use photos on their blogs than UK bloggers.

The top ten blogs with the most amount of words were those belonging to six US males, two UK males and one US female and one UK female. Again the blogs with the most words are dominated by those with a theme or focus. Of the two female bloggers, one discussed right-wing politics while the other reviewed crime novels. One of the two UK males wrote about military affairs, having been a soldier, while the other was a policeman writing anonymously about policing in the UK. Of the five US males, two were religious bloggers, one wrote about international affairs, one was the expert on the subject of American sports and betting, one was a solider writing about military affairs and one was the German expat blogging about international culture.

As can be ascertained from the above descriptions, many of the bloggers who were in the top ten for amount of words were also in the top ten blogs for either use of images and links. The two bloggers who were in the top ten for everything were a young, evangelical Christian male (US) and a blogger who wrote about international politics, with an emphasis on technology (US male). Bloggers in the top ten for at least two out of three: were the US male birding enthusiast; the German expat living in the US and writing about cultural issues; the group blog on American sports and betting (US male); a US male minister writing from a Christian viewpoint. Thus all the particular dominant bloggers in the survey, according to Surfwax data, were US males.

## 4 Conclusions

This project set out to compare and contrast the blogging techniques of UK and US bloggers. However, what it has discovered is noticeable differences between US males and the rest of the blogosphere. The US male bloggers surveyed dominated the rankings as far as links, use of images, amount of words and overall popularity (as defined by Technorati and The Truth Laid Bear) are concerned. The content of the US male blogs was also more likely to focus on an interest, business or hobby and less likely to be categorised as personal. They were least likely to write a blog purely for their own records or to see blogging as a form of diary keeping. In other words, US male bloggers were less likely to write 'journal' blogs, which confirms the findings of Herring et al in their analysis of such blogs. In terms of the debate about the dominance of male bloggers in the US blogosphere, it has been suggested that men are more likely to blog about external events, rather than personal ones, and are therefore more likely to be found by prospective readers when using a search engine and thus more likely to be linked to, raising their popularity ranking. The US male bloggers also seemed less concerned about privacy, giving their full name and showing identifiable photographs of themselves more frequently. Thus the main finding of the project is that the dominance of male bloggers in the US, as identified by many commentators in the last few years, also translates into a dominance of the international, anglophone blogosphere.

Some statements can be made about differences between the UK and US bloggers surveyed. The US bloggers were on average older than the UK bloggers and differences between the two countries were also found in terms of educational attainment, suggesting that the picture of bloggers as, on average, educated to graduate level gained from earlier US-based studies needs to be questioned by more research into the blogosphere outside the US.

Bloggers were much more likely to blog at home than at work, although US bloggers were more likely to blog in the morning than UK bloggers. This finding concurs with other research into Europeans' use of the Internet.

In terms of the promotion of their blog, UK respondents were more likely to use blogrings to promote their blog. Whilst US respondents tended to dismiss blogrings as of less use than blog directories, UK bloggers were happier to use them, in particular those that identified the blogger as part of the UK or its regions. Given the very different sizes of the US and UK blogosphere, this may well be in order for UK bloggers to identify each other and to maintain a sense of a UK identity against the overwhelming US group. UK bloggers were also more ready to make links to overseas blogs in their blogrolls, while US bloggers as a group were less ready. More US bloggers also had blogrolls which contained only local links. Obviously, a great part of the explanation for this is the size of the US blogosphere compared to the rest of the world. It will be interesting to see if this US-centric approach changes in the future as the blogosphere continues to expand.

US bloggers were more likely to see blogging as a useful activity, attracting readers for the intellectual or creative work. However, an equal number of bloggers in both countries identified financial gain as a motivation for blogging. This was particularly true of female bloggers and can probably be linked to the higher number of women bloggers who worked part time. Blogging is now being seen as a viable income generator for those who need a flexible approach to employment.

Overall, the project suggests that further research needs to be undertaken into the blogosphere outside the US. Since blogging started in the US, the majority of research into blogging so far has focused on the US and it is suggested that this focus has resulted in all bloggers being defined through the US experience. The findings of this project suggest that bloggers outside the US may have different approaches to blogging and find different satisfactions. It also suggests a new financial motivation for blogging, which had not previously been identified, and which may be an indication of the way in which the blogosphere is evolving.

## References

[1] BLOOD, R., Weblogs: a history and perspective. *Rebecca's Pocket*, 7 September 2000**.** http://www.rebeccablood.net/essays/weblog_history.html (accessed November 2006).

[2] SCHIANO, D., et al, Blogging by the rest of us. Conference on Human Factors in Computing Systems, 24-29 April 2004, Vienna. Published in CHI '04 extended abstracts on Human factors in computing systems, ACM Press, New York. 1143-1146.

[3] HERRING, S. C.; SCHEIDT, L. A.; BONUS, S.; WRIGHT, E., Bridging the gap: A genre analysis of weblogs. Proceedings of the 37th Hawaii International Conference on System Sciences, 5-8 January 2004, Big Island, Hawaii. Los Alamitos: IEEE Press.

[4] KRISHNAMURTHY, S., The multidimensionality of blog conversations: The virtual enactment of September 11. AOIR Internet Research 3.0: Net/Work/Theory. Maastricht. October 13-16 2002.

[5] SINGER, J. B., The Political J-Blogger: Normalising a new media form to fit old norms and practices, *Journalism*, 6(2), 2005, 173-198; Matheson, D., Weblogs and the epistemology of the news: some trends in online journalism, *New Media and Society*, 6(4), 2004, 443-468; Kahn, R. and Kellner, D., New media and internet activism: from the 'Battle of Seattle' to blogging, *New Media & Society* 6(1), 2004, 87-95; Wall, M., Blogs of War, *Journalism*, 6(2), 2005, 153-172; Pedersen, S. and Chivers, A., 'Empowering citizens to join the debate? What draws readers to news blogs?', submitted to *The International Journal of Technology, Knowledge and Society*, 2007.

[6] HERRING, S. C. et al, Conversations in the blogosphere: an analysis "from the bottom up". *Proceedings of the thirty-eighth Hawai'i International Conference on System Sciences (HICSS-38)*, 2005, Los Alamitos: IEEE Press; Schiano, D. J., Nardi, B. A., Gumbrecht, M. and Swartz, L., Blogging by the rest of us. *CHI 2004, April 24-29 2004, Vienna, Austria*, http://home.comcast.net/~diane.schiano/CHI04.Blog.pdf (accessed on 3rd June 2004); Nardi, B. A., Schiano, D. J. and Gumbrecht, M., Blogging as a social activity, or, would you let 900 million people read your diary? *Proceedings of computer supported cooperative work 2004*, http://home.comcast.net/%7Ediane.schiano/CSCW04.Blog.pdf (accessed 23 February 2006).

[7] HUFFAKER, D.; CALVERT, S., Gender, identity, and language use in teenage blogs, *Journal of Computer-Mediated Communication*. 10.2, 2005.

[8] MORTENSEN, T.; WALKER, J., Blogging Thoughts: Personal Publication as an Online Research Tool. In A. Morrison (Ed.), Researching ICTs in Context. InterMedia, University of Oslo, 2002, 249-279.

[9] THELWALL, M., Bloggers during the London attacks: Top information sources and topics. 15th International World Wide Web Conference, Edinburgh, Scotland, May 23-26, 2006, 2.

[10] TRAMMELL, K. D.; TARKOWSKI, A.; HOFMOKL, J.; SAPP, A. M., Rzeczpospolita blogów [Republic of blog]: Examining Polish bloggers through content analysis, *Journal of Computer-Mediated Communication, 11(3),* 2006; Abold, Roland, 1000 Little Election Campaigns: Utilisation and Acceptance of Weblogs in the Run-up to the German General Election 2005, 2006 ECPR Joint Sessions of Workshops, 25-30 April, 2006. Nicosia/Cyprus; Tricas-García, F., and Merelo-Guervos, J. J., The Spanish-Speaking Blogosphere: Towards the Powerlaw? IADIS International Conference WWW/Web Based Communities, Lisbon. 24-26 March, 2004.

[11] PEDERSEN, S.; MACAFEE, C.,The Practices and Popularity of British Bloggers. ELPUB2006. Digital Spectrum: Integrating Technology and Culture - Proceedings of the 10th International Conference on Electronic Publishing held in Bansko, Bulgaria 14-16 June 2006 / Edited by: Bob Martens, Milena Dobreva. pp. 155-164 http://elpub.scix.net/cgi-bin/works/Show?213_elpub2006; Auty, Caroline, UK elected representatives and their weblogs: first impressions. Aslib Proceedings: New Information Perspectives, 57.4 (August 2005). 338-355; Thelwall, M., Bloggers during the London Attacks.

[12] WRAY, R, Surfers glued to web for 11 hours a week. *The Guardian*, 29 November 2006. http://technology.guardian.co.uk/news/story/0,,1959548,00.html (accessed 29 November 2006).

[13] PEDERSEN, S., Women users' motivation for establishing and interacting with blogs (web logs). *International Journal of the Book 3(2),* 2005, 85-90.

[14] RILEY, D., Blog count for July: 70 million blogs. The Blog Herald. b5media, 2005. http://www.blogherald.com/2005/07/19/blog-count-for-july-70-million-blogs/.

[15] POLLARD, D., Is the blogosphere sexist? How to save the world, 30 October 2003. http://blogs.salon.com/0002007/2003/10/30.html (accessed 9 March 2006); Ratliff, C., Whose voices get heard? Gender politics in the blogosphere. Culture Cat, 25 March 2004. http://culturecat.net/node/303 (accessed 12 September 2005); Ratliff, ., *The* link portal on gender in the blogosphere. Culture Cat, 21 December 2004. http://culturecat.net/node/637 (accessed 12 September 2005); Garfunkel, J., Promoting women bloggers: a timeline of relevant discussions. Civilities media structures research, 15 March 2005. http://civilities.net/PromotingWomenBloggersTimeline (accessed 23 February 2006).

[16] RATLIFF, C., WATW by the numbers. Culture Cat, 18 February 2006. http://culturecat.net/node/1030.

[17] HENNING, J., The blogging iceberg. Perseus, 4 October 2003. http://www.perseus.com/blogsurvey/iceberg.html (accessed 5 October 2005).

[18] HERRING, S. C.; KOUPER, I.; SCHEIDT, L. A.; WRIGHT, E. L., Women and children last: the discourse construction of weblogs. *In*: Laura J. Gurak et al., eds. *Into the blogosphere: rhetoric, community, and culture of weblogs*. Minneapolis: University of Minnesota, 2004. http://blog.lib.umn.edu/blogosphere/women_and_children.html (accessed 7 March 2006).

# Enhancing Traditional Media Services Utilising Lessons Learnt from Successful Social Media Applications – Case Studies and Framework

*Asta Bäck ; Sari Vainikainen*

VTT, Media and Internet, P.O.Box 1000, FI-02044 VTT, Finland
e-mail: asta.back@vtt.fi; sari.vainikainen@vtt.fi

## Abstract

The paper presents a framework for describing electronic media services. The framework was created by utilising earlier models and case studies of successful social media applications. Wikipedia, YouTube and MySpace were analysed because they are among the most popular sites in the world and they highlight different aspects of social media applications. The proposed model consists of two main parts: Concept and system, and Content and user. Both of them were further divided into four subgroups. With the help of a radar view, various applications can be described and compared and their further development opportunities identified. A prototype application, StorySlotMachine, is used as a case example, where the framework is used.

**Keywords:** social media; YouTube; Wikipedia; MySpace

## 1 Introduction

The initial vision by Tim Bernes-Lee [1] was that internet would be a platform for interactive information sharing. During the recent years we have seen development that has made this vision true. The tools needed for digital content production have become easy to use and cheap enough for a large number of people. More and more people have broadband Internet access, which makes Internet the natural way to share digital media. Terms like Web 2.0, social media and user-generated content are being used to describe the services that have been created on top of this development during the last years.

This development has raised many questions among media companies. The traditional media approach has been product centred. The aim is to offer a product to as large audiences as possible. The role of a media company has been to organise and filter content and to package it into a marketable whole. Marketing is often done at two directions: the product is sold to readers or viewers and this audience is then sold to advertisers.

Media services can be roughly categorised into two main groups: information or entertainment. In practice, most traditional publications try to address both of these with the main emphasis on one or the other. We can also claim that many publications promote some sense of community among their readers or viewers by giving people common topics to talk about. Community building is also obvious in publications with relation to some religious, political or some other idealistic movement. But also in these cases, the role of individual end users has mostly been invisible and they have been seen as a target group.

Social media applications give people new ways to find information and entertainment and also to build communities, which has meant that the role of traditional media has become weaker. Traditional media companies do not host the most popular or most quickly growing sites on the net.

Several terms are being used in connection to these new services. The word participatory media has been used to emphasise the nature of user participation in media creation. Other frequently used terms are social media, social software and social networking. These terms bring out the social aspect – users not only act as content creators but social interaction between users has become possible and visible in the applications.

From the content or media object point of view, the most important change is the lengthened life cycle of media objects. In the traditional publishing models, the content selection at media object level was made by professionals who compiled the aggregations that then were offered to consumers as packages. As content has become digital and it has become easy to refer to and discuss single media objects, the life of a media object may be lengthened considerably (Fig 1). Media is also more and more consumed as smaller fragments - video clips instead of whole shows, single songs instead of whole albums.

This development raises the question of what traditional media companies could learn from the successful social media applications to revitalise their own business. The aim was to create a framework that would help in characterising social media applications and would be usable as a tool in finding new opportunities to developing new services to traditional media companies.

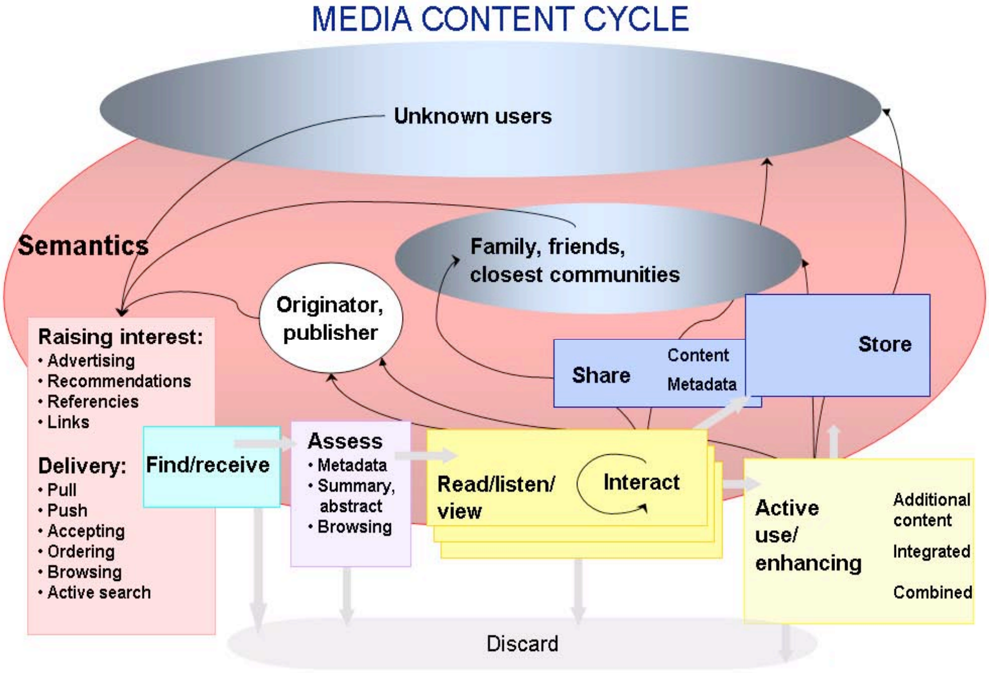


**Figure 1: The life content cycle is prolonged for the electronic media**

## 2 Methodology

The chosen research method was to make case studies of successful social media applications. In order to make the case studies in a systematic way, the main characterising features of social media applications were identified. Here, also the traditional media processes were taken into consideration in order to find and pinpoint the areas were the differences are greatest.

Case studies were made by going through three hugely popular applications, Wikipedia, YouTube and MySpace, and analysing their features. These three are among the most popular sites in the world, and they highlight different aspects of social media applications. Additional information and research findings were searched from the literature. No user studies were made, so the analysis is based on what can be seen at the website and what research results and other information is available.

## 3 Results

### 3.1 Characterisation of Social Media Applications

We can look for characteristics for describing social media applications from two directions: from general IT application adoption point of view and in comparison to traditional media. Technology acceptance model (TAM) has been applied successfully as a theoretical framework to explain the adoption of IT applications [2]. It was originally developed relating to office applications, but is has been found to apply to other IT applications as well.

The model is simple. It explains the acceptance with the help of two factors – ease of use and usefulness. People must perceive benefits from using an IT application and the application must be easy to use. These are important criteria for social media applications, since the user must be able to learn to use the service by him or her self, and the value of these applications increases as the number of users increases. The first seconds a first time visitor spends with the application are the most important ones, because if the user is confused and does not know right away how to use the service and what benefit there are available, he or she may never return. The traditional media process consists at top level of three main steps:

1. Concept and longer term product development: what kind of content products will be offered and to which target groups;
2. Marketing to customers and advertisers;
3. Production – creating and selecting the content for individual issues. Here, depending on the type of media, content is either made based on commissions or offered for publishing;
4. Feedback from customers is received two ways: direct feedback on single articles but mostly in the form of increasing or decreasing sales and subscriptions.

Since social media applications are IT applications, we need to pay attention to the ease of use aspect, which has not been relevant in traditional media products. The usefulness or value proposition is important in any service or product, since other features are irrelevant if there is not perceived value in using an application. In social media, the aspects are particularly worth addressing users and content. Based on these features, the following characteristics were chosen to be used in the analysis:

1. What is the main service concept, and which additional needs people may have in connection to media products (Concept and value proposition);
2. What kind of content is being created and shared and how users participate in content creation and management, what is required from users to be able to participate in content creation and management (Content and user participation);
3. How visible the users are in the service, does the service support identity and community creation (User identity and networking);
4. When and how the service is being used and marketed (Use and marketing).

## 3.2 Wikipedia

Wikipedia[1] has become an important collection of knowledge and it can be regarded as the open content counterpart to the open source development.

***Concept, value proposition***

The Wikipedia product concept – a collaboratively created encyclopaedia, was defined by its founders. The key promise was to make a free and open knowledge source and this way participating in Wikipedia entailed participating in a big common goods endeavour. The five pillars [3] depicting the key Wikipedia principles are as follows:

- Wikipedia is an encyclopaedia
- Wikipedia has a neutral point of view
- Wikipedia is free content
- Wikipedia has a code of conduct
- Wikipedia does not have firm rules

Larry Sanger, one of the Wikipedia founders, sees the following causes behind the Wikipedia success [4]:

1. Open content licence
2. Focus on encyclopaedia
3. Openness – anyone can contribute
4. Ease of editing
5. Collaborate radically, anybody can edit anybody's article
6. Offer unedited, unapproved content for further development
7. Neutrality
8. Start with a core of good people
9. Enjoy the Google effect

[1] http://en.wikipedia.org

***Content and user participation in content creation***
There are two main parts in the Wikipedia:

- the platform with many features supporting collaborative and unmanaged content creation;
- the content creation and management process that lets any user act as a content creator or editor.

The Wikipedia platform supports collaborative editing by storing all versions of the content together with the user name or IP address. All versions of the article are available and comparisons may be made between any of them. If one of the earlier versions is considered better than the current one, it is possible to revert to it. This can also be used to fight spam.

Anyone is able to create a new entry on any subject they see worth writing about. Existing articles can be used as examples. Most active users act as editors and, for example, mark articles that need further refinement. New articles are being created by making an internal Wikipedia reference to a topic that does not yet exist. This means that the person who creates or names a page, need not know or write about it but only to be of the opinion that this is a topic that should be included in the Wikipedia. There is no process for selecting which articles to write. However, users may create what is called a portal within Wikipedia to promote and support creating articles relating to a topic. Such portals have been created for multiple topics, for example for various sports or countries.

In order to promote high-quality articles, articles may be requested a peer-review, and an article may get evaluated as a good article or a featured article. The focal point of the users' work is creating a coherent and balanced article about a topic. Self-organising is the supported way to get things done. Active users may become selected into managerial tasks within the Wikipedia community.

***User networking***
Mediawiki[2], the platform used for creating Wikipedia, supports creating pages or articles. Information about single users may be presented in the same way: A user may write an article about him- or herself and give whatever information he or she wants to share with others - or remain unknown and participate anonymously. There is no direct support to find information about user connections, but users must explicitly create any such information.

***Use and marketing***
Users may access articles either by making a search, clicking the Random page link or following the links in the articles. Users are encouraged to embed internal Wikipedia links within the articles, and group the external links at the end of the article, if such links exist.

There is not direct information available about how much the pages have been viewed. There is a Statistics page that gives some information about the total number of pages, as well as lists the most popular pages. There is no direct support to invite other people to visit Wikipedia.

Currently search engines are an important driver to Wikipedia articles. Wikipedia articles are ranked highly, and some search engines support targeting searches directly to Wikipedia. Also, the free access to Wikipedia content has given it visibility, as various open initiatives, such as Semapedia[3], have utilised the Wikipedia content. Free content brings with it the opportunity to get free visibility.

## 3.3 YouTube

***Concept, value proposition***
YouTube[4] is technically a platform for sharing online videos. The service was created to support this particular media format at the point when people had started creating more videos, but there was no easy way to share them.

We can say that the value proposition has two levels: the immediate and practical value and potential value. The immediate value is being able to share content over the internet with family and friends, and the potential value is in being able to reach new people and contacts, and even international fame.

[2] http://www.mediawiki.org/wiki/MediaWiki
[3] http://www.semapedia.org/
[4] http://www.youtube.com

***Content and user participation***

The service relies completely on users to upload video clips. The platform is open for any person who wants to upload a video. At uploading, users give some basic metadata about their video (category, description, tags).

User participation is not restricted to bringing in new content, but users may participate in assessing and evaluating content. There are several opportunities to that such as rating, discussing, and adding to favourites. Also the mere act of viewing a video is utilised as a metadata, because number of times each video is viewed is shown and the most viewed videos can easily be found. Users' may collect videos into playlists and these lists may be made public, which supports finding related items. A user may also create a channel that others may subscribe. A collaborative way of aggregation videos is to set up a group with a special theme.

***User networking***

Each user gets a home page that gathers information about that user and his or her activities in the community. Users may easily communicate with each other and also create a permanent link between themselves by connecting as a friend. User favourites, subscriptions and subscribers are visible there. User networks support both finding other users with similar interests and also finding content.

***Use and marketing***

The YouTube website shows about each video how many times it has been viewed. Also information is given on which websites have links to each video, as well as honours, if the video has received such. Once a video has been viewed the system suggests additional videos for viewing in order to make people stay longer at the website. User networks and the ability to embed any YouTube video in other websites have played a key role in making YouTube known and popular. YouTube fits well to the online communication where users post links to each other.

Jawed Karim, one of the YouTube founders considers the ease of linking to the videos as well as the opportunity to embed a YouTube video by copying a piece of HTML code to ones web page as factors contributing to the rapid growth that the site experienced [5]. Additional key feature was the ease of use: videos could be viewed immediately without downloading a viewer or codec and discussions could be made around videos, so that viewing a video was not any longer a single, detached event, but became a social event with possible many steps.

## 3.4 MySpace

***Concept, value proposition***

MySpace[5] is a social networking site. It main value proposition is to offer a public web space to present oneself and to connect to other people with similar interests. Also here, the opportunity to become famous, particularly in the music scene, is an important part of the value proposition.

***Content and user participation***

The content in MySpace is gathered around profile pages. People add media that they find interesting and supportive to their profile and personality. Also various commercial products like cars or films may have a profile page that people may connect to, if they find the product valuable to them. Users may, for example, rate videos, but these ratings are not published as directly and openly as for example in YouTube.

***User*** marketing. The break through however was made when small offline communities between 100 and 1000 members were attracted to the users of the site. This contributed to launching the network effect.

## 3.5 Lessons from the Social Media Applications

***networking***

User networking is the key functionality in the service.

***Use and marketing***

According to danah boyd [6] there are several reasons why MySpace became so popular, particularly among 14-24 people. This age group has a bigger need for creating a public profile than other age groups. Youth also needs their own public space, and MySpace has offered this in virtual space.

[5] http://www.myspace.com

Also, the communication opportunities offered by MySpace fit the existing communication patterns: the youth communicates via instant messaging for immediate communication needs, and MySpace is the complementing asynchronous communication channel. MySpace has become part of their daily routine, which means that when people have their computer on, one of the sites they have open, is MySpace.

Gabbay in his case study of MySpace [7] claims as one of the decisive factors contributing to the MySpace success the freedom that the platform and its developers offered to the users. Photos and music were according to him the most important content elements. As to the initial marketing of the site, various kinds of marketing actions were utilised including contests and email
Each of these three successful applications has a clear focus or value proposition, which is complemented with additional values.

Wikipedia is the most idealistic application where a common goal to create an important knowledge mass motivates people. Here a single user may remain anonymous, if he or she wishes so. And also, there is least support for making and showing user connections. Wikipedia differs from the other two also in that a common media object, an encyclopaedia article is being created in collaborative fashion. The model is disruptive in the sense that users are encouraged to modify and alter other users' texts. Also, the self-organising way of creating content can be regarded as disruptive. The critical point in Wikipedia is user motivation and maintaining it. There are little opportunities to external rewards for Wikipedia content creators and editors.

YouTube offers both immediate value (video sharing) and potential value as a platform in seeking feedback, popularity and fame. User information and profile are visible and even though the content is shared freely among users, each user may utilise the platform for making him or herself more known. There are several ways for users to connect with each other. Users may also create groups that concentrate on some special interest. The role of users in organising the content is also important: the video related metadata becomes richer as people rate and comment videos and add them to their favourites. Finding videos would be practically impossible without this additional user generated metadata.

MySpace turns the relation between users and media objects into reversed order. Here, the users or user profiles are in the focus of attention and media objects are used to complement the user profiles. MySpace lets users play with their creativity and use the system in many, also in unforeseen ways.

In traditional media, products like an issue of a magazine or book, are being created by utilising well-defined processes with input and output. Web-based social media applications have a very different life cycle: a platform is offered to users and as the service gets used, its starts to take shape and may get new form.

## 3.6 Framework

Based on the initial characteristics used in describing the social media applications and the findings that were made, the characteristics were further elaborated into a framework. The framework supports getting a quick overview of the features a service and pinpoints to areas where there could be additional opportunities for further development by taking into account the features that have been successfully utilised in social media applications. The features were grouped into two groups: Concept and system, and Content and user. There were further divided into four subgroups each.

*Concept and system* (Fig 2) is divided into the following subgroups:

1. Main attention in the service
   - Content
   - Users
   - Other

2. Value proposition timeframe
   - Immediate
   - Long term, cumulative

3. Value proposition type
   - Social
   - Emotional
   - Rational, practical

Social and emotional values are indicated as separate, even though they are somewhat overlapping. We wanted, however, to give the opportunity to separate them. For example, a movie typically offers emotional value without strong social aspect.

4. Usability and improvements
    - Ease of use
    - Evolutionary development

*Content and use* (Fig 2) is divided into four subgroups:

1. Content enhancements
    - Aggregating content (e.g. playlists)
    - Modifying content
    - Opinion expression (ratings, metadata) and its visibility

Aggregating content refers to the opportunities of combining available content into aggregations or combinations that a user finds meaningful or practical to him- or herself, for example making a playlist of videos or combining bookmarks into a readlist. Modifying content refers to changing the content or features within a media object, for example editing the text or picking a part of a video and combining it with some other video. Opinion expression refers to user generated metadata, which may be utilised to make recommendations, and/or shown explicitly to other users.

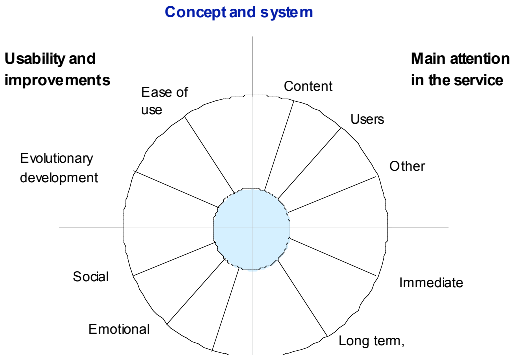


**Figure 2: The framework for describing a media service features. The more a certain feature or functionality is supported a position closer to the outer circle should be chosen**

2. Content type and sharability
    - User-generated
    - Commercial, professional
    - Exportability (embedding, APIs)

3. Creation opportunities
    - Alone
    - Small group
    - Community

Here, the term ‘Small group’ refers to the opportunity of organising groups with special focus or aim, or with one’s friends and family. The ‘Community’ refers to larger scale collaborative work and sense of community within the service.

4. User visibility and networking
    - Identity building
    - User networking
    - Inviting and attracting new users

By going through the whole framework, it is possible to get an overview as to where the main opportunities are for further development. Also more detailed analysis may be carried out by picking two of the subgroups and looking at their interrelations as a quadrangle. For example, the value proposition time frame and type could be analysed as a quadrangle, or Content enhancement opportunities and Content type and sharability, or Content enhancement opportunities and Creation opportunities. The framework does not imply some ideal value, even though the general interpretation is that the larger a figure comes out when evaluating the features of a service, the more opportunities there are for user interaction and value creation. It may not, however, be sensible or even possible to combine all the features in one service.

Figure 3 shows two of the analysed services described with the help of the framework. We can see that YouTube has more support on the user and user connectivity side. The clear difference in the approach to content enhancement is that YouTube supports aggregating and playing with video clips where as the main focus of Wikipedia in the modifying and working on the actual content, mainly the articles. Regarding concept and use, we can see that YouTube includes many more of the identified features than Wikipedia.

In Figure 4, an application prototype, called StorySlotMachine, is described with the help of the framework. StorySlotMachine [8] was built in one of our earlier research projects. The project aimed at exploring the opportunities of utilising semantic metadata and user-created content together with commercial media content. The application lets users explore and play with media content. Content may come from various sources: user's own content may be complemented with that from other users and what a commercial media company offers. The users are encouraged to add some information about where their photos were taken, and what is shown in the photos. Based on these clues, additional related content is offered. Users may also create their own collections out of the existing material, either travelling plans, guides or reports.

The prototype deals with travelling related content and focus. Many media companies have extensive archives that are not effectively utilised as end user services. The StorySlotMachine is an example of how their content could be offered in a more interesting ways than as mere searches and search result lists.

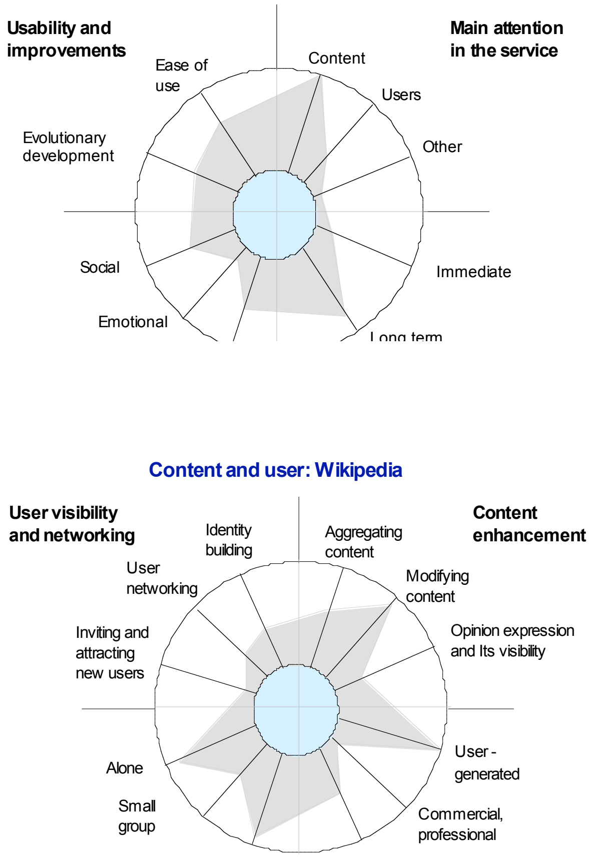


**Figure 3a-b: Two social media applications, Wikipedia and YouTube characteristics depicted in the proposed framework. The differences in the emphasis of the applications come clearly out**

The prototype development was started three years ago. The idea of combining user-generated content with professionally created commercial content was emphasised as well as the aspect of being able to easily play with content components. The social aspects of such an application were not taken so strongly into the focus in system development. In user tests, users liked the idea to get information as ready made stories and to combine their own content with commercial content, and were interested in features that let them view and utilise aggregations that other users had created. We clearly see the opportunities in further development by adding more visibility as to what other people are doing in the service, and also being able to connect to other users.

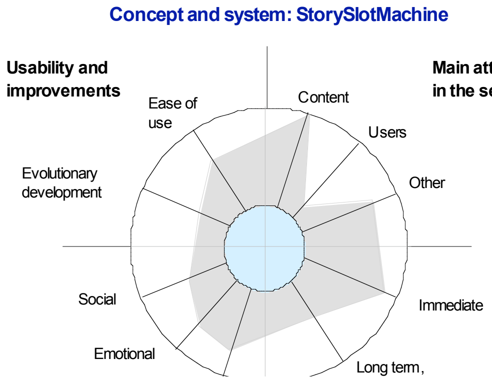


**Figure 4: A prototype application, StorySlotMachine described with the framework. We can see that the development opportunities exist particularly in adding support for user visibility and social aspects of content creation and sharing**

## 4 Discussion

The framework gives a quick way of capturing the features of a media service and comparing them with the op– portunities and the lessons learned from successful social media applications. The framework does not address the business side. Currently advertising is the key business model, and people rarely pay for the access to content [9].

Traditional media companies may benefit from the web experiments that other companies have made, but there are also many challenges. The most successful applications have originated from new companies, and not from traditional media companies. New entrants have not had any existing business and this had given them the opportunity to adopt very new and even disruptive models. During the first Internet wave in late 1990'ies, the belief was that the key was to attract as many users as possible, and that the business would come along with it. This belief is still valid in many cases. Also, creating new services on the web for new customers is in many cases easier and more sensible than trying to offer web-based services to users who are more accustomed and happy with their traditional media products.

As to the framework, we see that it should be tested further. More services should be described with it in order to validate it more. Also comparisons should be made with not so successful applications – does this framework help in differentiating these from the more successful ones. Also, wider application of the model as a tool in exploring the development opportunities for existing media services should be done.

The focus is utilising the framework should not be in debating where the exact position on each axis is but to discuss, whether these different aspects should be supported in a service, and if so, what would be the best way to do it.

## 5 Conclusions

Social media applications have brought the users into focus in media applications. The traditional model where users were regarded as a target group remains also valid but the growth opportunities have been in activating and giving tools to users. New opportunities have also been found in taking something else than content as a starting point: in MySpace the users are the centre piece, and in our StorySlotMachine the travelling sights.

The framework gives a good starting point for exploring the development opportunities based on success stories. But, we must pay attention to new opportunities that may emerge but which have not been explored yet. Adding the mobile dimension and utilising context can be mentioned as two areas where to look for new opportunities.

Involving users creates the opportunity to grow the sense of community and ownership of the service. The most critical point is the start – how to get the community to attract and active users. Also, connectivity to other users and content on the net is needed and helps in creating successful applications.

## Notes and References

[1] BERNERS-LEE, T., Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web. Harper San Francisco; 1st edition (September 22, 1999).

[2] DAVIS, F. D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quartely:13/1989, pp. 319–339.

[3] Wikipedia: Five pillars. http://en.wikipedia.org/wiki/Wikipedia:Five_pillars. Accessed April 10, 2007.

[4] ANON. The Early History of Nupedia and Wikipedia, Part II. http://features.slashdot.org/article.pl?sid=05/04/19/1746205&tid=95. Accessed April 10, 2007.

[5] KARIM, J., YouTube: From Concept to Hyper-growth. http://www.youtube.com/watch?v=nssfmTo7SZg. Accessed April 10, 2007.

[6] BOYD, D. "Identity Production in a Networked Culture: Why Youth Heart MySpace." *American Association for the Advancement of Science*, St. Louis, MO. February 19. 2006. Available at http://www.danah.org/papers/AAAS2006.html. Accessed April 10, 2007.

[7] GABBAY, N. MySpace Case Study: Not a purely viral start. September 2006. http://www.startup-review.com/blog/myspace-case-study-not-a-purely-viral-start.php. Accessed April 10, 2007.

[8] BÄCK, A.; Vainikainen, S.; Näkki, P.; Reti, T.; Sarvas, R.; Seppälä, L.; Hietanen, H.; Turpeinen, M. Semantically supported media services with user participation. Report on the RISE-project. VTT, Espoo. VTT Publications: 612. 2006.

[9] KANGAS, P.; TOIVONEN, S.; BÄCK, A.; (ed.). Google advertisements and other social media business models (in Finnish, Googlen mainokset ja muita sosiaalisen median liiketoimintamalleja) VTT, Espoo. VTT Tiedotteita - Research Notes: 2369. 2007.

# The Fight against Spam - A Machine Learning Approach

*Karel Jezek; Jiri Hynek*

Department of Computer Science & Engineering, Faculty of Applied Sciences, University of West Bohemia
Univerzitní 22, 306 14 Pilsen, Czech Republic
e-mail: {jezek_ka, jhynek}@kiv.zcu.cz

## Abstract

The paper presents a brief survey of the fight between spammers and antispam software developers, and also describes new approaches to spam filtering. In the first two sections we present a survey of the currently existing spam types. Some well-mapped spammer tricks are also described, although the imagination of spam distributors is endless, and therefore only the most common tricks are covered. We present some up-to-date spam blocking techniques currently integrated into today's spam filters. In the Methodology and Results sections we describe our implementation of Itemsets-based, Naïve Bayes and LSI classifiers for classifying email messages into spam and non-spam (ham) categories.

**Keywords:** unsolicited mail; spam filter; machine learning; latent semantic indexing; classification

## 1 Introduction

The term "electronic publishing" commonly refers to the distribution of e-books and periodicals, as well as websites, blogs, etc. E-mail is just another means of information dissemination. It thereby demonstrates the features of electronic publishing. If used properly, it perfectly serves for information exchange among individuals, but when used maliciously (which is more often the case), it serves for broadcasting of (mis)information to the general public. What we have in mind is, of course, spam, also known as Unsolicited Bulk Mail (UBM), Excessive Multi-Posting (EMP), Unsolicited Commercial Email (UCE), spam mail, junk mail, or bulk email, as opposed to the term "ham" used for legitimate mail.

The very first spam was distributed in 1978 via ARPANET, notifying all network users of the newly developed DEC-20 computer. The first unsolicited mail that was actually labeled "spam" for the first time in history was distributed in 1993 by Richard Depew, who mistakenly distributed 200 messages to newsgroups administered by him. He apologized immediately, using the word "spam".

The antispam industry is constantly developing new techniques to fight sophisticated tricks used by spammers. On January 24, 2004, Microsoft chairman Bill Gates presumptuously announced that "spam will be solved by 2006". However, neither Microsoft, nor any other company, has yet found a solution. The spam-filter-review statistics of 2006 (see http://spam-filter-review.toptenreviews.com/spam-statistics.html) show the following data: spam constitutes 40% of all email messages, there are 12.4 billion spam emails distributed per day, and 2,200 spam messages are received per person per year. The most common categories of spam are product advertisements (25%), financial (20%), adult (19%), scams (9%), and health (7%).

Spam messages pose a serious problem due to multi-billion dollar costs. The MSSP Survey of 2006 claims that unsolicited emails now consume approx. 819 terabytes of bandwidth every day, representing 85% of global mail traffic. Fortunately for email users, antispam software has become increasingly effective in recent years

Many computer users have been given hope by antispam laws, such as the US federal Can-Spam Act [1] of 2003, but only 26 states had implemented any antispam legislature by 2006. This required spam senders to allow recipients to opt out of receiving future messages. It also prescribed imprisonment for violators. According to the US Federal Trade Commission, the volume of spam declined in the first eight months of 2005, but the decline was short-lived. At the beginning of 2006, spam was again out of control.

An example of another useful initiative is the OECD's "Task Force on Spam" [2]. The OECD (Organization for Economic Co-operation and Development) has launched its Anti-Spam "Toolkit" as the first step in a broader initiative to help policy makers, regulators and industry players orient their policies relating to spam solutions and restore trust in the Internet and email.

# 2 State of the Art

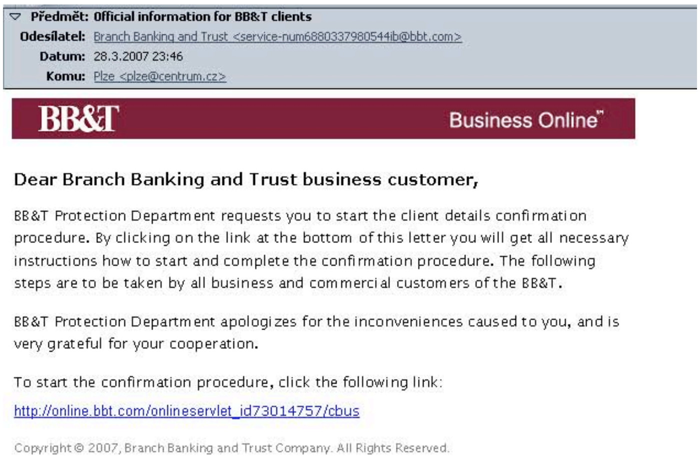
## 2.1 A Survey of Currently Existing Spam Types

### 2.1.1 Stock Spam, „Pump and Dump“

The term “pump and dump” on the Internet represents unsolicited mail offers of very inexpensive goods (typically below $1), urging mail recipients to quick purchase. This evokes massive demand for goods which have already been sold in most cases. Nonetheless, the price of the goods is gradually increased (“pumped”). This type of unsolicited mail often includes links to small or non-existing companies, as it is almost impossible to track any information on the company making the attractive deal. In some cases, “pump and dump” spam is designed to hurt the good name of an existing company, as the consequences of illegal business deals are borne by the actual company, not the spammers.

### 2.1.2 Phishing

Phishing (see Figure 1) is used for messages designed to elicit personal data (such as bank account numbers, credit card numbers, passwords, etc.) from email recipients. The term is derived from “fishing”, which is exactly what spammers do – distribute “bait” and wait to see what happens. Spammers commonly use exploits such as using the company’s image, inserting links to the real company site, or using email that appears to be from the spoofed company.



▽ **Předmět:** Official information for BB&T clients
**Odesílatel:** Branch Banking and Trust <service-num6880337980544ib@bbt.com>
**Datum:** 28.3.2007 23:46
**Komu:** Plze <plze@centrum.cz>

Dear Branch Banking and Trust business customer,

BB&T Protection Department requests you to start the client details confirmation procedure. By clicking on the link at the bottom of this letter you will get all necessary instructions how to start and complete the confirmation procedure. The following steps are to be taken by all business and commercial customers of the BB&T.

BB&T Protection Department apologizes for the inconveniences caused to you, and is very grateful for your cooperation.

To start the confirmation procedure, click the following link:

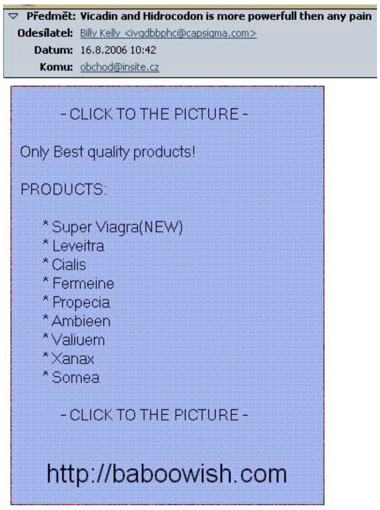
http://online.bbt.com/onlineservlet_id73014757/cbus

Copyright © 2007, Branch Banking and Trust Company. All Rights Reserved.

**Figure 1: Example of a phishing spam**

### 2.1.3 Image-Based Spam

Tricks used to distribute unsolicited mail get more and more sophisticated. The best way to get around statistical text filters is to use images instead of text (see Figure 2). Image handling is quite difficult for antispam software, regardless of the actual image form – plain text converted into an image, various interference items on the background, use of animations, etc. Although use of images for spamming is not a new concept, it is definitely gaining popularity. According to various studies, approximately one-third of all unsolicited mail was represented by image-based spam at the end of 2006. It seems that spammers are quite content with the hit rate of their messages, and keep converting all their text-based mails into images.



▽ **Předmět: Vicadin and Hidrocodon is more powerfull then any pain**
**Odesílatel:** Billy Kelly <ivgdbbphc@capsigma.com>
**Datum:** 16.8.2006 10:42
**Komu:** obchod@insite.cz

- CLICK TO THE PICTURE -

Only Best quality products!

PRODUCTS:

* Super Viagra(NEW)
* Leveitra
* Cialis
* Fermeine
* Propecia
* Ambieen
* Valiuem
* Xanax
* Somea

- CLICK TO THE PICTURE -

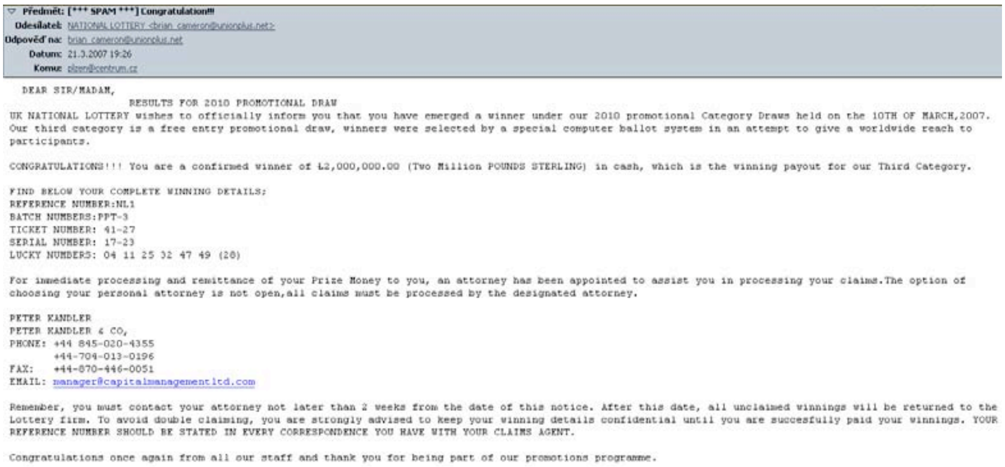
http://baboowish.com

**Figure 2: Example of a clickable image-based spam**

### 2.1.4 Text Spam

Text spam is just unsolicited commercial mail distributed in textual form (see Figure 3). Typical features of the text spam are listed below (please note that the majority of these features are language-independent):

- HTML text contained in message body,
- High proportion of capital letters (usually more than 30%),
- Exclamation mark(s) in the message subject,
- Instructions on how to unregister from the distribution list,
- Instruction to click on a link,
- Text lines longer than 200 characters,
- High priority assigned to the message,
- Nonsense date of sending (such as 1st January 1970),
- Disclosed message sender,
- More (or disclosed) message recipients.



▽ **Předmět: [*** SPAM ***] Congratulation!!!**
**Odesílatel:** NATIONAL LOTTERY <brian_cameron@unionplus.net>
**Odpověď na:** brian_cameron@unionplus.net
**Datum:** 21.3.2007 19:26
**Komu:** plzen@centrum.cz

```
  DEAR SIR/MADAM,
                  RESULTS FOR 2010 PROMOTIONAL DRAW
UK NATIONAL LOTTERY wishes to officially inform you that you have emerged a winner under our 2010 promotional Category Draws held on the 10TH OF MARCH,2007.
Our third category is a free entry promotional draw, winners were selected by a special computer ballot system in an attempt to give a worldwide reach to
participants.

CONGRATULATIONS!!! You are a confirmed winner of £2,000,000.00 (Two Million POUNDS STERLING) in cash, which is the winning payout for our Third Category.

FIND BELOW YOUR COMPLETE WINNING DETAILS;
REFERENCE NUMBER:NL1
BATCH NUMBERS:PPT-3
TICKET NUMBER: 41-27
SERIAL NUMBER: 17-23
LUCKY NUMBERS: 04 11 25 32 47 49 (28)

For immediate processing and remittance of your Prize Money to you, an attorney has been appointed to assist you in processing your claims.The option of
choosing your personal attorney is not open,all claims must be processed by the designated attorney.

PETER KANDLER
PETER KANDLER & CO,
PHONE: +44 845-020-4355
       +44-704-013-0196
FAX:   +44-870-446-0051
EMAIL: manager@capitalmanagementltd.com

Remember, you must contact your attorney not later than 2 weeks from the date of this notice. After this date, all unclaimed winnings will be returned to the
Lottery firm. To avoid double claiming, you are strongly advised to keep your winning details confidential until you are succesfully paid your winnings. YOUR
REFERENCE NUMBER SHOULD BE STATED IN EVERY CORRESPONDENCE YOU HAVE WITH YOUR CLAIMS AGENT.

Congratulations once again from all our staff and thank you for being part of our promotions programme.
```

**Figure 3: Example of a text spam**

## 2.2 Common Spammer Tricks

### 2.2.1 How to Get More Victims: Email Address Harvesting

Having enough email addresses to distribute spam to is the basic prerequisite for the success of any "advertising campaign" on the Internet. Spammers must therefore adopt various high-tech tricks to identify as many email recipients as possible, often looking for publicly available emails posted on the web. According to the CAN-SPAM Act [1], advertisers are prohibited from "harvesting" email addresses from web sites in the first place. However, such activities are difficult to monitor or penalize.

An “Internet bot” is a standalone Internet application designed to perform predefined tasks. The largest group of bots is represented by web spiders, whose task is to collect information from web pages. “Positive spiders” present no problem, as they focus namely on page indexing for Internet search engines. On the other hand, “spambots” are designed to search pages and look for email addresses only. A set of robots, i.e. computers hosting the same bot, represents a “botnet”, i.e. a network of spambots, that can be utilized for coordinated attacks, namely for high-volume spam distribution. Spammers thus misuse the computers of other people on the Internet to commit their illegal activities. They are, of course, immune to blacklists. Networks consisting of thousands of computers are available on the Internet to be leased for distributing unsolicited mail.

### 2.2.2 Traditional Tricks Used by Spammers to Fool Spam Filters

Over time, spammers have adopted many more or less sophisticated tricks to fool spam filters, namely those that are based on statistical parameters of spam messages. Here are some examples:

- Avoidance of keywords (such as *stock*, *Viagra*, etc.),
- Frequent change in sender’s address,
- Message encoding (such as base64, commonly used for secure message transfer),
- Hashing (e.g. insertion of HTML tags into messages),
- Use of images instead of plain text (namely GIF, JPEG, and PNG).

### 2.2.3 New Spammer Tricks

In the following paragraphs you will find a sample survey of new tricks used by authors of “new generation” spam. Unfortunately, the list is far from exhaustive, as new approaches are constantly being developed to obfuscate spam filters.

**Character hashing in words**

Spammers use this trick to make typical spam keywords illegible for a filter, although they present no problem for a human brain. Should the user label such a message as spam manually, a few new keywords are added to the keyword database used by the antispam software, with no effect until re-training the filter.

Example of a message with character hashing:

```
I finlaly was able to lsoe the wieght I have been sturggling to lose for
years! And I couldn't bileeve how simple it was! Amizang pacth makes you
shed the ponuds! It's Guanarteed to work or your menoy back!
```

**HTML code interleaving**

HTML code is inserted into the middle of words. This presents no problem for email clients with HTML code support, as the message is kept in perfectly legible form. However, it is difficult for the filter to detect keywords split by HTML code. On the other hand, this HTML code interleaving trick is quickly losing popularity among spammers. Here is an example of an email encoded in HTML table:

```
<table cellSpacing=0 cellPadding=0 align=center border=0>
      <tr vAlign=bottom>
      <td rowSpan=2>Inc</td>
      <td rowSpan=2>reas</td>
      <td rowSpan=2>e&nbsp;S</td>
      <td rowSpan=2>exual&nbsp;Desi</td>
      ...
```

**Commercial attachments in the form of Microsoft Office documents**

This is a way to avoid contents analysis by spam filters altogether; the message is passed as long as it can stand an antivirus check. On the other hand, the user must be curious enough to open the attachment, which is rarely the case, as the message comes from an unknown sender, and usually contains neither body text nor subject line in order to pass the spam filter.

**Keyword masking by repeating characters**

Spammers try to obfuscate keywords by repeating some characters. The message remains legible for humans, but makes detection by statistical filters difficult.
Here is an example: `Buuuyyyy cheeeeaaap viaaagraaa.`

**Word obfuscation by replacing characters by punctuation marks, spaces or images**
Statistical spam filters typically look for certain keywords such as *Viagra*, *tablets*, or *watches*, so spammers have adopted techniques to obfuscate them by using spaces and various punctuation marks, while preserving the legibility of their messages for humans. However, heuristics (i.e. sophisticated lexical analysis) can be integrated into text-based spam filters to fight this technique.

Examples of word obfuscations:

```
\/laGr@
Need a{} Dpiloma?
sh1pp1ng //orldwide
S0ft T4bs
Ci@li$
repl1ca w4tches from r0lex
```

**Use of CSS styles for color setting and/or visibility of letters**
The widespread application of CSS styles for web page formatting gives spammers a new opportunity to use the same technique to format their messages and circumvent spam filters based on statistical parameters.
Example – Insertion of CSS styles into HTML tags to "encode" the word *Cialis*:

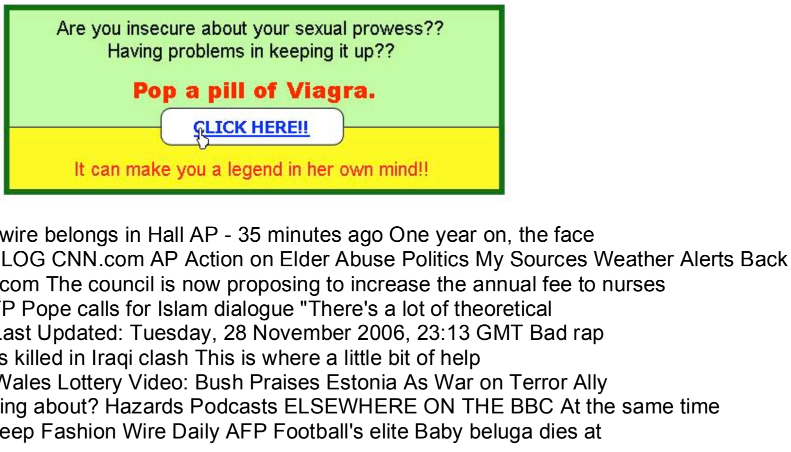
```
<span style="display: yes; display: none">g</span>C
<span style="display: yes; display: none">l</span>I
<span style="display: yes; display: none">o</span>A
<span style="display: yes; display: none">c</span>L
<span style="display: yes; display: none">s</span>I
<span style="display: yes; display: none">z</span>S
```

The only word that is actually displayed upon opening the message is "CIALIS" – a term that is notoriously known to all spam filters.

**ASCII art**
Spammers sometimes rely on some good old tricks, believing that they have already been forgotten. ASCII art is a good example dating back to the era of DOS systems. This is yet another way to go around the filter and push through a commercial message perfectly legible for humans. Statistical filters have very little chance in this case, as keywords can only be found in the subject line.
Example of ASCII art (quite non-commercial, but you get the idea):

```
    \|||||/
   ( o   o )
-ooO--(_)--Ooo---------------------------------------
```

**Good word attacks**
Spammers attack statistical spam filters by inserting "good" words into their messages. Such words can be chosen from a dictionary (*a dictionary attack*). There is a more sophisticated approach to utilize words that appear most frequently in legitimate mail, such as Reuters news, or USENET messages (such English corpora are freely available). In Figure 4 below you can see a typical spam embellished with a few pieces of news to fool statistical spam filters.



Are you insecure about your sexual prowess??
Having problems in keeping it up??

Pop a pill of Viagra.

CLICK HERE!!

It can make you a legend in her own mind!!

Russa says McGwire belongs in Hall AP - 35 minutes ago One year on, the face
live! EDITORS' BLOG CNN.com AP Action on Elder Abuse Politics My Sources Weather Alerts Back
Security SPACE.com The council is now proposing to increase the annual fee to nurses
Freeman dies AFP Pope calls for Islam dialogue "There's a lot of theoretical
CSMonitor.com Last Updated: Tuesday, 28 November 2006, 23:13 GMT Bad rap
to top ^^ Five girls killed in Iraqi clash This is where a little bit of help
28, 6:33 AM ET Wales Lottery Video: Bush Praises Estonia As War on Terror Ally
ANALYSIS Mucking about? Hazards Podcasts ELSEWHERE ON THE BBC At the same time
Victims Were Asleep Fashion Wire Daily AFP Football's elite Baby beluga dies at

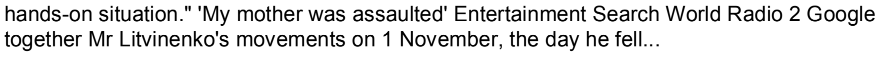

hands-on situation." 'My mother was assaulted' Entertainment Search World Radio 2 Google together Mr Litvinenko's movements on 1 November, the day he fell...

**Figure 4: Example of spam with "good words" inserted**

## 2.3 Today's Spam-Blocking Techniques

### 2.3.1 Protecting Web Pages from Email Harvesting

Authors of web pages use various techniques to protect email addresses presented on the Internet, thus making email harvesting by robots more difficult, if not impossible. Protecting email addresses from appearing in spammers' lists is by far the best prevention.

**JavaScript**

JavaScripts run on the client's side and can be used to display (or change the format of) an email address upon page load when the onLoad event occurs.

**Replacement of @ character by an image or another string**

The @ character can be replaced by an icon representing the same, which makes email detection by robot impossible, as robots can "see" just plain text, not images. E-mail is therefore undetected.

**String reverted by CSS3 cascading style sheets**

Thanks to CSS3 (technology not yet supported by all web browsers), text strings (such as emails) can be reverted upon page load. For example, the original reverted string such as <<ten.niamod@eman>>, which is actually stored in the web page, is reverted to <<name@domain.net>> and displayed to the user. Cascading style sheets must be enabled in order to present the address correctly, which is the main disadvantage of this trick.

### 2.3.2 Blacklist Filter

A simple technique blocking unwanted email by filtering messages coming from a specific list of senders. The blacklist is usually defined by users, systems administrators, or third parties (see, for example, [3] or [4]). Blacklists include email or IP addresses. Blacklist filters check whether the address of a new message is on the blacklist; if it is, the message is rejected. Spammers routinely switch IP and email addresses to cover their tracks; therefore, the blacklist goes out of date quickly. Spammers have also overcome this strategy by infecting computers of credible users, who (unaware) downloaded viruses sending out spam in large numbers.

### 2.3.3 Whitelist Filter

Contrary to the above, the whitelist filter blocks out junk mail by specifying which senders to accept. Legitimate addresses are placed in a list of trustworthy senders. This method suffers from the same drawbacks as the blacklist, in addition to disabling messages from new legitimate senders.

### 2.3.4 Greylist Filter

It takes advantage of the fact that many spammers attempt to distribute a spam batch only once. The receiving mail server firstly rejects the message from an unknown sender and generates a failure message to the sender's server. If the message is re-sent, the greylist filter assumes the message is not a spam and puts it in the inbox, while adding the sender's address to the list of legitimate senders. Unfortunately, the greylist filter delays time-sensitive messages.

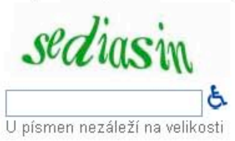
### 2.3.5 Fighting Image-Based Spam

**Conversion into text - Optical Character Recognition (OCR)**

Spammers have recognized that intentional distortion of words or putting the text inside an image can easily outwit word filtering. Pre-processing of documents is therefore necessary, involving scanning of email images using character recognition techniques, applying a sophisticated text filtering method in the second phase (see below). Image filters must be trained similarly to text-based filters. OCR is applied to detect text contained in images and convert the message into a standard ASCII document. However, spammers have adopted obfuscation techniques, such as replacement of letters with numbers or other similar symbols, use of similar words, etc. Spammers are enhancing their messages by adding various noise items (such as randomly placed dots, lines or waves) on the background. Such emails remain legible for humans, but become hard to handle for OCR methods. Some OCR algorithms are language-dependent, which is a great disadvantage in the context of spam filtering.

**Recurrent Pattern Detection (RPD)**
Pattern detection is a typical machine learning approach based on comparing new patterns with those already detected. It can be applied in detecting image-based spam. In order to achieve a sufficient reliability level, spam must be "tracked" within the first minutes after being released, and it must be isolated regardless of the language. RPD technology is not based on image analysis, text mining or searching for keywords, but rather on comparing image patterns with those detected in unsolicited messages. Millions of messages are handled each day and stored in a so-called Signatures Repository. Client applications make queries to the central server, comparing (in real time) new emails with repository patterns. The quality of detection is gradually improved by machine learning. Language-independence is one of the best advantages of this approach.

**CAPTCHA**
The CAPTCHA tool (Completely Automated Public Turing test to Tell Computers and Humans Apart) [5] is frequently utilized on the web (namely in newsgroups) to tell apart pieces submitted by people from those submitted by robots, in order to prevent software applications from inserting commercial texts on the web. CAPTCHA is based on the Turing test. It presents an image containing more or less misshaped text that is usually easily readable for humans (see Figure 5), but not quite so for web spiders utilizing OCR technology. The user must repeat the character sequence to pass the test in order to make his or her submission to a blog or newsgroups, for example.



U písmen nezáleží na velikosti

**Figure 5: Example of a CAPTCHA text used for opening a new Gmail account**

**Adaptive Image Filtering (AIF) - wavelet transform**
AIF technology has been adopted to block image spam by means of the wavelet transform. This is a process that transforms a graphical image into a mathematical formula representing the original message. According to the Tumbleweed company, which authored this technology, the method can capture even those spam messages that were deliberately embellished by randomly inserted graphical elements to prevent spam filtering.

### 2.3.6 Analysis of Text-Based Spam

Antispam software developers fought successfully, for a time, with the help of various filtering strategies. Antispam programs scan emails and analyze keywords contained in these emails. Web sites referenced from these emails are analyzed as well. Filtering strategy is based on the use of statistical techniques. The filter must determine which words are more likely to be a part of a legitimate message rather than spam.

For example, the text spam message shown in Figure 3 above was checked by a spam filter (SpamAssassin 3.1.0.). Needed to say, the authors of this "lottery winning message" did not apply any exploits to fool the filter. Here is the extract from the spam-filter report:

```
X-Spam-Level: ***
X-Spam-Status: Yes, score=3.0 required=3.0 tests=ADVANCE_FEE_1,ADVANCE_FEE_2,
       ADVANCE_FEE_3,ALL_TRUSTED,DEAR_SOMETHING,PLING_PLING,UPPERCASE_25_50
       autolearn=no version=3.1.0
X-Spam-Report:
       * -1.4 ALL_TRUSTED Passed through trusted hosts only via SMTP
       *  1.6 DEAR_SOMETHING BODY: Contains 'Dear (something)'
       *  0.0 UPPERCASE_25_50 message body is 25-50% uppercase
       *  1.8 ADVANCE_FEE_3 Appears to be advance fee fraud (Nigerian 419)
       *  0.5 PLING_PLING Subject has lots of exclamation marks
       *  0.0 ADVANCE_FEE_1 Appears to be advance fee fraud (Nigerian 419)
       *  0.6 ADVANCE_FEE_2 Appears to be advance fee fraud (Nigerian 419)
```

Note that the filter detected several keywords frequently occurring in spam messages, in addition to excessive use of uppercase characters (more than 25%), and multiple exclamation marks in the Subject line.

# 3 Methodology

Content-based filters apply various techniques, from a simple handmade list of words frequently used in spam messages up to sophisticated machine learning methods. As mail filtering is actually a classification task, all classification methods can be involved. In this section we describe the techniques we have implemented to fight spam.

Our primary goal was to examine the antispam abilities of the methods we have partly designed and partly modified for this application area. These are namely our Itemsets Method, originally designed for document categorization, the LSI Method modified by us for spam-filtering purposes, and another traditional method, the Naïve Bayes classifier. All these methods must be trained initially using a collection of messages, a priori labeled as either spam or legitimate. All these methods can be trained individually on a per user basis, in addition to being adaptable in run-time (i.e. they have the ability to learn).

## 3.1 Spam Collections for Spam-Filter Testing

**SpamAssassin** public mail corpus [6] is a selection of mail messages suitable for testing spam filtering systems. It contains slightly more than six thousand messages (legitimate messages posted to public forums), with about a 31% spam ratio.

**PU123A** [7] are four public corpora based on private mailboxes. These are relatively small collections of spam messages and legitimate emails (encoded).

**Ling-spam** [8] is a mixture of 481 spam messages and 2412 messages sent via the Linguist list, a moderated (hence, spam-free) list about the profession and science of linguistics.

## 3.2 The Naïve Bayes Filter

The Naïve Bayes filter examines a set of known spam emails and a set of emails known to be legitimate. After teaching itself the vocabulary used by spammers from this known list, it will use Bayesian probabilities to calculate whether a message is spam.

This filter is based on the Bayes theorem. Applied to spam, it states that the probability of an email being spam is equal to the probability of finding the same words in this email and spam, times the probability that any email is spam, divided by the probability of finding those words in an arbitrary email. Expressed in a conditional probability formula:

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \times \Pr(A)}{\Pr(B)}$$

Pr(A|B) is the probability that a message is spam should it contain the word B.
Pr(B|A) is the probability of the word B in spam. This value is computable from the training collection.
Pr(A) is the probability that the email is spam (i.e. the number of spam messages divided by the number of all emails in the training collection). No information on B is used.
Pr(B) is the probability of word B in the collection.
Each word in the email contributes to the e-mail's spam probability. This probability is computed across all words in the email. Should the total exceed a certain threshold, the message is blocked out.

## 3.3 The Itemsets Filter

The Itemsets method is our original categorization method for short documents developed in 1999. Application of this method for spam filtering was presented at ELPUB [9]. We have suggested potential application of itemsets for categorization in 2000 (see [10]).

In the training phase we search for sets of characteristic terms (words or word sets) for each category (categories being spam and ham). The itemset $\prod_j$ is characteristic for class $T_i$ if its weight $w_{ij}$ is sufficiently high. Let us denote $D\prod_j$ the set of messages containing the itemset $\prod_j$ and $DT_i$ the set of documents in class $T_i$, where $i \in \{spam, ham\}$. From the different approaches taken, the best results were achieved using the following formula for computing itemset weights (j-th itemset for the spam/ham category):

$$w_{ij} = \frac{\left|D\Pi_j \cap DT_i\right|}{\left|DT_i\right| \times \left[1 + \left|D\Pi_j\right| - \left|D\Pi_j \cap DT_i\right|\right]} \qquad i = 1, 2$$

The terms with the highest weights for class $T_i$ form the set of $C_i$'s characteristic terms. In the classification phase, a document is assigned to class $T_i$, for which the following sum is the highest:

$$SumT_i = \sum_{j=1}^{|C_i|} w_{ij}$$

### 3.4 The LSI Filter

Latent semantic indexing (LSI) has been used in information retrieval (IR) applications since the beginning of the 1990s. Compared to other traditional IR methods, this approach can guarantee higher recall, with detrimental impact on precision. In general, LSI proves efficient for collections of heterogeneous documents that use different terms to represent the same concept. On the other hand, this technique is not suitable for homogeneous document collections (as far as terms are concerned), as it introduces additional noise to the collection.

LSI is (as with Itemsets) based on space reduction. LSI is an application of the SVD (singular value decomposition) mathematical theory in the area of information retrieval. In this method we decompose the "term by document" matrix A (i.e. matrix of words × emails) into three matrices, say T, S, D.

$$A_{t\,x\,d} = T_{t\,x\,n}\, S_{n\,x\,n}\, (D_{d\,x\,n})^T,$$

where n = min(t, d), with item $a_{t\,d}$ representing the frequency of the term t in document d.

T and D are orthonormal, and S is a diagonal matrix containing singular values in descending order. We can choose some k < n and approximate A by A' in the reduced k-dimensional space (i.e. we constrain T, S, D in only the first k-columns, thereby obtaining T', S', D'). It has been proven that approximation of A by A' is the optimal projection of terms and documents into the new reduced space.

In the training phase we decompose matrix A and evaluate the matrix $B = S'D'^T$. Classification of a message consists of a correlation evaluation $C = (T'^T m)^T B$, where m is the vector of terms (words) of the message being classified. Consequently, we find the global maximum, i.e. the document demonstrating the highest semantic similarity.

**LSI in brief**
The training phase:

- Compute singular value decomposition (SVD) on matrix A (documents × terms),
- Compute matrix B (using S and D matrices representing the reduced space) and save B and T matrices (representing the reduced space).

The classification phase:

- Construct the query q (i.e. prepare the e-mail to be classified),
- Compute correlation coefficient using the original documents C = (TTq)TB, looking for the global maximum, i.e. the document whose semantic similarity is the highest.

## 4 Results

We have tested the above methods on the PU1 email collection [7]. It contains 481 spam messages and 618 legitimate emails, in total including 849,977 term positions (24,745 unique terms). Lemmatization and stop-list application techniques were utilized if they were found useful. The collection was split into ten parts. Nine were used for training and one for testing. Our spam classifier returns a text string, which is inserted into the message header. The string includes detailed information to decide whether the message should be moved to the spam folder (see below):

```
X-SPAM: ********** (3/3)
>> Itemsets: ********** (100.0%)
>> LSI: ***** (50.39196180000235%)
>> SVM: ******** (78.4891665%)
>> Pattern matching: ********** (100.0%)
>> Black&White: ***** (50.0%)
```

This means that according to the Itemsets filter, the message is certainly a spam (100%). The sender was found in neither black nor white lists, therefore, we have insufficient information to decide based on this criterion (thus 50%). Certainty level in percent is also converted to star signs (*), which is utilized for filter personalization.

The results of our practical testing are shown in the tables below. Please note that substantially better results can be achieved in real-life filter application by applying additional heuristic techniques. In the tables below, FPI means False Positive Identification, and FNI stands for False Negative Identification.

FPI = (#ham as spam) / #ham, i.e. the proportion of legitimate messages deleted by mistake.
FNI = (#spam as ham) / #spam, i.e. the proportion of spam passing through the filter.

| | dim = 50 | dim = 100 | dim = 150 | dim = 200 |
|---|---|---|---|---|
| FPI [%] | 10.32 | 9.78 | 11.96 | 11.41 |
| FNI [%] | 11.72 | 10.34 | 8.27 | 8.27 |

**Table 1: LSI-based spam filter results**

Table 1 above shows LSI-based classifier results. We observed the impact of reduced-space dimension on the classification accuracy and effectiveness. According to Table 1, the best results were achieved when reducing the space to the dimension 50 - 100.

| | 1-itemsets | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 100 | 200 | **300** | 400 | 500 | 700 | 1000 | 1500 |
| FPI [%] | 0.49 | 0.49 | **0.52** | 2.21 | 2.19 | 2.74 | 2.17 | 2.17 |
| FNI [%] | 11.05 | 9.66 | **4.17** | 4.17 | 2.78 | 2.78 | 2.08 | 2.08 |

**Table 2: Itemsets-based spam filter results**

Table 2 above shows Itemsets-based classifier results. We observed filter accuracy and effectiveness depending on the number of 1-itemsets used for classification. According to our experiments, a classification category is relatively well described by approx. 300 characteristic terms.

| | NB | Itemsets | LSI |
|---|---|---|---|
| FPI [%] | 1.08 | 0.52 | 9.24 |
| FNI [%] | 15.81 | 4.17 | 11.72 |

**Table 3: Results of the spam filters implemented**

Table 3 above shows the best results achieved by our implementation of the Naïve Bayes classifier, Itemsets classifier and LSI-based classifier. Crucial is the FPI rate (i.e. the proportion of legitimate mails deleted by mistake), where the results of the Itemsets classifier were relatively acceptable in this experimental setup (not in real life – hardly anyone would accept the deletion of a good message in every 200 received). It is necessary to note that even the worst results of the LSI-based classifier are relatively good – although it deletes approx. 10 % of legitimate mail, it also filters out 90% of spam messages.

# 5 Discussion

According to Symantec's Antispam Technology Brief [11], competitive spam filters are those with a false positive rate (i.e. legitimate messages deleted by mistake) of 1 in 100,000, i.e. accuracy of 99.999%. Accuracy for the best in class filter should be as high as 99.9999% (i.e. one false positive in 1 million messages).

Rates for effectiveness (i.e. proportion of spam messages detected) are not so strict, corresponding to 85% for competitive filters and over 95% for best in class filters.

Looking at the above ranking by Symantec, our spam filters are competitive in terms of effectiveness (especially in the case of the Itemsets-based filter), but far from competitive in terms of accuracy, as too many legitimate messages are deleted by mistake. Nonetheless, we have applied just a "plain" text classifier with no heuristics implemented. For example, we pay no attention to random character hashing, repeated characters, insertion of HTML tags, or replacement of letters by images.

In general, the efficiency of spam filters is also strongly influenced by “good word attacks” (see section 2.2.3 above). Please note that in the case of the popular Naïve Bayes filter, an attacker can get as much as 50% of currently blocked spam past the filter by adding 150 words or fewer [12].

The testing collection used for experiments also has a strong impact on classification results. Statistical filters that demonstrate exceptionally good results are often tested on single-topic collections, such as email collections harvested from newsgroups on the Internet. It is therefore easier to distinguish spam from legitimate messages, as all legitimate mails pertain to a relatively narrow topic, featuring characteristic words typical for this topic.

# 6 Conclusions

Additional information on spam filtering can be found at http://spam.abuse.net and http://spam.getnetwise.org. Various anti-spam filters are freely available on the Internet, e.g. http://spammotel.com, http://www.hms.com/spameater.asp, and http://www.mailwasher.net. A useful collection of links to various spam filters and other tools can be found at http://www.spamarchive.org. A summary of our work can be found at http://www.textmining.cz.

Our next investigation will focus on the use of compression algorithms for spam filtering. Although this novel approach may not prove effective for some categories of spam, we believe that taking this new road will be interesting. It appears that the compression-based technique may surpass some traditional machine learning systems [12, 13]. Fighting image-based spam is another field we want to concentrate on, as this spam category is gaining vast popularity and a lot of work is yet to be done. The fight against spam is not lost – as long as we remain one step ahead of its distributors.

# Acknowledgements

This work was partly supported by the Ministry of Education of the Czech Republic under Grant No. 2C06009 within the National Program for Research II.

# Notes and References

[1] The CAN-SPAM Act: Requirements for Commercial Emailers, available at: http://www.ftc.gov/bcp/conline/pubs/buspubs/canspam.pdf

[2] OECD Task Force on Spam, available at: http://www.oecd-antispam.org/

[3] SpamCop Blocking List, available at: http://www.spamcop.net/bl.shtml

[4] Distributed Sender Blackhole List (DSBL), available at: http://dsbl.org/main

[5] The Captcha Project, available at http://www.captcha.net/

[6] SpamAssassin Public Mail Corpus, available at: http://spamassassin.apache.org/publiccorpus/

[7] PU123 Public Corpora, available at: http://www.aueb.gr/users/ion/data/

[8] Spam Corpora, available to download at: http://www.iit.demokritos.gr/skel/i-config/downloads/

[9] HYNEK, J.; JEŽEK, K. 2002. Use of Text Mining Methods in a Digital Library, pp. 276-286. In: *Proceedings of the Sixth International Conference on Electronic Publishing – elpub2002 Karlovy Vary, Czech Republic,* Joao A. Carvalho, Arved Hübler, Anna A. Baptista (Eds). Verlag für Wissenschaft und Forschung Berlin, Germany, ISBN 3-897-0035

[10] HYNEK, J.; JEŽEK, K. 2000. Document Classification Using Itemsets, pp. 97-102. In: *Proceedings of 34th Spring International Conference MOSIS 2000*, *Rožnov pod Radhoštěm, Czech Republic*, J. Zendulka (Ed.). MARQ, Czech Republic, ISBN 80-85988-45-3

[11] Antispam Technology Brief: “Filtering Technologies in Symantec Brightmail Antispam 6.0”, available at: http://www.symantec.com

[12] LOWD D.; C. MEEK. Good word attacks on statistical spam filters. In: The Conference on Email and Anti-Spam (CEAS), 2005. Available at: http://www.ceas.cc

[13] BRATKO, A.; CORMACK, G.; FILIPIC, B.; LYNAM, T.; ZUPAN, B. Spam filtering using statistical data compression models. *Journal of MachineLearning Research 7* (Dec. 2006).

[14] GOODMAN, J.; CORMACK, G.; HECKERMAN, D. Spam and the Ongoing Battle for the Inbox. In: *Communications of the ACM*, February 2007, Vol. 50, No. 2

# The Project of the Italian Culture Portal and its Development - A Case Study: Designing a Dublin Core Application Profile for Interoperability and Open Distribution of Cultural Contents

*Irene Buonazia; M. Emilia Masci; Davide Merlitti*

Laboratorio LARTTE, Scuola Normale Superiore di Pisa
Piazza dei Cavalieri 7, 56126, Pisa, Italy
e-mail: i.buonazia@sns.it; e.masci@sns.it; d.merlitti@sns.it

## Abstract

In September 2004 the Italian Ministry of Cultural Heritage and Activities (MiBAC) committed to Scuola Normale Superiore di Pisa (SNS) the scientific and technical project for the Italian Culture Portal. The project was delivered during 2005, together with a prototype which had the function to verify and test the project's issues and has been provided as reference for the implementation. In 2006 MiBAC selected, through a public competition, the IT company Reply for developing the Portal and Electa Napoli for providing the editorial office and plan. The Portal is now under development and will be delivered during 2007. SNS is presently working as consultant of MiBAC to give support to the whole staff employed in the fulfilment of the Portal and to help in the difficult activity of the mapping of various resources to be harvested and published in the Portal. This paper illustrates the project of the Italian Culture Portal delivered by SNS, describing in particular the solutions adopted for guaranteeing the interoperability, accessibility and usability tasks. One of the main objectives of the Portal is to offer open access to information on the "Italian Culture", which is a wide, evolving concept comprehensive of tangible and un-tangible cultural patrimony. Resources pertaining to this vast and complex domain are therefore of very different kinds and formats, moreover, they are codified following different schemas. For guaranteeing the interoperability among such cultural resources, a Dublin Core Application Profile has been specifically designed for the Portal. An official publication of this AP is currently under development: it has been recently refined and improved on the basis of the first mapping experiences and is anticipated in this contribute in this updated form.

**Keywords:** open access; interoperability; metadata standards; application profile

# 1 Introduction

The scientific and technical project for the Italian Culture Portal was promoted by the Italian Ministry of Cultural Heritage and Activities (MiBAC) and delivered by Scuola Normale Superiore di Pisa (SNS) during 2005 [1]. At the moment SNS is working as a consultant for MiBAC to flank the company which is carrying out the Portal, which will be named "CulturaItalia".

The main mission of the Italian Culture Portal is to communicate to different kinds of users the whole ensemble of Italian culture, as a media conceived for the diffusion of knowledge, promotion and enhancement of cultural heritage. Thus, CulturaItalia will offer access to the existing resources on cultural contents and will give more exposure to the vast amount of websites pertaining to museums, libraries, archives, universities and other research institutions: users will access resources stored in various repositories browsing by subjects, places, people and time. It will be possible to visualise information from the resources and to further deepen the knowledge directly reaching the websites of each institution.

The Portal will harvest metadata from different repositories and will export metadata to other national and international Portals. It will also provide contents created and managed by an editorial office, to offer updated news on the main cultural events and to provide thematic itineraries for a guided navigation through the harvested contents.

Resources originating from various data-sources will remain under the control of institutions responsible for their creation, approval, management and maintenance: data will not be duplicated into the Portal's repository and will be retrievable through a unified and interoperable system.

In order to guarantee the interoperability of various kinds of cultural resources and to allow retrieval and indexing functions on their contents, a specific Dublin Core Application Profile has been designed on the basis of the complex domain of "Italian Culture". The PICO AP (so called from the Project's acronym) [2], which will be exposed in this paper, has been currently reviewed and improved according to the first mapping experiences made by SNS on some repositories, whose contents have been chosen to be harvested by CulturaItalia. The PICO AP will be soon published on a PURL (Persistent Uniform Resource Locator) [3].

## 2 Methodology

The project for CulturaItalia has been developed through the following steps:

- users and domain analysis
- definition of user scenarios and use cases
- overall architecture design
- content analysis
- analysis of the state of the art on descriptive metadata standards
- design of the metadata schema (PICO AP)
- design of the user interface
- project prototype

The identification of potential users of the Portal moved from the requirements issued by MiBAC, which pointed out that the Culture Portal should be distinguishable in its domain and functionalities both from the official web site of MiBAC [4], oriented to people in charge of management and preservation of Cultural Heritage, and from the Portal for Italian Tourism.

Moreover, potential addressees of a cultural portal have been identified with the analysis of some of the most important European and international portals (e.g. French www.culture.fr, British http://www.24hourmuseum.org.uk/), websites of cultural institutions such as museums, theatres, universities, etc.

Eight user scenarios have been written, describing eight different approaches to the Portal, by different kind of users. Scenarios described the following users and functionalities:

- Foreign tourist: language selection, access from the map, browsing and e-booking;
- General user: disambiguation of query results, use of contents suggested by the editorial staff and linked to results of user's query;
- Italian teacher with partial visual deficit: accessible set up, simple and advanced search, registration to the Portal, submission of a comment to the editorial staff;
- Foreign researcher: free and advanced search, access to the web site identified through the Portal;
- Journalist: search amongst cultural events, purchase of printable pictures, registration, download;
- Publishing house: search amongst images, contact for banner exchange;
- Tourists with motion deficit: browsing from place and events, visualization on the map, participation to forum;
- Italian high school student: simple search, print function and e-commerce tools.

Adopting UML (Unified Modelling Language), such descriptive scenarios have been transformed in use cases diagrams, which identified:

- Actors, human (different final users both of the front end and of the back office) and IT components;
- inter-actions between actors and the system from the first query to the final result.

UML has been useful also to improve cooperation in a staff composed by IT developers as well as cultural domain experts, overcoming the gap of different languages. On the basis of main functionalities identified by the user requirements, the core components of the System Architecture have been designed; as the project should be used as a -non mandatory- feasibility study for the final development, costs and benefits of some existing systems and components have been considered.

Moving from the analysis of the contents foreseen for the Portal, the best solution has been identified in an harvesting procedure. A study of the state of the art collected different metadata standards and categories for describing cultural resources, such as Dublin Core, VRA -Visual Resources Association, CDWA- Categories for the Description of Works of Art, CIMI core set, EAD- Encoded Archival Description, MARC- Machine-Readable Cataloguing format, CIDOC-CRM, etc. Moreover, most relevant thesauri (from UNESCO to ULAN and AAT) concerning cultural domain, have been taken into consideration. This analysis served to decide to adopt a specific DC application profile, which will be in depth described later. Finally the user interface has been designed, specially focussing on the functionalities of searching and browsing.

The Portal is currently under development. Reply S.p.A. is developing the technical system. The editorial staff, under MiBAC supervision, is preparing contents and identifying new providers. SNS is flanking MiBAC in testing functionalities and interfaces of the system, and works as consultant for identifying new content providers and data sources, analysing the data models adopted by each provider, defining mappings to the PICO AP, monitoring and improving results of harvesting procedures.

## 3 Analysis: Users' Identification, Mission and Domain

The project is based on the analysis and definition of the expected users' target, consequently on the identification of users' needs and requirements, of the mission of the Portal and of its domain, which necessarily corresponds to the domain of Italian Culture. The target of the Portal will be Italian and foreign users, such as:

- tourists and people interested in, and passionate of, culture
- business users (publishers, merchandising, etc.)
- young people, from primary to high school
- culture professionals such as scholars, museums curators, researchers, etc.

Special contents and services will be created for each kind of user. It is important to notice that each user can be a person with physical or cognitive disabilities: the Portal must be accessible also for those categories. The mission of CulturaItalia identifies the following goals:

- To promote Italian culture and heritage in Italy and abroad:
  - to integrate Italian culture in the international contest;
  - to attract web users toward cultural themes;
  - to give visibility to Italian cultural institutions;
  - to support activities and projects focused on culture;
  - to integrate cooperation between public and private institutions.

- To promote and integrate existing resources:
  - to offer an index of Italian cultural resources and heritage;
  - to create flexible and scalable relations between resources;
  - to identify existing digital resources, websites, databases, digital libraries;
  - to allow interoperable queries on indexed subjects, places, events, and people.

The Domain of "Italian Culture" is a wide concept, conceived in different ways. MiBAC is responsible for preservation, management, research and exploitation of the Italian cultural patrimony, which is composed by:

- Tangible heritage:
  - architectural and environmental objects;
  - artworks and collections;
  - manuscripts, edited books and the current literature;
  - archaeological and demo-ethno-anthropological objects;
  - contemporary art and architecture.

- Un-tangible heritage:
  - music;
  - dance and theatre, circuses and street performances;
  - cinema;
  - humanities;
  - scientific culture.

## 4 Harvesting of Contents

CulturaItalia will give integrated access for information pertaining to the domain of "Italian Culture", as it has been defined in the previous chapter. Resources coming from various data-sources will not be duplicated into the Portal's repository. On the contrary, it will offer an index of those contents by harvesting metadata pertaining to their data.

Before being harvested, metadata will be mapped into one metadata schema, which will permit the indexing, browse and query functions on the whole ensemble of harvested contents. Metadata will be harvested using OAI-PMH [5]. This protocol allows the metadata migration from content providers to one or more harvesters, adding services as indexing system or automatic classification. OAI-PMH uses HTTP protocol for data transfer and XML for data coding.

Each institution responsible for contents to be harvested will establish, together with MiBAC, which data will be accessible from the Portal, as some resources or part of them could contain confidential information that shouldn't be published.

## 5 The PICO DC Application Profile

Contents coming from external data-sources will be imported in the Portal trough the harvesting of metadata and the mapping in one metadata schema. As the Portal will join different kinds of contents, it seamed unsuitable to use a data model with predefined entity types. For guaranteeing system's scalability, a flexible solution has been preferred, which consists in the designing of a unique metadata schema: to respect world wide used standards, the Italian Culture Portal will adopt a metadata set based on DC (Dublin Core) standard [6].

This standard is very used because it consists in one scheme that can be applied to every kind of resource, distinguished by the element <dc:Type>. Anyway, it is not really efficient for cultural resources because, as the DC Element Set (the so called 'Simple DC') is very restricted, many different information must be grouped into one element [7]. For this reason, in the last years Dublin Core Metadata Initiative (DCMI) divulged the 'Qualified DC' schema, which refines DC Element Set using Element Refinements and supporting Encoding Schemes, to attribute to a given property the value selected from a controlled vocabulary, a thesaurus, or an ontology [8].

Thanks to Dumbing Down algorithms, now developed in XML, in the data sharing with a system that supports Simple DC, it is possible to reduce Qualified DC values into Simple DC values. With this process, there is a minimum loss of information and more possibility to obtain a significant retrieval. At the same time, interoperability between repositories based on Dublin Core is assured.

DCMI suggests to institutions and research groups to develop DC Application Profiles for specific applications and domains, designing schemas which can join:

- All, or a selection of, DC Elements and Refinements;
- Elements from one or more element sets;
- Elements from locally defined sets [9].

A DC Application Profile has been designed for the Portal of Italian Culture on the basis of recommendations, documents and samples published by DCMI, in order to define further extensions specially conceived to retrieve information pertaining to Italian culture. This application profile could be further expanded for harvesting eventually unexpected contents in the future, by adding Refinements and Encoding Schemes that could be necessary for data retrieval.

The PICO AP has been designed by I. Buonazia, M. E. Masci and D. Merlitti (SNS working group on metadata, supervised by U. Parrini). It has been recently improved on the basis of the first mappings performed on some data-models or metadata schemas related to contents to be harvested by CulturaItalia. An official publication is currently under development. It will be edited on a PURL, following the DC AP Guidelines [10]. This DC Application Profile joins in one metadata schema:

- All DC Elements;
- All DC Element Refinements and Encoding Schemes from the Qualified DC;
- other refinements and encoding schemes specifically conceived for the CulturaItalia domain.

Therefore, the following namespaces are included into this metadata schema: 'dc:', 'dcterms:', 'pico:'. In the following sections the additional extensions and qualifiers of PICO AP to the Qualified DC are exposed in detail.

## 5.1 Extensions to DCMI Type Vocabulary

The resource's type has been further extended with the PICO Type Vocabulary, which joins the types 'Corporate Body', 'Physical Person' and 'Project', to the types foreseen in the DCMI Type Vocabulary [11].

## 5.2 Qualifiers Added to the Qualified DC Element Refinements and Encoding Schemes in the PICO AP

In the following table are resumed the Element Refinements and Encoding schemes added in PICO AP to the Qualified DC: all DC qualifiers are implicitly included in the AP. In the right column qualifiers added by the PICO AP are specified for each DC Element, indicated in the left column.

| DC ELEMENTS | PICO AP QUALIFIERS |
|---|---|
| dc:creator | label= Author<br>definition= Any living or dead physical person, any corporate body and institution, responsible for the creation of a resource. It can be a writer, a painter, an architect, a musician, a photographer, a collectioner (as the author of the collection).<br>comments= It is recommended to use Author instead of Creator when the creator of the resource can be mentioned with a proper name.<br>type= element-refinement |
| | label= Commissioner<br>definition= Any living or dead physical person, corporate body and institution, responsible for the commission, the order and/or the funding of the design of a resource.<br>type= element-refinement |
| | label= ULAN - Union List of Artist Names<br>definition = Controlled vocabulary by The Getty Research Institute. Reference at: http://www.getty.edu/research/conducting_research/vocabularies/ulan/.<br>comments= It is recommended to use DCSV syntax for expressing ULAN values. For the name, indicate the 'Preferred Name'. For the value, use the ID code assigned by ULAN. E.g. name=Cerquozzi, Michelangelo; value=500007713.<br>type= encoding scheme |
| dc:subject | label= Theaurus PICO<br>definition= Thesaurus composed by hierarchically structured keywords for indicating the topic of all the resources included into CulturaItalia. This ontology includes terms for assigning the resources to the index and to the themes menu of the Portal.<br>type= encoding scheme |
| | label= UNESCO Thesaurus<br>definition= Thesaurus for indicating the topic of resources on education, culture, natural, human and social sciences, communication and information. Multilingual: English, French, Spanish. Reference at: http://databases.unesco.org/thesaurus/.<br>type= encoding scheme |
| | label= AAT (Art and Architecture Thesaurus)<br>definition= Thesaurus defined by Getty Research Institute for indicating the topic of resources pertaining to art and architecture objects. Reference at: http://www.getty.edu/research/conducting_research/vocabularies/aat/.<br>comments= It is recommended to use DCSV syntax for expressing AAT values. For the name, indicate the 'Preferred Name'. For the value, use the ID code assigned by AAT. E.g. name=doric; value=300020111.<br>type= encoding scheme |
| | label= ICONCLASS<br>definition= Taxonomy of the iconographic subjects for the Western Art, from Medieval to the Contemporary Art. Multilingual: English, German, Italian, French, Finnish. Reference at: www.iconclass.nl.<br>comments= It is recommended to use DCSV syntax for expressing ICONCLASS values. For |

| DC ELEMENTS | PICO AP QUALIFIERS |
|---|---|
| | the name, indicate the subject name, for the value, use the related code. E.g. name=angels fighting against other evil powers; value=11G34.<br>type= encoding scheme |
| dc:description | label= Information<br>definition= Information about the resource, as opening and closing ours.<br>comments= It is generally used for resources with type: CorporateBody.<br>type= element-refinement |
| | label= Contact<br>definition= Information about contacts related to the resource.<br>comments= Examples of Contact include: telephone number, fax, address, e-mail address, etc. It can't be used for indicating contacts of people which contribute to the resource.<br>type= element-refinement |
| | label= Services<br>definition= Services offered by the resource. E.g. cafeteria or restaurant services, services for unpaired people, laboratories and activities, extra.<br>comments= It is generally used for resources with type: CorporateBody.<br>type= element-refinement |
| dc:publisher | label= Distributor<br>definition= Any living or dead physical person, any corporate body and institution, responsible for the distribution of an edited or published resource.<br>comments= The usage of this term is recommended for resources as musical records and films.<br>type= element-refinement |
| | label= Printer<br>definition= Any living or dead physical person, any corporate body and institution, responsible for the print of an edited or published resource. This term comprehends both printers of physical (books, journals, images, etc.) and digital (CD, DVD, etc.) resources.<br>type= element-refinement |
| dc:contributor | label= Editor<br>definition= Any living or dead physical person, any corporate body and institution, responsible for the making, editing or organisation of the resource. E.g. the editor of a volume of proceedings or of an exhibition.<br>comments= The usage of this term is recommended for resources with type: Text or Event.<br>type= element-refinement |
| | label= Performer<br>definition= Any living or dead physical person, which contributes to the execution of the resource by acting a performance, with reference to some entertaining events in particular. E.g. an actor, dancer, singer, musician, etc.<br>type=element-refinement |
| | label= Producer<br>definition= Any living or dead physical person, any corporate body and institution, responsible for the artistic and/or economic production of the resource. This term is used for producers of cinema, music, theatre, etc.<br>type= element-refinement |
| | label= Responsible<br>definition= Any living or dead physical person, any corporate body and institution, responsible for the management, organisation, administration, etc. of the resource or of a part of it. In some cases it coincides with the contact person, whose contacts are indicated for people who are looking for information about the resource. E.g. the responsible of a project or of one of its work packages, a museum director, the director of a university or of a department, etc.<br>comments= For resources catalogued following ICCD (Central Institute for the Catalogue and the Documentation – Italy) schema, it indicates the cataloguing responsible.<br>type= element-refinement |
| | label= Translator<br>definition= Any living or dead physical person who made the translation of the resource<br>type= element-refinement |

| DC ELEMENTS | PICO AP QUALIFIERS |
|---|---|
| | label= ULAN - Union List of Artist Names (see above, under dc:creator) |
| dc:type | label= PICO Type Vocabulary<br>definition= Controlled vocabulary which includes some resource types specifically conceived for the Italian Culture Portal domain: Corporate Body, Physical Person, Project. Those types are not foreseen by the DCMI Type Vocabulary.<br>type= encoding scheme |
| | label= CDType - Collection Description Type Vocabulary<br>definedBy= http://purl.org/cld/terms/<br>definition= A list of types that categorize a collection.<br>comments= Reference at: http://www.ukoln.ac.uk/metadata/dcmi/collection-application-profile/#cldCLDT<br>type= encoding scheme |
| dc:format | label= Material And Technique<br>definition= The material of the object and of its support and the technique of execution of a resource with type: PhysicalObject<br>type= element-refinement |
| dc:identifier | label= ISBN - International Standard Book Number<br>definition= The International Standard Book Number is an uniform and persistent identifier for a given title or for the edition of a title pertaining to a given publisher. Reference at: http://www.isbn.it/.<br>comments= It is generally used for resources with type: Text.<br>type= encoding scheme |
| | label= ISSN - International Standard Serial Number<br>definition= The International Standard Serial Number is the international identifier for serial publications such as printed or digital newspapers and periodicals. Reference at: http://www.issn.org/.<br>comments= It is generally used for resources with type: Text.<br>type= encoding scheme |
| dc:relation | label= Preview<br>definition= Any form of abstract, reduction, image, video streaming used as anticipation of the resource.<br>type= element-refinement |
| | label= Promotes<br>definition= The described resource promotes and/or organizes the referenced resource.<br>type= element-refinement |
| | label= is Promoted By<br>definition= The described resource is promoted and/or organized by the referenced resource.<br>type= element-refinement |
| | label= Manages<br>definition= The described resource manages with different responsibilities (scientific, administrative, technical, etc.) the referenced resource.<br>type= element-refinement |
| | label= Is Managed By<br>definition= The described resource is managed with different responsibilities (scientific, administrative, technical, etc.) by the referenced resource.<br>type= element-refinement |
| | label= Is Owner Of<br>definition= The described resource owns the referenced resource.<br>type= element-refinement |
| | label= Is Owned By<br>definition= The described resource is owned by the referenced resource.<br>type= element-refinement |
| | label= Produces<br>definition= The described resource produces in its physical, or administrative, or any other issue, the referenced resource. |

| DC ELEMENTS | PICO AP QUALIFIERS |
|---|---|
| | comments= It is recommended to express value as URI. This relation should be used when it is possible to point at another resource, pertaining to the work produced by the described resource.<br>type= element-refinement |
| | label= Is Produced By<br>definition= The described resource is produced in its physical, or administrative, or any other issue, by the referenced resource.<br>comments= It is recommended to express value as URI. This relation should be used when it is possible to point at another resource pertaining to who produced the described resource; otherwise it is recommended to use Producer.<br>type= element-refinement |
| | label= Performs<br>definition= The described resource performs, directly participating (e.g. as actor or musician) to, the referenced resource.<br>comments= It is recommended to express value as URI. This relation should be used when it is possible to point at another resource pertaining to the work performed by the described resource.<br>type= element-refinement |
| | label= Is Perfomed By<br>definition= The described resource is performed by the referenced resource.<br>comments= It is recommended to express value as URI. This relation should be used when it is possible to point to another resource, pertaining to who performs the described resource; otherwise it is recommended to use Performer.<br>type= element-refinement |
| | label= Is Responsible For<br>definition= The described resource is anyhow responsible for, or is the contact person of, the referenced resource.<br>comments= It is recommended to express value as URI. This relation should be used when it is possible to point to another resource, the described resource is responsible for.<br>type= element-refinement. |
| | label= Has As Responsible<br>definition= The described resource has as responsible and/or contact person the referenced resource.<br>comments= It is recommended to express value as URI. This relation should be used when it is possible to point at another resource, pertaining to who is responsible for the described resource. Otherwise it is recommended to use Responsible.<br>type= element-refinement |
| | label= Contributes To<br>definition= The described resource contributes anyhow to the referenced resource.<br>comments= It is recommended to express value as URI. This relation should be used when it is possible to point at another resource, pertaining to something/someone that receives contributions from the described resource.<br>type= element-refinement |
| | label= Has As Contributor<br>definition= The described resource is produced, managed, organized with the contribution of the referenced resource.<br>comments= It is recommended to express value as URI. This relation should be used when it is possible to point at another resource, pertaining to something/someone that is giving contributions to the described resource. Otherwise it is recommended to use Contributor.<br>type= element-refinement |
| | label= Digitises<br>definition= The described resource is responsible of the digitisation of the referenced resource.<br>comments= It is recommended to express the value as URI. This relation is generally used for resources with type: Physical Person or Corporate Body.<br>type= element-refinement |
| | label= Is Digitised By |

| DC ELEMENTS | PICO AP QUALIFIERS |
|---|---|
| | definition= The described resource is digitised by the referenced resource.<br>type= element-refinement |
| | label= Anchor<br>definition= Reference to the URL of the web-page publishing the resource described by the metadata record.<br>comments= It is recommended to use DCSV syntax as follows: title= e.g. Website of the Scuola Normale Superiore of Pisa; URL=http://www.sns.it<br>type= encoding scheme |
| dc:coverage | label=Date of Birth<br>definition= Date of Birth pertaining to resources with type: Physical Person.<br>type=element-refinement |
| | label= Date of Death<br>definition= Date of Death pertaining to resources with type: Physical Person.<br>type= element-refinement |
| | label= Place of Birth<br>definition= Place of Birth pertaining to resources with type: Physical Person.<br>type= element-refinement |
| | label= Place of Death<br>definition= Place of Death pertaining to resources with type: Physical Person.<br>type= element-refinement |
| | label= ISTAT Code<br>definition= Code assigned by Istituto Nazionale di Statistica italiano (Italian National Institute for Statistics), which identifies inhabited places in the Italian territory. Reference at http://www.istat.it/strumenti/definizioni/comuni/<br>comments= ISTAT code must be composed by 8 numbers: first 2 identify the Region; following 3 identify Province; final 3 identify the City (o smaller inhabited place) within the Province.<br>type=encoding scheme |
| | label= Postal Address<br>definition= Postal address of a resource with type: Physical Object or Corporate Body. It is expressed with the DCSV syntax as specified in the following example: PlaceType=Via /piazza / Largo, etc.; PlaceName=Dante; PlaceNumber=26; ZipCode - CAP=57124; City=Roma; Province=RM; Region=Lazio; Country=Italia.<br>type=encoding scheme |

# 6 User Interface

*CulturaItalia* will publish different kinds of contents:

- static contents: Head and logo, access to multilingual versions, credits, contact information, mission, site map, copyright;
- dynamic contents, from CMS: news, itineraries, focus, press release, forum, FAQ, newsletter, specific areas (e.g.: young users);
- dynamic contents, from harvesting: metadata harvested from external repositories;
- business logic contents, depending on the user session: search results, bookmarks, etc.;
- user inputs: layout personalization controls (font, contrast, colour), registration area to access in a private area to save bookmarks, annotate events in agenda, etc.

The interface will allow data retrieval on those contents trough different possibilities for searching and browsing. User will access contents through three kinds of searches:

- free search: user composes one or more words, using Boolean syntax;
- advanced search: user refines the query in the catalogue, selecting if the item to be retrieved is “place”, “person”, “event”, or “object”;
- geographic search, selecting a place on a list or on a map related to a GIS system.

It will be possible to browse the catalogue through the Main Menu or the Theme Menu. According to the 4 High Level Elements of DC Culture, defined by Aquarel project and approved by MINERVA project, the Main Menu of the catalogue is structured in:

- Who: people, institutions, administration offices, museums, archives, libraries, universities, etc.;
- What: art objects, monuments, documents, books, photos, movies, records, theatre and music productions, etc.;
- When: contents retrievable trough temporal periods;
- Where: browse by region, province, town, on a controlled list or directly on a GIS.

User will browse the catalogue using a 'facettes' system: he can start the query from one of the four elements and further refine the results range. A simplified alternative for browsing is the Themes menu. It groups the resources according to the following arguments: Archaeology, Architecture, Visual Arts, Environment and Landscape, Cinema and Media, Music, Entertainment, Traditions, Humanities, Scientific Culture, Education and Research, Libraries, Literature, Archives, Museums, Exhibitions.

The Portal will not publish only resources harvested from external repositories, but will produce also new contents: an editorial office will prepare and manage contents to provide interesting relations between resources and make the user discover them through links among different kinds of information. Those new contents will be tailored on different users' targets and will be distributed into the following sections of the Portal:

- Itineraries: articles focused on a theme, aimed at suggesting a virtual tour through some resources selected from the catalogue;
- Focuses: short monographs on a single argument;
- Events: information on cultural events (exhibitions, concerts, theatre, conferences);
- News: selected news on Italian culture.

Finally the project recommends that the Portal would provide the following services, to be eventually implemented in a later phase after the first realization:

- Multilingual versions;
- Newsletter;
- Forum;
- Young users area.

## 7 Recommendations for Usability and Accessibility

The project for the Italian Culture Portal deals with recommendations both for usability and for accessibility by impaired people. The Portal, which will be maintained by a public institution, must be usable by impaired people, e.g. by people with visual, auditive, motion and cognitive deficit, through the use of assistive devices and technologies.

In Italy such recommendation is ruled by the law n. 4 issued on 2004/01/09, "Disposizioni per favorire l'accesso dei soggetti disabili agli strumenti informatici" [12] (Recommendations for favouring access to IT tools by impaired people). This law imposes specific obligations both for the purchase of goods and for the providing of services, even with the possibility of making the service void.

The abovementioned law refers to the Italian Constitution (art. 3 "Every citizen has equal social dignity and is equal for the law, without distinctions of […] personal and social conditions. It is duty of the Republic to remove economic and social impediments which […] prevent the full development of the human being and the complete participation to the political, economical and social organization of the State.") and to norms issued by the Ministry of Public Administration (Ministero della Funzione Pubblica) and by the AIPA – Authority for IT in Public Administration, often ignored.

The law, together with recommendations for technical requirements for hardware (e.g. keyboards, devices for remote control) and software (e.g. user interface, maintenance of set up defined by users, textual information, buttons for accessing the assistive devices, etc.) rules also how web sites must be created (and tested) for guaranteeing accessibility.

Concerning web sites accessibility, the law is based on international recommendations as chapter 1194.21 of Section 508 of the USA Rehabilitation Act [13], and on guidelines provided by international bodies such as World Wide Web Consortium (W3C [14]) and, in particular, the recommendations of Web Accessibility Initiative (WAI [15]).

Moreover the project for the Italian Culture Portal adopts the guidelines proposed by MINERVA EU Project, concerning requirements and testing methodologies for the creation of "good quality web sites", to be not only technically accessible but also completely usable. The handbook for "Quality in cultural web sites[1]" recommends a technical test for accessibility to contents and a subjective test for the usability of information and services.

The minimum level of accessibility for all users (including people with complete or partial visual disabilities) takes into consideration what appears in the browser window, for technical aspects and for the contents, and imposes (amongst others) the following requirements:

- It is mandatory to adopt a DTD Strict and to use XHTML. Such recommendation implies to separate content from layout, and forbids to open new windows within the present one. Such an issue imposes some specific constraints specially when pieces of contents are imported from other web sites (e.g. through RSS feed);
- The use of frames must be avoided;
- Every non textual object must have an equivalent textual alternative. Therefore, images, audio and video streamings must be integrated with a text (from simple captions to a complete synchronized under-titling) in order to make assistive devices able to read all the objects of the page;
- Sensible maps must be client side or, if not possible, they must be linked to textual alternatives;
- It must be possible to easily distinguish main information from the background, both for graphic or audio components. This remarkably impacts on the use of colours and backgrounds;
- Layout and contents must be resizable, without overlapping or loss of information;
- Table-based layouts should be avoided; it is recommended to adopt a CSS based layout, using the element <div>
- Tables of data must be provided with information to be correctly interpreted by assistive devices, such as screen readers. Forms too must be created taking into consideration that they could cause problems when read by assistive devices;
- Pages must be usable even when scripts and applets are disabled or not supported;
- Links must be understandable even if read out of their context. User must be able to click them even through keyboard commands, technologies of keyboard emulation and pointing devices alternative to the mouse. This implies constraints of their position in the page, as they must not be too close each other.

Evaluation procedures are based both on automatic and semiautomatic validation systems and on the analysis carried on by an expert in web technologies and accessibility. The project suggests to evaluate the web site with the cooperation of a user panel, including impaired users, according to the following (subjective) quality criteria:

- perception
- comprehensiveness
- efficiency
- consistency
- safety
- security
- transparency
- easiness of learning system functionalities
- availability of helps and documentation
- tolerance to errors
- look and feel
- flexibility

Such criteria must be taken into consideration both during the design of the web site and in the development of the interface after this first evaluation.

[1] http://www.minervaeurope.org/publications/qualitycriteria.htm/

## Notes and References

[1] Project responsibles for MiBAC: A. P. Recchia, R. Caffo. Project responsible for SNS: S. Settis. Coordinators: B. Benedetti, U. Parrini. Working group: P. Baccalario, I. Buonazia, M. Delcaldo, M. E. Masci, D. Merlitti. Consultants: G. Cresci, O. Signore, P. Valentino.

[2] PICO AP – Portal of Italian Culture Online - Application Profile.

[3] PURL - Persistent Uniform Resource Locator: http://purl.oclc.org/

[4] MiBAC website: http://www.beniculturali.it/

[5] OAI-PMH – Open Archive Iniziative – Protocol for Metadata Harvesting: http://www.openarchives.org/OAI/openarchivesprotocol.html/

[6] DCMI - Dublin Core Metadata Initiative: http://dublincore.org/

[7] For DCES – Dublin Core Element Set, see: http://dublincore.org/documents/dces/

[8] DC elements and terms are defined in: http://dublincore.org/documents/dcmi-terms/. See also *DC Metadata Registry*: http://dublincore.org/dcregistry/. For Qualified DC, see: *Using DC Qualifiers*: http://dublincore.org/documents/usageguide/qualifiers.shtml/; *Expressing Qualified DC in RDF/XML*: http://dublincore.org/documents/dcq-rdf-XML/

[9] Definition of 'Application Profile' from the DCMI Glossary: "In DCMI usage, an application profile is a declaration of the metadata terms an organization, information resource, application, or user community uses in its metadata. In a broader sense, it includes the set of metadata elements, policies, and guidelines defined for a particular application or implementation. The elements may be from one or more element sets, thus allowing a given application to meet its functional requirements by using metadata elements from several element sets including locally defined sets. For example, a given application might choose a specific subset of the Dublin Core elements that meets its needs, or may include elements from the Dublin Core, another element set, and several locally defined elements, all combined in a single schema. An application profile is not considered complete without documentation that defines the policies and best practices appropriate to the application". See: http://dublincore.org/documents/usageguide/glossary.shtml

[10] This document is downloadable at: ftp://ftp.cenorm.be/PUBLIC/CWAs/e-Europe/MMI-DC/cwa14855-00-2003-Nov.pdf.

[11] See: *DCMI Type Vocabulary*: http://dublincore.org/documents/dcmi-type-vocabulary/

[12] Disposizioni per favorire l'accesso dei soggetti disabili agli strumenti informatici - http://www.pubbliaccesso.gov.it/normative/legge_20040109_n4.htm

[13] See: http://www.pubbliaccesso.gov.it/normative/rehabilitation_act/index.htm

[14] W3C website: http://www.w3.org/

[15] WAI website: http://www.w3c.it/wai/

# Building Bridges with Blocks: Assisting Digital Library and Virtual Learning Environment Integration through Reusable Middleware

*Santiago Chumbe*[1]*; Roddy MacLeod*[2]*; Marion Kennedy*[2]

[1] Institute for Computer Based Learning, School of Mathematical and Computer Sciences
Heriot Watt University, Edinburgh, EH14 4AS, UK
e-mail: santiago@macs.hw.ac.uk
[2] Heriot-Watt University Library
Heriot Watt University, Edinburgh, EH14 4AS, UK
e-mail: r.a.macleod@hw.ac.uk; m.l.kennedy@hw.ac.uk

## Abstract

Various studies have reported that achieving effective use of increasingly heterogeneous scholarly objects within institutional learning and teaching frameworks is becoming critical to the performance of educational institutions. The integration of digital information environments, such as a University library, within a virtual learning environment (VLE) encapsulates this challenge. This paper presents reusable middleware to achieve effective digital library (DL) and VLE integration. The aim of the study is to demonstrate that the use of open standards and service-oriented architectures (SOA) to build "light" web-services-based middleware is a suitable alternative for embedding digital library information sources in learning and teaching frameworks. We argue that by using open-source and open-standards approaches rather than software and practices developed specifically for a particular VLE product, it is possible to obtain open reusable middleware that can simplify the DL-VLE integration and bridge the functionality of both environments. We hope that our methodology can provide a common foundation on which a variety of institutions may build their own customized middleware to integrate scholarly objects in VLEs. The study has assessed the impact of the VLE-library integration on academic users of both the library and the VLE. Performance issues of the proposed digital library-VLE integration are also discussed. A secondary but important finding of our study is that much more effort is required to open and standardize the closed, restricted and proprietary approach of digital publishers to the reuse of scholarly material. This approach can be a serious obstacle to effective digital library-VLE integration and can limit the publishers' ability to allow the discovery, integration and reuse of scholarly material. Current research in this area is analyzed and discussed.

**Keywords:** reusable middleware; SOA; SRU/SRW; federated search; VLE; open standards

## 1 Introduction

It is becoming clear from a number of perspectives that allowing effective discovery and use of scholarly objects within learning and teaching frameworks such as VLEs and institutional portals will be critical to the performance of educational institutions [1-7]. However, as Low B. [8] has noticed, resource discovery has been overlooked as a function of VLEs by vendors. We believe that this deficiency needs to be addressed urgently and with an "open standard" perspective. Digital libraries (DL) and VLEs both support learning and teaching in academic institutions. Institutions use library management system or digital libraries (DLs) to gain access to the content of scholarly objects from local databases such as institutional repositories or other collections of research papers, e-theses, technical reports, OPACs, image banks, etc., as well as from subscribed external content such as scientific papers provided by journal publishers or aggregators, and remote digital libraries, directories and online databases. On the other hand, VLEs are integrated environments of components (e.g. online discussions, course materials, e-mailing communication, submission of assignments, assessment, etc.) in which learners and tutors participate in "online" interactions of various kinds, involving online learning (VLEs are also known as Learning Management Systems (LMS) outside the UK.) However, despite the fact that both DLs and VLEs are oriented to support learning and teaching, previous studies have reported that the process of integrating DL and VLE systems can raise technical issues that require in depth investigation and complex solutions [9-11] (non-technical but important issues are beyond the scope of this study.) For example, systems run on different operating systems, use different data formats, have different authentication requirements and different web interfaces, etc. This paper sheds some light on a cost-effective methodology for overcoming such technical issues and confirms that a service-oriented approach combined with web services technology that makes use of standards or specifications for interoperability is a simple solution for achieving effective DL-VLE integration.

This paper first briefly describes the PerX toolkit, an open-source federated search software application produced by the *Pilot Engineering Repositories Xsearch* (PerX) Project [12]. It then presents the web services-based and open shareable SOA-compliant middleware used to embed Library functionality within the VLE. When describing the work done for encapsulating the middleware in the VLE used in this study, the paper mentions the commercial VLE *Blackboard* platform [13]. However, the work presented is not dependent on *Blackboard,* because it uses open standards and open source. Thus, our work can provide a common foundation on which a variety of institutions may build their own customized middleware to integrate their scholarly objects in their own VLEs. The only requirement is that the VLEs support XML-based retrieval via HTTP, preferably using standard web services communication. Further information and discussion on the middleware implementation and details of the "*Building Block*" encapsulation is presented in section three.

We also discuss the current status of digital publishing with respect to DL-VLE integration, finding that, within this context, most digital publishers have adopted a closed, restricted and non-standard approach. Publishers of scientific papers are one of the main sources of DL content and their lack of participation in sharing and reusing of scholarly metadata via open standard mechanisms can have a negative impact on DL-VLE integration success. Some recommendations for increasing interoperability and reuse in digital publishing are outlined at the end of section four.

The study has assessed the impact of the proposed VLE-library integration on academic users of the VLE and library services. Use case scenarios highlighting experiences gained and implications for stakeholders arising from the pilot are described in section five. The outcomes of these experiences are used as a basis for recommendations for future development of the pilot as well as for institutions planning to integrate their library with institutional VLEs.

After a discussion of the implications and some performance issues of the proposed digital library-VLE integration, the paper ends with conclusions obtained from the study.

## 2 The PerX Federated Search Toolkit

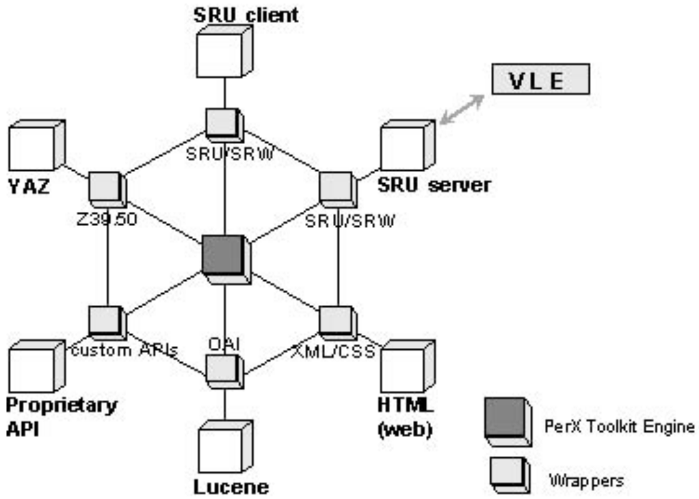


**Figure 1: The PerX Toolkit Architecture**

The core software component of this pilot is an open-source federated search toolkit produced by the *Pilot Engineering Repositories Xsearch* (PerX) Project, funded by the JISC Digital Repositories Programme. We chose this federated search software (referred to as the PerX toolkit) because it uses both an XML-based technology for system integration and a service-oriented architecture (SOA) [14] for achieving greater "loose" separation between its software components. In fact, the PerX toolkit is a reusable library of open source software applications integrated by a SOA model. It is a loosely coupled collection of proven, scalable and reusable software libraries and APIs (Application Program Interfaces) that have been combined via XML messages. Figure 1 represents the PerX toolkit architecture. Its main component is the *PerX Toolkit Engine*, which communicates with the rest of the software components via *wrappers*. The *wrappers* use XML-messaging for handling requests from/to the reusable APIs, which in turn deal with the database sources. The toolkit allows remote and local heterogeneous database sources to be cross-searched from one access point. It uses open standard technology for metadata exchange such as OAI-PMH (Open Archive Initiative-Protocol for Metadata Harvesting, [15]) and the search protocols SRU (Search/Retrieve via URL, [16]) and Z39.50 (International Information Retrieval Standard ISO 23950, [17].) Chumbe et al [18] presents a full description of searching databases with the PerX Toolkit as well as the processes involved in metadata exchange with data providers (harvesting, normalization, searching and rendering.)

# 3 Reusable Middleware Approach for Embedding DL Functionality within a VLE

Heriot-Watt University Library has recently collaborated with the Institute for Computer Based Learning (ICBL) on a *Blackboard* VLE - e-Library integration pilot. The core software component of this pilot is the PerX toolkit described in the previous section. However the key player, or broker, of the integration itself is the reusable web-services based and SOA compliant middleware used to embed the toolkit functionality within the VLE. An important condition for the pilot was that the middleware should know nothing about the hosting VLE environment and thus can potentially be reused within any VLE framework. Its only function was to provide a "live bridge" between the toolkit functionality and the VLE system.

Traditional client/server middleware has typically been deployed in a 2-tier, point-to-point architecture [19], which in our case would involve the installation of a proprietary API Client on the VLE server, and an API server on the PerX server machines. However, this is an expensive and inflexible model, because neither of the Client and Server APIs can be reused. A step forward in flexibility is an n-tier model, where an XML-based middleware API is installed between the client (VLE) and the server (DL) systems. The n-tier model offers the benefit of avoiding the development of two different APIs and the need to access source codes on both sides to enable interoperability. The XML-based middleware API is a kind of *wrapper* that hides the complexity of the native APIs of both server and client because it uses web services technology for exposing their services. This XML-based n-tier model has been used by the LEBONED Project to integrate the *eVerlage* Digital Library product into the *Blackboard* VLE [20]. However, the cost-effective factor is still unresolved by this approach, because such a specific *wrapper* will need to be written again for any other digital library system to be integrated into *Blackboard*. We present here a further step towards achieving inexpensive, reusable and flexible DL-VLE integration. Our approach is also based on a three-tier design pattern using an XML-based middleware API that sits between the *Blackboard* VLE (front end) and the PerX toolkit (back end) systems, but we do use the open standard SRU/SRW protocol for interoperability and XML message exchanging between the systems. The use of proven open standards effectively turns our XML-based middleware into a reusable *wrapper* or message broker. This approach would allow organizations to access virtually any SRU/SRW compliant system from within any SRU/SRW compliant DL system through a scalable service-oriented architecture. While simple, the middleware constitutes the basic infrastructure behind the DL-VLE integration, becoming a versatile alternative for integration. The basic deployment architecture of the proposed approach is shown in Figure 2.

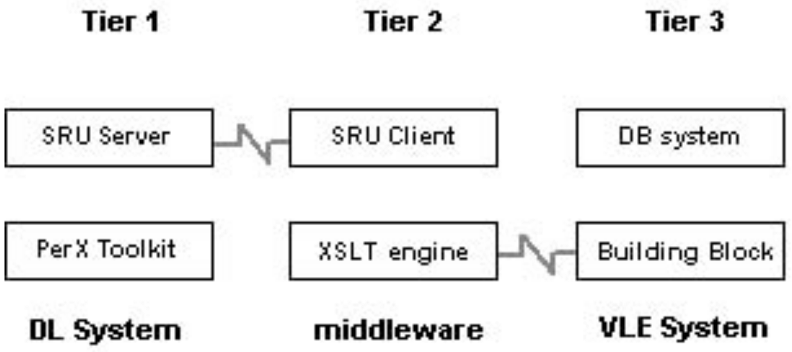


**Figure 2: DL-VLE integration architecture using a three-tier design model**

The approach described above offers a potentially rich system-level DL-VLE integration because it uses a standard specification for interoperability such as the Search/Retrieve via URL (SRU) protocol encapsulated in a reusable middleware. The SRU/SRW protocol is simple to implement because it is a standard REST-ful specification for providing Web Services functionality without the complexity of tightly coupled designs as found in remote procedure calls such as SOAP [21]. A REST-based protocol uses the HTTP mechanism to implement a client/server model using TCP/IP sockets [22]. The encapsulated middleware (the HTTP client) opens a connection to the PerX toolkit's SRU server (the HTTP server) and sends a request message consisting of a search query using the HTTP GET method. The HTTP server then returns a response XML message with the search results using the POST method and then closes the connection. The middleware then reformats the

XML message and puts the search results into the VLE database system, so they can be shareable in the VLE modules. The middleware is a kind of proxy or intermediary software that handles requests on behalf of the systems that it is bridging.

In order that the middleware be recognized by the VLE as one of its components, it needs to be encapsulated in a *building block* of the *Blackboard Learning System*. This is accomplished by issuing an XML configuration file (*manifest*) to identify the middleware as a “bridge type” *Blackboard* "*building block*," and by including *Blackboard* proprietary Java class Tag libraries to abstract user interface components [23]. The middleware implementation has followed as strictly as possible the current Java Servlet specifications for Web applications [24]. The *Blackboard* system includes a portal running on a *Tomcat servlet* [25] and in fact its “*building blocks*” are just local *portlets* that can be handled as web applications individually deployed on the local *servlet*. These *building blocks* do not adhere to the web services specifications for remote *portlets* (known as Web Services for Remote Portlets WSRP specification [26]), so they are not shareable from other portals or remote systems. However, this is not an issue for our implementation as we supply the share-ability via the REST-ful model described above. At its very core, the middleware is an SRU client that provides standards-based technology to achieve integrated behavior and performance at the system-level across diverse environments such as the federated search toolkit and the VLE system.

Unlike other open standards for interoperability, such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), implementations of the SRU/SRW protocol are poorly reported in the digital library literature. Equally, while OAI-PMH has attracted much attention in providing interoperability, despite reports of a number of important issues concerning OAI-PMH in the literature [18, 27-29], practical examples of SRU and its relevance in the digital library are seldom discussed. To the best of our knowledge the approach of combining SOA with SRU to embed DL functionality within a VLE for cross-searching remote databases and local repositories has not yet been reported fully, and practical implementations have not received much attention. An attempt to integrate VLEs and digital repository systems using the SRU protocol in the open source d+ toolkit has been reported by Low B. [8]. However, some issues with the software were uncovered when using d+ for interoperability with VLEs. Despite claims of adherence to the current service-oriented trend, it was found that, apart from the SRU functionality provided by licensed OCLC software, deployment and use of the d+ toolkit required hardcode configuration of the software components as well as of the digital repositories. Also, at the time of testing, d+ could only query one database at the time (a sequential searching approach in contrast to the desirable "simultaneous" cross-searching approach.) Performance issues were also noted. It seemed that the ability of d+ for searching Z39.50 targets was bound to the limitations of the JAFER toolkit [30], which is still not fully available for production. Other open source alternatives considered before PerX and found unsuitable for the work discussed here, were the MDC toolkit, MyLibrary and the software suite Greenstone. In the UK, the JISC – DiVLE research strand involved a number of projects looking at how library resources can be integrated into VLEs using open standards. Thus, for example, between 2002 and 2004, the OLIVE project has been focused on how the OpenURL standard can be used to link Reading lists and Learning objects from the VLE. It also explored the use of Web Services (SOAP.) However, little practical achievement was reported [31], and unfortunately most of the plug-in software developed by the project was dependent on the commercial platforms used for integration (*MetaLib*, *Blackboard*, *Aleph*, *Discover*, *SFX*, etc.) For example, the method for implementing OpenURL is tightly coupled to the search form in the *Building Block* and cannot be reused for other applications. Also, the approach of the OLIVE project of loading functionality on the *Building Block* for metadata management raised many interoperability issues, as the *Blackboard* metadata functionality proved to be unusable and inaccessible to other areas within *Blackboard*. In Australia, Richardson J. [32] also reported on a project at Griffith University to integrate library resources into the *Blackboard* system. She recognized the power of commercial products in this arena, such as Sentient *Discover*, which supports OpenURL and Z39.50, but also highlights the “cognitive disconnect” faced by users of *Blackboard* when are taken away to the *Discover* user-interface environment from the *Blackboard* user-interface environment.

On the other hand, SOA approaches in e-learning are being promoted as suitable alternatives by important organizations such as JISC (the United Kingdom’s Joint Information Systems Committee), DEST (the Australian Department of Education Science and Training), ADL (the US Advanced Distributed Learning Initiatives), IMS (the Innovation Adoption Learning global learning consortium), NSDL (the US National Science Digital Library) and IC (Industry Canada). Need for stable and coherent technical frameworks or infrastructures where e-learning services can inter-operate harmoniously have been highlighted [33-36]. Our work is firmly in harmony with the above approach and recommendations, and it would be part of any standard e-learning framework where its functional components expose service behavior via loosely coupled interfaces. In this context, we follow with interest the work being carried out by related projects, such as the open source digital library architecture Fedora [37] and the NSDL Data Repository Architecture [38], as well as any research outcome from the JISC e-Framework for Education and Research Programme [39].

## 4 Issues of the Digital Publishing Model Regarding Reuse of Scholarly Material

An additional finding of our study is that integration of digital publishing is made difficult by the fact that publishers rarely use open standards to make their metadata available to third parties. Many publishers currently rely on large external aggregators in order to expose their scholarly contents to a wider audience. Frequently, digital libraries need to deal with these external aggregators in order to gain access to subscribed scholarly material using expensive commercial software tools, which in most cases do not use open standards. The consequence is that it is often difficult for institutions to get access to publishers' metadata and databases using suitable open standards and protocols for interoperability. The reality is that progress towards integration of scholarly digital information within VLEs is slowed down by commercial publishers and aggregators by not offering machine-to-machine access to their databases using open standards. Figure 3 illustrates a simplified view of a typical digital publishing model within the digital library context. Clearly noticeable is a need for a "consolidator" point for effective inter-operation between digital library and the rest of components of the model. Integration in a component-by-component basis would be unfeasible. Figure 4 sketches an alternative model where effective integration is enable by a suitable "middleware consolidator" created using technology presented in previous section. Advocacy for open standards is not about encouraging free access to resources but simply about providing effective ways to find (discover) and reuse resources. The PerX Project has produced a relevant report on the benefits for publishers of exposing their metadata via open standards [40]. Also other works [41, 42] advocate the use of best practices among data providers and argue that the business strategies of digital publishing in fact can benefit from the standards that are part of the digital library.

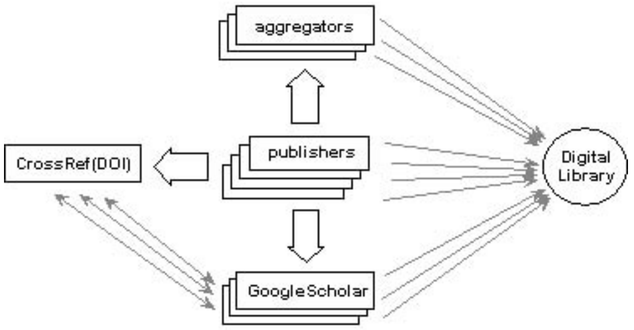


*Figure 3* A typical digital publishing model.

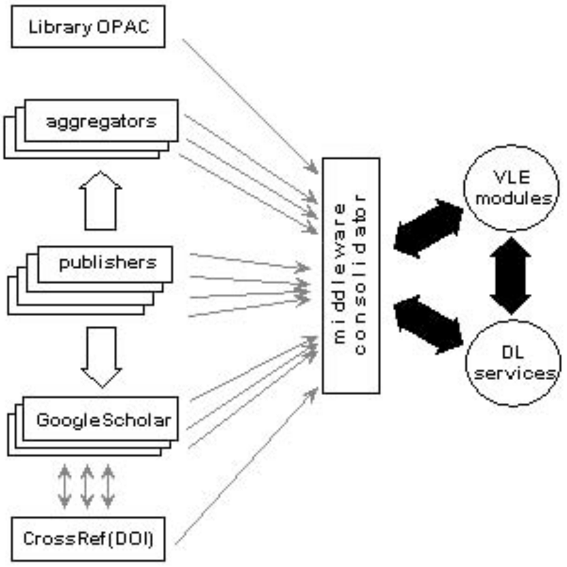


**Figure 4: Publishing model with "middleware consolidator"**

Currently there is a large movement towards openness in almost all aspects of digital publishing. Promising initiatives for solving important interoperability issues are not only coming from organizations that advocate open standards. Thus two technologies for enabling easier scholarly resource discovery have emerged, one from the publisher's side (CrossRef) and another one, Google Scholar, from Google, the leading commercial search engine. (Microsoft Windows Live Academic Search and Scirus services could also be mentioned here.) CrossRef is being promoted by the publishing industry to make possible standard scalable linkage of scholarly material through Digital Object Identifiers (DOI) [43]. We have been investigating the feasibility of using the CrossRef OAI service to cross-search metadata for a selection of the 23 million records hosted by CrossRef as well as to provide openURL linkage via the CrossRef openURL resolver. Unfortunately access to the CrossRef OAI repository is not open to everyone, which again puts a limitation to the reuse of metadata. Also, the CrossRef OAI service uses a somewhat limited subject classification that makes subject-based implementations difficult. The usefulness of being able to only search on title, authors and citation of papers could be also challenged (CrossRef does not store abstract and keywords.) Google Scholar [44] via direct agreements with publishers is in fact crawling and cross-searching a very important and increasing portion of scholarly resources (e.g. peer-review articles, theses, preprints, technical reports, etc.) Google Scholar also works with CrossRef to use the DOI as the primary means to link to an article. Despite the facto that Google Scholar is still a very broad commercial oriented solution that includes any material that "looks scholarly" and that can come from unknown sources, it offers something more than other open access federated search services such as the DOAJ (Directory of Open Access Journals [45]) or Scirus [46]. It offers enhanced and fast search capabilities, cited references and links to subscribed resources through local link resolvers. On top of that, Google Scholar has the advantage of being quickly associated with the de facto ubiquitous discovery tool: Google is everywhere. As the value and usage of Google Scholar is significant, we have integrated Google Scholar in the DL-VLE prototype, via a custom API. Some researchers think that Google Scholar could be the possible solution to the cluttered access provided by traditional gateway and hosting providers used by the library. However, Google is still a commercial initiative, produced without open standards and developed in a way in which not everyone can participate, which raises concerns over obsolescence and dependency issues. For example, experience suggests that only file formats that use open standards can secure long-term preservation of scholarly material and avoid software dependencies.

In fact, the publisher with relatively cost-effective and simple solutions can produce chunk-able, reuse-able and embed-able metadata using open standards. The technical undertaking by publishers need not be large, though the potential benefits are. Publishers are already using complex and tailored mechanism to expose their data on demand basis. This approach can be ineffective and expensive. Let us consider, for example, the case of a publisher that wants its metadata to be included in CrossRef, and also be available (optionally the full-text, too) from various aggregators (MetaPress, ingentaConnect, Ovid, ProQuest, SwetsWise, etc.) and indexed by Google Scholar. Without standards, the publisher will need to set up and maintain different XML metadata files for Google, for CrossRef and for each aggregator. It will need to use an FTP-based mechanism for uploading data on the aggregators or allow them to crawl their servers hosting its data. All that could be avoided if the publishing industry agree on using a set of open standards, and better still if they work with librarians for enabling easy resource discovery, as both of them share the same goal: to make scholarly content available for the users that need it. The benefits of making online search a pleasing experience are for both the publishers and the digital libraries. We suggest that publishers start by implementing “light” open standards such as SRU/SRW, openURL, RSS feeds, and Dublin-Core (DC) metadata format. It is worth it for publishers to consider redirecting some of their IT resources to implementing open standards, automatic machine-to-machine access and simplified user interfaces. Diverse studies have already suggested that what online users want is fast and effortless access to the resources they need [47-50]. Users give little value to sophisticated user interfaces provided by publishers’ web sites. Publishers should take notice of the behavior of users. On the other hand, commitment to protocols or specifications that do not adhere in full to the open standard concept [51] should be avoided if possible, in case that “cutting edge” technology that is not backed by mature open standard bodies is abandoned. For example it can be instructive to follow the discussions on the reasons for the apparent decline in the use of the CORBA protocol [52], which has been providing interoperability for more than a decade. In summary, it would make a positive impact on interoperability in general, and possibly in their revenues too, if publishers implement open standards for enabling institutional and individual users to gain quick access to the content they need with almost no effort.

## 5 Study of the Impact of DL-VLE Integration on Library Users

A prototype working system demonstrating the VLE-Library system interoperability has been implemented and made available to stakeholders (students, academic and library staff) at the Heriot Watt University. It is being used to asses the impact of the VLE-library integration on academic and library users as well as a basis for gathering suggestions and recommendations for future developments to benefit institutional planning for library and institutional VLE integration. The prototype system, named as *PerX Building Block*, provides distributed searching of a sample of subscribed e-journals, the local library catalogue (OPAC) and the Google search engine. A facility for bibliographic export in RDF-based format is being added in the prototype. Testing is being carried out with a group of academic library users, and feedback is being gathered using a short questionnaire and informal interview. So far our study has confirmed the perception that in particular under-graduated students tend to ignore searching in databases subscribed by the library and prefer the ease of using Google [53]. Post-graduate researchers also feel attracted to Google capabilities. The current searchable web based interface of the library does not include links to Google or Google Scholar. If even lecturers and librarians use Google in their work, we expect that users will appreciate having Google Scholar embedded in the prototype. In fact Google Scholar can be used to drive users to the library web site and add value to the sometimes ignored library catalogues at not cost.

The reuse and sharing of DL content among the different VLE components is being explored with particular interest. We have had high interest in finding out how users rate the usefulness of cross searching from within the VLE and the convenience of onward use of search results in other VLE functions e.g. exporting, saving, emailing and posting them to discussion boards.

In parallel to our work, the University VLE Educational Support Team has been conducting consultation meetings with lecturers who are using the VLE in their courses to give them the opportunity to bring up any problems and provide feedback. In some of the meetings various issues were mentioned by lecturers that in fact would be solved by enabling machine-to-machine inter-operability between the VLE and the rest of University systems. This prompted the possibility of expanding the applicability of the reusable middleware for *bridging* systems such as the Students Registration System with the VLE.

Regarding possible performance issues of the proposed DL-VLE integration, we have noticed that SRU/SRW is not necessary relatively slow. We were expecting that the SOA-based prototype be significantly slower than fast, general purpose search engines because it uses XML-based messaging services, which typically consume more computing resources. However, after assessing the performance of the search services when searching various heterogeneous scholarly objects, users noticed that speed and performance were not issues in the prototype.

Finally, some recommendations for increasing the usability and the effectiveness of the prototype have been identified. In addition to more sophisticated retrieval and searching algorithms (e.g. full common Boolean support across heterogeneous databases), there are key operational enhancements that have been acknowledged as desirables. Enhancements include:

- Combining search results from multiple databases, which involves unified ranking;
- Comparing and consolidating search results (simplest case: removing duplicate search results; more complex case: fussy techniques for combining several databases´ results);
- Discovering inconsistencies and removing them in the search results (for example search results that seems to be different but in fact point to same resources).

# 6 Conclusions

By using open-source and open-standards approaches rather than products and practices developed specifically for an individual VLE product, we have obtained a reusable middleware that can provide a common foundation on which a variety of institutions may build their own customized middleware to integrate their scholarly objects in VLEs. Our study hopes to demonstrate that the use of service-oriented architectures (SOA) and REST-ful based (SRU) open source middleware is a cost effective, simple and open alternative for embedding digital library services within learning and teaching frameworks.

We have described relevant related works and software solutions. We have highlighted shortcomings and pros of those studies. Most of these studies have tended to produce solutions tied to commercial platforms or have given priority to questionable standard such as OAI-PMH, for achieving interoperability, as it was in the case of the BRICKS Project [54], which it seemed promising when presented web services based concepts for achieving integration. However the SOA factor and the ease of alternatives such as SRU/SRW were unnoticed by these projects.

Although SOA middleware reduces the need for system development and also management and maintenance burdens, the performance of SOA-based search services need to be monitored for large production services, because XML-based messaging services typically consume more computing resources and are slower than fast general purpose search engines. Early testes suggested that users have not found performance issues using the DL-VLE integration prototype system.

A key requirement for VLEs should be integration, and the tendency of using VLEs that do not support SOA, open standards and Web Services should be reversed. Main global and national organizations are working towards SOA e-Frameworks, where monolithic and centralized architectures are no longer taken into account for effective delivery of services. The ultimate aim of a VLE should be to provide a framework where service applications are embedded and integrated through agreed behaviors and interfaces using open web services technology to achieve interoperability.

The combination of open standards, “light” web services and SOA can produce powerful platforms that can help to develop information environments that are responsive to new generation of library users (the Net generation) that expect to find ubiquitous discovery tools, such as Google Scholar, in their learning environment systems.

Publishers and libraries share the ultimate goal of making scholarly content available for the users that need it. Both of them also face the same challenge produced by the movement towards openness in almost all aspects of e-learning. Clearly it has been demonstrated that both can benefit from open standards. Libraries and publishers no longer can expect that their users adapt to and learn about their existing closed, restricted and non-standard systems. It is them who need to provide open access to their assets for interoperability purposes.

Our study concludes that without open standards, any middleware used to integrated different systems is likely to become rather cumbersome and infeasible. The use of open standards reduces dependency and heterogeneity and it is a key facilitator for systems integration and for making reality service-oriented systems.

## Acknowledgements

The study described here is based upon work supported by the PerX research project of the JISC-funded Digital Repositories Programme. The authors acknowledge the efforts and help of the Heriot Watt University Library staff as well as give special recognition to Dr. David A. Cole for his fundamental role in the installation of the middleware on the VLE described here.

## Notes and References

[1] ROSENBERG, M. J. *e-Learning - Strategies for Delivering Knowledge in the Digital Age*. McGraw-Hill, 2001

[2] PINFIELD, S. *The changing role of subject librarians in academic libraries*. Journal of Librarianship and Information Science 33(1) pp. 32-38. 2001

[3] ALEXANDER, W. *Adaptive developments for learning in the hybrid library.* Ariadne Issue 24. 2000 [Available at http://www.ariadne.ac.uk/issue24/sellic/intro.html]

[4] MACCOLL, J. *Virtuous learning environments: the library and the VLE*. Program: electronic library & information systems, Volume 35, Number 3, pp. 227-239(13) 2001

[5] MCLEA, N.; LYNCH, C. *Interoperability between Library Information Services and Learning Environments - Bridging the Gaps*. A white paper on behalf of the IMS Global Learning Consortium [Available at http://www.imsglobal.org/digitalrepositories/CNIandIMS_2004.pdf]

[6] CARLSON, S. *The Deserted Library: As Students Work Online, Reading Rooms Empty Out Leading Some Campuses to Add Starbucks*. Chronicle of Higher Education Nov 16 2001 [Available at http://chronicle.com/free/v48/i12/12a03501.htm]

[7] FAULHABER, C. *Distance Learning and Digital Libraries: Two Sides of a Single Coin*. Journal of the American Society for Information Science, v47 n11 pp. 854-56 Nov 1996

[8] LOW, B.; MACCOLL, J. *Searching Heterogeneous e-Learning Resources*. Paper presented at the DELOS Workshop 2005, Digital Repositories: Interoperability and Common Services, May 2005

[9] BLACK, NANCY E. *Distance Library Services in Canada: Observations and Overview of Some of the Issues*. New Review of Libraries and Lifelong Learning 4, pp. 45-62. 2003

[10] FLECKER, D.; MCLEAN, N., *Digital Library Content and Course Management Systems: Issues of Interoperation.* Report of a study group funded by the Andrew W. Mellon Foundation. July 2004 [Available at http://www.diglib.org/pubs/dlf100/cmsdl0407.pdf]

[11] WHITING, J.; KARTUS, E.; RUNNER, E. *Challenges of Integrating Learning Resources within the Learning Management System at Deakin University*. Proceedings of the EDUCAUSE Australasia Conference, pp. 159-171. 2003.[Available at http://www.caudit.edu.au/educauseaustralasia/2003/EDUCAUSE/PDF/AUTHOR/ED030070.PDF]

[12] Pilot Engineering Repositories Xsearch (PerX) Project: http://www.icbl.hw.ac.uk/perx

[13] Blackboard VLE: http://www.blackboard.com/us/index.Bb

[14] HE, H. What Is Service-Oriented Architecture. Sept. 2003 [Available at http://www.xml.com/pub/a/ws/2003/09/30/soa.html]

[15] Open Archives Initiative: http://www.openarchives.org

[16] SRU/SRW Protocol: http://www.loc.gov/standards/sru

[17] Z39.50 Protocol: http://www.loc.gov/z3950/agency

[18] CHUMBE, S.; MACLEOD, R.; BARKER, P.; MOFFAT, M.; RIST, R. *Overcoming the obstacles of harvesting and searching digital repositories from federated searching toolkits, and embedding them in VLEs*. Proceedings of the 2nd International Conference on Computer Science and Information Systems, Greece. 2006

[19] ALONSO, G.; CASATI, F.; KUNO, H.; MACHIRAJU, V. *Web Services: Concepts, Architectures and Applications*. ISBN: 978-3-540-44008-6 Springer 2004

[20] OLDENETTEL, F.; MALACHINSKI, M.; REIL, D., *Integrating digital libraries into learning environments: the LEBONED approach.* Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries. 2003

[21] ZUR MUEHLEN, M.; NICKERSON, J.; SWENSON, K. *Developing Web Services Choreography Standards – The Case of REST vs. SOAP.* Decision Support Systems 37. Elsevier. 2004.

[22] FIELDING, R. *Architectural Styles and the Design of Network-based Software Architectures*. Doctoral Dissertation. University of California, Irvine, CA (180p.) 2000

[23] Blackboard Building Blocks. Developer Guide. Blackboard Learning System Release 7.1 2006

[24] JSR 154: JavaTM Servlet 2.4 Specification: http://jcp.org/en/jsr/detail?id=154

[25] Apache Tomcat Servlet: http://tomcat.apache.org/tomcat-6.0-doc/index.html

[26] POLGAR, J.; BRAM, R.; POLGAR, A. *Building and Managing Enterprise-wide Portals*. Idea Group Inc (IGI) 2006

[27] LAGOZE, C.; KRAFFT, D.; CORNWELL, T.; DUSHAY, N.; ECKSTROM, D.; SAYLOR, J. *Supporting education: Metadata aggregation and "automated digital libraries": a retrospective on the NSDL experience*. Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries JCDL '06, June 2006

[28] STER,N R.; MCELROY, M. *Virtual Collections: Challenges in Harvesting and Transforming Metadata from Harvard Catalogs for Topical Collections*. Proceedings of the DLF Fall 2006 Forum. Harvard University. Boston MA. 2006

[29] FOULONNEAU, M. at al. *Strand : Open Archives Protocol for Metadata Harvesting*. The Knowledge Exchange workshop on Institutional Repositories Report. Danish Library Agency in Copenhagen, Denmark. Feb. 2007 [Available at http://knowledge-exchange.net.dynamicweb.dk]

[30] JAFER Toolkit Project: http://www.jafer.org

[31] OLIVE Project Final Report (2004) [Available at http://www.jisc.ac.uk/uploaded_documents/OLIVE_Project_Report.pdf]

[32] RICHARDSON, J. *Building Bridges between Learning Management Systems and Library Content Systems*. Presented at the 11th Australian World Wide Web (AusWeb05) Conference, Gold Coast. 2005 [Available at http://ausweb.scu.edu.au/aw05/papers/refereed/richardson/index.html]

[33] WILSON, S.; BLINCO, K.; REHAK, D. *Service-Oriented Frameworks: Modelling the Infrastructure for the Next Generation of e-Learning Systems*. Presented at Alt-I-Lab Conference, 2004

[34] POWELL, A. *A 'service oriented' view of the JISC Information Environment*. UKOLN Report. Nov. 2005 [Available at http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/soa/jisc-ie-soa.pdf]

[35] PAYETTE, S. *Choosing Technology that can Evolve With User Needs. A service-oriented approach to e-research, e-scholarship, and advanced scholarly publication*. VALA 2006 Conference. Melbourne, Australia. February 2006 [Available at http://www.valaconf.org.au/vala2006/papers2006/97_Payette_Final.pdf]

[36] HUNTER, J. *Scientific Models – A user-oriented approach to integrating scientific data and digital libraries*. VALA 2006, Melbourne. February 2006 [Available at http://www.valaconf.org.au/vala2006/papers2006/55_Hunter_Final.pdf]

[37] JOHNSTON, L. *Development and Assessment of a Public Discovery and Delivery Interface for a Fedora Repository*. In D-Lib Magazine, Vol. 11, Number 10. 2005 [Available at http://www.dlib.org/dlib/october05/johnston/10johnston.html]

[38] LAGOZE, C; KRAFFTI, D.; PAYETTEI, S.; JESUROGAII, S. *What Is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL*. In D-Lib Magazine Vol. 11, Number 11. 2005 [Available at http://www.dlib.org/dlib/november05/lagoze/11lagoze.html]

[39] The JISC e-Framework for Education and Research: http://www.jisc.ac.uk/whatwedo/programmes/programme_eframework.aspx

[40] MOFFAT, M. *'Marketing' with Metadata - How Metadata Can Increase Exposure and Visibility of Online Content*. PerX Project Report. 2006 [Available at http://www.icbl.hw.ac.uk/perx/advocacy/exposingmetadata.htm]

[41] BRANTLEY, P., Greenbaum D. and Yee R. *Emerging Best Practices for Integrating Library Content and Services with Educational Technology*. Presented at EDUCAUSE Annual Conferences. 2003

[42] COLEMAN, R. *Publication, Business and the Digital Library*. Sydney University Press. 2006 [Available at http://www.valaconf.org.au/vala2006/papers2006/58_Coleman_Final.pdf]

[43] CrossRef: http://www.crossref.org

[44] Google Scholar: http://scholar.google.com/intl/en/scholar/about.html

[45] Directory of open access journals: http://www.doaj.org

[46] Scirus: http://www.scirus.com

[47] XIE, H.I. *Supporting ease-of-use and user control: desired features and structure of web-based online IR systems.* Information Processing and Management, Vol. 39, pp.899–922. 2003

[48] VILAR, P.; ZUMER, M. *Comparison and evaluation of the user interfaces of e-journals*. Journal of Documentation. Vol. 61, No. 2, pp.203–227. 2005

[49] JACSO, P. *Google scholar: the pros and the cons*. Online Information Review, Vol. 29, No. 2, pp.208–214. 2005

[50] MARKLAND, M. Embedding online information resources in Virtual Learning En*vironments: some implications for lecturers and librarians of the move towards delivering teaching in the online environment* Information Research, 8(4), paper no. 158. 2003. [Available at: http://informationr.net/ir/8-4/paper158.html]

[51] COYLE, K. *Open Source, open standards*. Information Technology and Libraries. Vol. 21, No. 1 pp. 33-36. 2002

[52] CORBA: http://www.corba.org

[53] LIPPINCOTT, J. *Net Generation Students and Libraries*. Chapter 13 in *Educating the Net Generation*. Diana G. Oblinger and James Oblinger Eds. E-Book. 2005 [Available at http://www.educause.edu/NetGenerationStudentsandLibraries/6067]

[54] BERTONCINI, M.; MASCI, M.; RONCA, A. *Paving the Way for the Next Generation Cultural Digital Library Services: The Case Study of 'Fortuna visiva of Pompeii' within the BRICKS Project*. Proceedings of the 10th International Conference on Electronic Publishing ELPUB2006. Bulgaria, June 2006 ISBN 978-954-16-0040-5, 2006, pp. 5-16 [Available at http://elpub.scix.net/cgi-bin/works/Show?245_elpub2006]

# Designing Metadata Surrogates for Search Result Interfaces of Learning Object Repositories: Linear versus Clustered Metadata Design

*Panos Balatsoukas; Anne Morris; Ann O'Brien*

Department of Information Science, Loughborough University
Loughborough, Leics, LE 11 3TU, UK
e-mail: {p.balatsoukas; a.morris; a.o-brien}@lboro.ac.uk

## Abstract

This study reports the findings of a usability test conducted to examine users' interaction with two different learning object metadata-driven search result interfaces. The first was a clustered metadata surrogate interface (where metadata elements were divided into sections), and the second a linear or single metadata surrogate interface (where all metadata elements were listed in a single record). The objectives of this research were: firstly, to investigate the time needed by learners to identify a relevant learning object, using both interfaces; secondly, to examine learners' subjective satisfaction for both interfaces; and finally, to study the impact of task complexity on users' interaction with both interfaces. To facilitate the objectives of the study, twelve postgraduate students participated in a user study which employed a multi-method approach and involved observation of users' interactions, subjective satisfaction questionnaires and semi-structured interviews. Data collected included the time needed for users to identify relevant learning objects in both interfaces and the rating of users' subjective satisfaction. In addition, qualitative data were collected based on interviews and the think aloud protocol. Parametric analysis (ANOVA tests) was conducted to identify statistically significant differences between the two interfaces in terms of time, user satisfaction and the impact of task complexity. The data analysis revealed that users needed less time to perform the tasks using the clustered metadata surrogate interface. This difference, however, was not significant. In addition, there was no significant impact of task complexity on user's performance. In terms of subjective satisfaction, however, the participants perceived the clustered metadata surrogate interface to be significantly more satisfying, stimulating and easy to use (F=89.690, p.<0.01). The findings of this study provide useful recommendations for the design of search result interfaces in learning object repositories.

**Keywords:** user-centred metadata; metadata design; e-learning; interface design

## 1 Introduction

Research on user-centred metadata surrogate design in search result interfaces can be divided in two research strands: Firstly, the presentation of metadata in search result interfaces, and secondly, the content of metadata in search result interfaces. The former covers aspects such as, interface layout, presentation, display format, interactivity and 'tailorability' of metadata in search result interfaces, while the latter examines the users' level of understanding, the usefulness and the quality of metadata semantics and vocabularies for relevance judgment.

### 1.1 Presentation of Metadata in Search Result Interfaces

Research in this area has been focused on a comparison between list, categorical/clustered and dynamic displays of metadata [1]. These studies employed a controlled – laboratory based experimental design. Most researchers concluded that users were more satisfied, selected fewer non-relevant documents and performed the task of judging the relevance of documents faster using the category-based interfaces [2-6]. Other researchers, however, did not observe significant differences between category-based and list-based interface designs [7], while other researchers revealed that users preferred dynamic rather than category-based interfaces [8].

Research has also investigated the way metadata elements should be arranged within metadata surrogates. A logical pattern has been identified, such that, metadata elements providing access or arranging access to the resource should follow content related elements, such as, the title, abstract, subject headings or keywords [9].

## 1.2 Content of Metadata in Search Result Interfaces

In addition, to these studies, the research area of user-centred relevance judgment has significantly advanced knowledge about the content and types of metadata that should be included in metadata surrogates in search result interfaces.

Researchers in this field have advocated the importance of the presence of an abstract in the metadata surrogate when users judge the relevance of documents [10-12]. More recent studies by Drori [13] and Paek et al [14] have also revealed that abstracts should include information related to the users' search query (contextual information) rather than simply presenting the first sentences of the document. Researchers have also revealed that topical or subject relevance is not the only criterion users employ to judge relevance when examining metadata surrogates or documents. Other criteria, were: the purpose and scope of the document, objectives, recency, source quality, reputation of the author, accessibility information and cost [11, 12, 15].

## 1.3 Design of Learning Object Metadata Surrogates in Search Result Interfaces

Although previous research has covered a wide range of issues related to metadata design in search result interfaces, the study of the presentation and content of learning object metadata has been neglected in the e-learning literature with only a few studies attempting to explore the phenomenon in some depth.

For example, the 'MetaTest' project investigated users' interaction with GEM-based metadata records. The findings of the study revealed that users were significantly more satisfied and made better relevance judgments when an abstract with information about the contents of the learning object was included in the metadata surrogate. Further clues that could be included in metadata surrogates were: relevance rankings, reviews and comments from others who did similar searches [16-18].

In another study, researchers evaluated the usability of the SearchLT learning object repository [19]. The study confirmed that well established heuristics, such as, the need for visibility and user-centred terminology should be applied in the design of learning object metadata surrogates in the search result interfaces of learning object repositories. The study also suggested that the contents of metadata records should be divided into clusters/categories and not be displayed as a list within a single and information cluttered surrogate. This issue had not been raised in earlier studies. Researchers in the SearchLT study did not compare different metadata interface designs (for example linear versus clustered metadata surrogates), did not employ parametric techniques for testing the statistical significance of the finding and neglected the impact of task complexity on user performance when using both linear and clustered metadata surrogate interfaces.

# 2 Aim and Objectives

The aim of this research was to examine users' interaction with two different learning object metadata-driven search result interfaces. The first was a linear metadata surrogate interface (where metadata elements were included within a single surrogate in a linear form), and the second a clustered metadata surrogate interface (where metadata elements were divided into sections). In particular, the objectives of this study were:

- To investigate the time needed by learners to identify a relevant learning object, using both interfaces;
- To study the impact of task complexity on users' interaction with both interfaces; and
- To examine learners' subjective satisfaction for both interfaces.

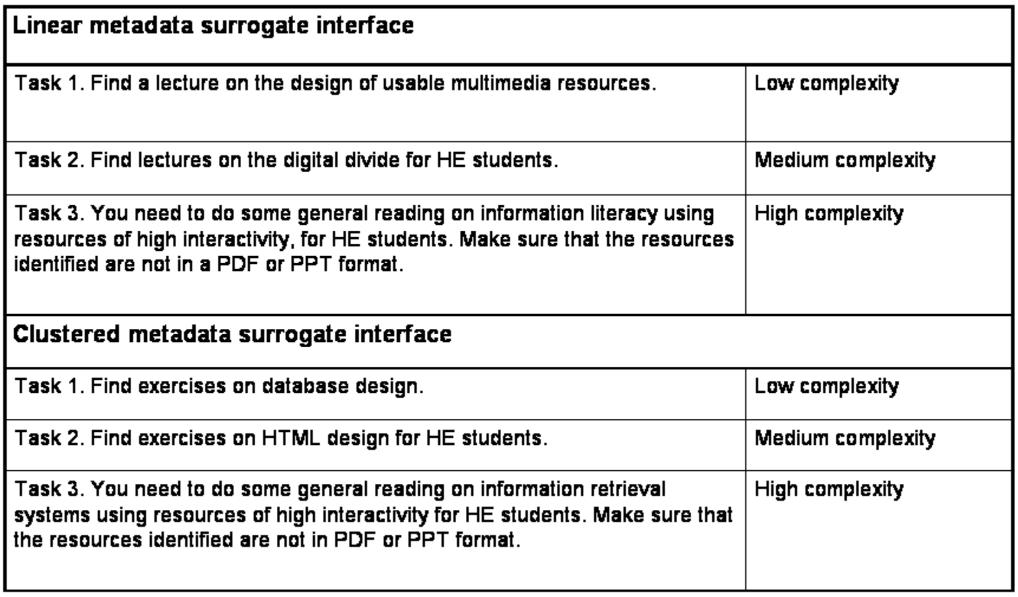
# 3 Methodology

## 3.1 Research Design

To address these objectives a usability test was conducted. A total of 12 postgraduate students in Information Science participated in the study. All participants were frequent users of Electronic Information Services (EIS) and the WWW. Students were recruited by means of e-mails and announcements on University notice-boards. A background questionnaire was completed by candidate participants before the usability tests. The background questionnaires facilitated the final selection of the participants in the study based on their familiarisation with EIS and the WWW.

The usability test employed a "within-subjects" design that required all participants to perform a similar set of tasks with both interfaces. The sequence with which the interfaces and the tasks were presented to the subjects was randomly altered for counterbalancing the effects of 'learning transfer'. Users had to perform three tasks in each interface. The tasks differed in terms of complexity. Table 1 presents the tasks assigned to the linear and clustered metadata surrogate interface and the level of complexity that each task represents.

During the testing, the 'think aloud' protocol was employed to elicit further qualitative and concurrent data about how users used both interfaces for identifying relevant learning objects. In addition, users' actions were captured through the use of screen recording software (Camtasia studio, v.4). After each 'task test' session users were asked to complete a subjective satisfaction questionnaire for each interface and took part in a semi-structured interview. In the interviews users were asked about their satisfaction with the learning object metadata elements, what additional metadata elements could be included in the records for facilitating relevance judgment, which style of presentation users liked the most, and how they would like metadata records to be displayed in search result interfaces. The research implements were piloted and then the testing took place during November 2006, at the Research School of Informatics of Loughborough University.



| **Linear metadata surrogate interface** | |
|---|---|
| Task 1. Find a lecture on the design of usable multimedia resources. | Low complexity |
| Task 2. Find lectures on the digital divide for HE students. | Medium complexity |
| Task 3. You need to do some general reading on information literacy using resources of high interactivity, for HE students. Make sure that the resources identified are not in a PDF or PPT format. | High complexity |
| **Clustered metadata surrogate interface** | |
| Task 1. Find exercises on database design. | Low complexity |
| Task 2. Find exercises on HTML design for HE students. | Medium complexity |
| Task 3. You need to do some general reading on information retrieval systems using resources of high interactivity for HE students. Make sure that the resources identified are not in PDF or PPT format. | High complexity |

**Table 1: Task table**

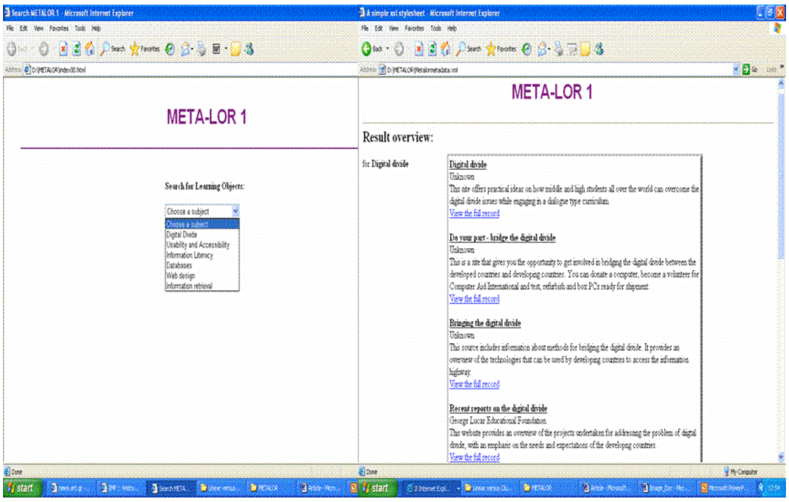
The data analysis included estimation of the means and statistical analysis (ANOVA tests) for the time needed for users to complete the tasks and users' ratings of subjective satisfaction. In addition, content analysis was performed on the qualitative data, such as the data collected from the interviews and the transcripts of the think aloud protocols.

## 3.2 The Meta-Lor 1 Interface

For the needs of this study a prototype learning object metadata repository system was set up using HTML, JavaScripts and XML technologies. The system stored and provided access to 60 learning object metadata records coded in XML. Metadata records included an identifier that provided access to the learning object itself. The data structure of the META-LOR 1 system is based on 19 elements derived from the LOM standard. The number and terminology of the elements was finalised after the pilot testing.

The prototype consisted of three main interfaces. The first was a simple search interface (see Figure 1). The second was the 'the search result overview' that provided a list of the retrieved results. The list included only the title of the learning object, the name of the author of the learning object, an abstract of the contents of the learning object and a link to the metadata record preview (see Figure 1). The third interface was 'the metadata surrogate preview' which included all 19 metadata elements. This interface comprised two different designs. The first presented metadata elements in a list (linear metadata surrogate interface) (see Figure 2) and the second divided metadata elements into three sections: General, Educational, and Technical metadata (clustered metadata surrogate interface) (see Figure 3).

In order to improve users' understanding of the metadata vocabularies, a definition of each metadata element was provided in a pop up box. The pop up boxes were contextually displayed every time the user selected a particular element with the mouse.



Search METALOR 1 - Microsoft Internet Explorer
A simple xsl stylesheet - Microsoft Internet Explorer

META-LOR 1

Search for Learning Objects:

Choose a subject
- Choose a subject
- Digital Divide
- Usability and Accessibility
- Information Literacy
- Databases
- Web design
- Information retrieval

META-LOR 1

Result overview:

for Digital divide

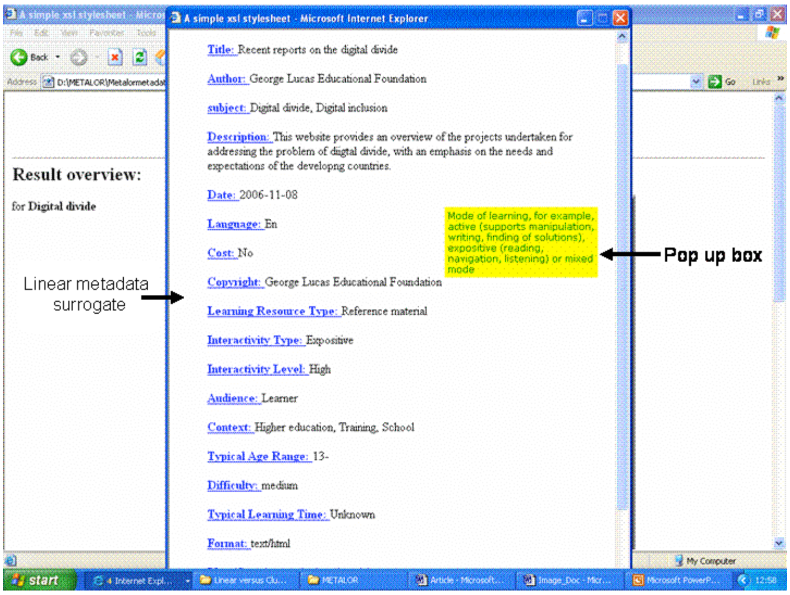
**Digital divide**
Unknown
This site offers practical ideas on how middle and high students all over the world can overcome the digital divide issues while engaging in a dialogue type curriculum.
View the full record

**Do your part - bridge the digital divide**
Unknown
This is a site that gives you the opportunity to get involved in bridging the digital divide between the developed countries and developing countries. You can donate a computer, become a volunteer for Computer Aid International and test, refurbish and box PCs ready for shipment
View the full record

**Bringing the digital divide**
Unknown
This source includes information about methods for bridging the digital divide. It provides an overview of the technologies that can be used by developing countries to access the information highway.
View the full record

**Recent reports on the digital divide**
George Lucas Educational Foundation
This website provides an overview of the projects undertaken for addressing the problem of digital divide, with an emphasis on the needs and expectations of the developng countries.
View the full record

**Figure 1: The search and search result overview interfaces**



Result overview:

for Digital divide

Title: Recent reports on the digital divide

Author: George Lucas Educational Foundation

subject: Digital divide, Digital inclusion

Description: This website provides an overview of the projects undertaken for addressing the problem of digital divide, with an emphasis on the needs and expectations of the developng countries.

Date: 2006-11-08

Language: En

Cost: No

Copyright: George Lucas Educational Foundation

Learning Resource Type: Reference material

Interactivity Type: Expositive

Interactivity Level: High

Audience: Learner

Context: Higher education, Training, School

Typical Age Range: 13-

Difficulty: medium

Typical Learning Time: Unknown

Format: text/html

Mode of learning, for example, active (supports manipulation, writing, finding of solutions), expositive (reading, navigation, listening) or mixed mode

Linear metadata surrogate

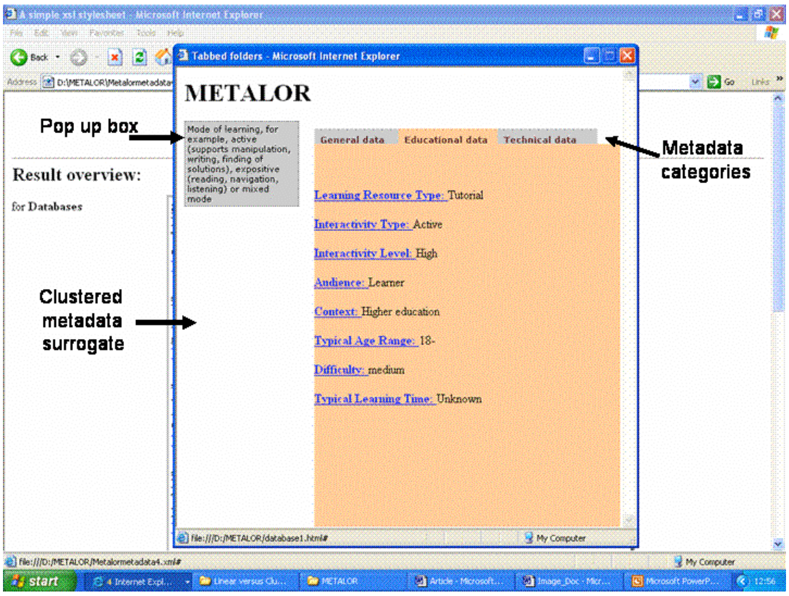
Pop up box

**Figure 2: Linear metadata surrogate interface**

# 4 Results and Discussion

## 4.1 Differences Between the Interfaces in Terms of Time

The analysis of time data revealed that participants performed the three tasks slightly faster using the clustered metadata surrogate interface. Participants needed an average time of 314 seconds to perform a task using the linear metadata surrogate interface and 301 seconds using the clustered metadata surrogate interface. This difference in time, however, was not statistically significant (p.>0.720).


A simple xsl stylesheet - Microsoft Internet Explorer
Tabbed folders - Microsoft Internet Explorer
Address D:\METALOR\Metalormetadata
Pop up box
Clustered metadata surrogate
Metadata categories
General data
Educational data
Technical data
file:///D:/METALOR/database1.html#
file:///D:/METALOR/Metalormetadata4.xml#
My Computer
start

METALOR

Result overview:

for Databases

Mode of learning, for example, active (supports manipulation, writing, finding of solutions), expositive (reading, navigation, listening) or mixed mode

Learning Resource Type: Tutorial

Interactivity Type: Active

Interactivity Level: High

Audience: Learner

Context: Higher education

Typical Age Range: 18-

Difficulty: medium

Typical Learning Time: Unknown

**Figure 3: Clustered metadata surrogate interface**

## 4.2 Impact of Task Complexity on the User Performance

Figure 4 summarizes the mean time needed for users to perform the three tasks in both interfaces. Based on these results it can be implicitly suggested that users completed the low and medium complexity tasks faster (Task 1 and Task 2) when they used the clustered metadata surrogate interface. On the other hand, users found it more efficient to identify relevant learning objects using the linear metadata surrogate interface (Task 3). These differences, however, were not statistically significant and further research is needed to investigate this further (p.>0.203).

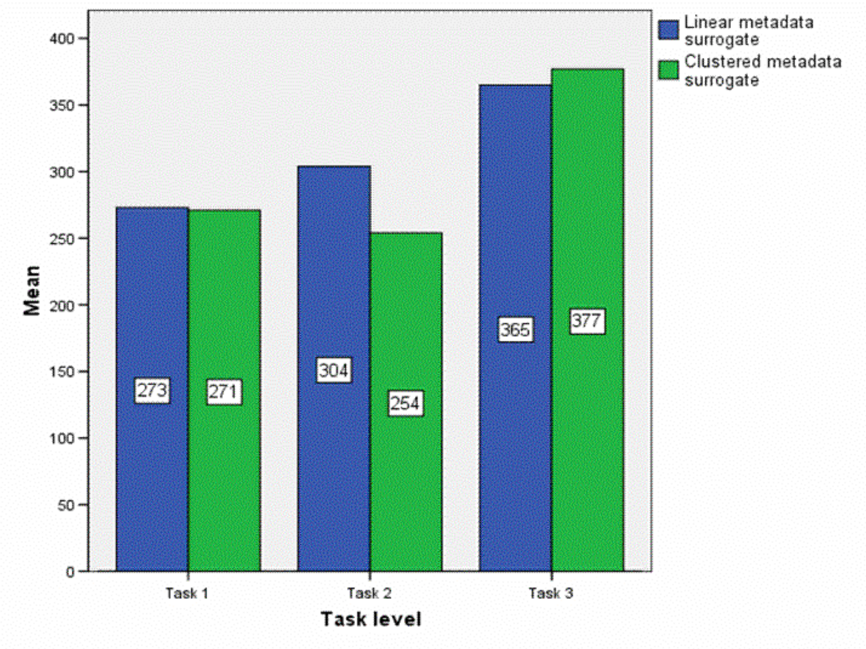


**Figure 4: Difference in time across the three tasks**

## 4.3 Subjective Satisfaction Questionnaire

Users' overall satisfaction with the two types of interfaces was measured in terms of four subjective measures: 1. Satisfaction, 2. Stimulation, 3. Easy of use and 4. Satisfaction with the presentation of metadata elements in the surrogate. The results revealed that participants' overall satisfaction (across the four subjective measures) was significantly higher in the case of the clustered metadata surrogate interface (mean overall satisfaction = 7.8) rather than the linear metadata surrogate interface (mean overall satisfaction = 6.3) (F=105.308, p.<0.01). Figure 5 presents the statistically and non-statistically significant differences between the two interfaces for each subjective measure ('Satisfaction': F= 6,796, p.>0.05; 'Stimulation': F=68,200, p.<0.01; 'Easy of use': F=35.200, p.<0.01; 'Satisfaction with metadata presentation in the surrogate': F=22,184, p.<0.01).

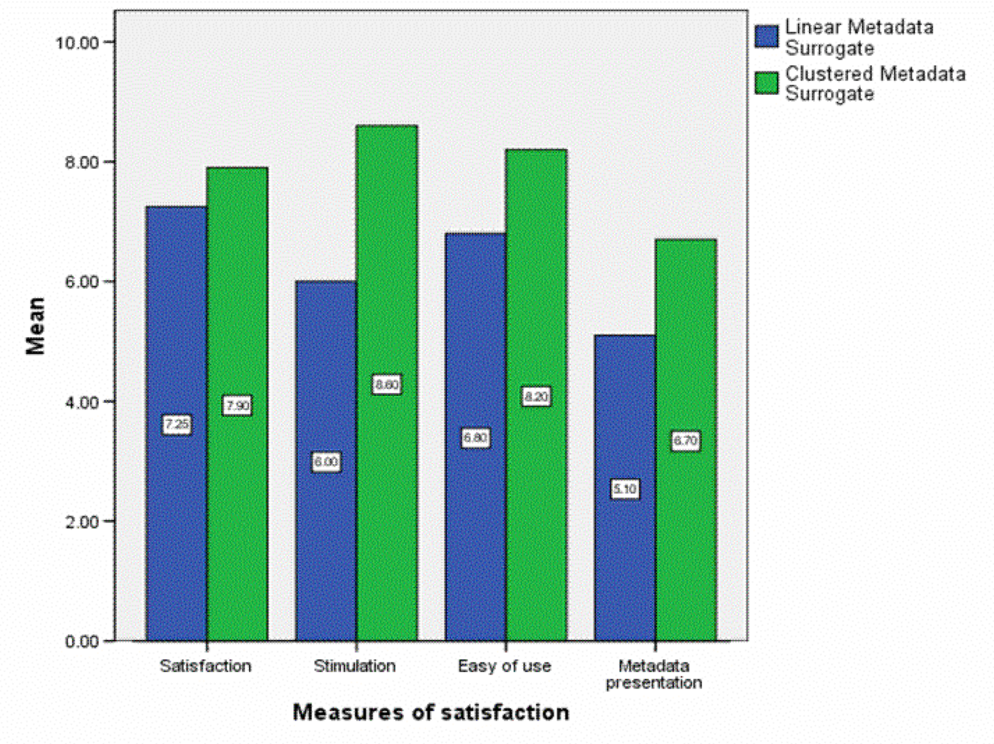


**Figure 5: Differences in subjective satisfaction.**

## 4.4 Qualitative Data (Interviews and Think Aloud Protocol)

### 4.4.1 Which Metadata Elements Users Liked The Most?

Participants in the study found the following metadata elements most useful for judging the relevance of learning objects: Title (12 users), Subject (12 users), Description (12 users), Difficulty level (12 users), Cost (10 users), Format (10 users), Identifier (10 users), Interactivity level (7 users), Audience (6 users). Based on these findings, it can be concluded that users favoured most of the 'General' and 'Technical' metadata categories. The former category was useful for users to judge the topical relevance of the learning object, while the latter provided users with information about how to access or download it. This observation agrees with previous studies that revealed users' preference towards content related metadata for evaluating the relevance of information and technical metadata for accessing it (Fraser and Gluck, 1999).

It is also worth mentioning that seven participants liked the use of pop up boxes that displayed information about learning object metadata elements and improved their understanding of some complex educational metadata elements, such as the 'interactivity type' and 'context' metadata. On the other hand, most of the educational elements were mentioned by only a few participants as useful. This, however, does not hold in the case of the 'difficulty', 'interactivity level' and 'audience' elements. These elements were regarded as useful by half or more participants in the study. This finding may be explained in two ways:

1. The laboratory and controlled nature of the study. This study employed a controlled task test with three predetermined tasks that were focused on specific metadata elements in order to be accomplished. More naturalistic studies that reflect real user needs in a variety of learning situations are needed to enhance the knowledge about how students use educational metadata to search for and judge the relevance of learning objects;

2. The semantic ambiguity of most of the educational metadata elements. During the task test a total of eight users found the meaning of some educational elements difficult to understand even though an explanatory pop up box accompanied these metadata elements. Future research should be focused on the investigation of the user-centeredness of the terminology and semantics of educational metadata. Such metadata should take into account learners' vocabularies and learning experiences.

### 4.4.2 What Other Metadata Elements Users Would Like to be Included?

The think aloud transcripts revealed that users did not like the inclusion of many metadata elements in the preview surrogates. This was more evident in the case of the linear metadata surrogate interface. Some of the comments users made regarding this issue were:

*"There is too much information [...] Should I read all of it?"*

[Male participant, linear metadata surrogate]

*"I'll only select to read some of them [...]. Is this information really needed?"*

[Male participant, linear metadata surrogate]

Although the length of the metadata records was criticized by some participants, few subjects mentioned that the metadata records should be enhanced with some additional information. Some of the information identified by participants is already included in the LOM standard. This includes: information about other similar resources available (Relation metadata category of LOM) and other people's comments about the resource (Annotation metadata category of LOM). It was interesting, however, that users identified the need for some information that is not explicitly covered by the existing LOM standard, such as, information about the time it takes for a learning object to be downloaded, accessibility needs (for example, how the learning object meets the needs of an heterogeneous community of learners) and information about the quality of a learning object.

### 4.4.3 Which Style of Metadata Presentation Users Liked the Most?

The majority of participants in the interviews (n=10) liked the way metadata was presented in the clustered metadata surrogate interface. The reasons that justify this preference are summarised in two categories:

1. *Plausibility and engagement*. The clustered metadata surrogate interface was characterised as more pleasant on the eye and more engaging. In addition, learners had the opportunity to focus on a specific category of metadata rather than the whole record;
2. *Structure and organisation of information*. The use of categories provides more structure to the metadata record. This minimises users' cognitive and memory load by making metadata more visible.

Two participants preferred the linear interface; they were more familiar with linear search result interfaces and metadata records and they did not like the way information was categorised in the clustered metadata surrogate interface. Further research needs to examine how metadata elements should be categorised in clustered metadata surrogates and whether the a priori categorisation of metadata elements as 'General', 'Educational' and 'Technical' should change in order to be better aligned with the mental model of learners.

### 4.4.4 Alternative Ways For Displaying Metadata Records

Although the majority of participants in the study preferred the clustered metadata interface, many users (n=9) mentioned that learners should be provided with the opportunity to control the display and the content of metadata records. The personalisation of metadata displays in search result interfaces could improve the relevance judgment process. For example, learners could select the number and type of metadata elements in search result interfaces. Such information could be either stored in learner profiles or generated dynamically during a search. In the latter case, the methodologies applied in relevance judgment research, sense-making and information foraging theory can provide a useful framework for predicting the dynamic nature of users' preferences.

## 5 Conclusions and Recommendations

This study suggests that learning object metadata elements in search result interfaces should be grouped into metadata categories. Although there is no significant impact of metadata interface design (linear or clustered metadata surrogate) on users' performance, users were significantly more satisfied with the clustered metadata surrogate interface which they perceived as stimulating and engaging to use. Future research should investigate the impact of different learning and cognitive styles on the design of learning object metadata-driven search result interfaces as well as employ more naturalistic research design methods. In addition, research is needed to challenge the a priori triple categorisation of learning object metadata as "General", "Educational" and "Technical". Although the findings of this study have implications in the design of metadata surrogates in other types of repositories and e-publishing systems, such as e-prints, institutional repositories, digital libraries, portals, library OPACs and WWW search engines, future research should focus on the needs of users of these systems as well. The paper concludes with some recommendations for the design of search result interfaces of learning object repositories. These are:

- The provision for alternative displays of metadata surrogates, for example, both in linear and clustered forms;
- The design of adaptive interfaces that present the content and format of metadata surrogates according to learners' needs;
- The use of pop up boxes for documenting and presenting the meaning of learning object metadata elements to users;
- The population of learning object repositories with less conventional educational metadata elements of LOM, such as, the 'difficulty level', 'interactivity level' and 'Audience'. In addition, there is a need for extending the LOM standard with information that users perceive as important to judge the relevance of learning objects, such as, 'the time it takes for a learning object to be downloaded', 'accessibility needs' information, as well as 'information about the 'quality' of a learning object.

## References

[1] IHADJADENE, M.; CHAUDIRON, S. The effect of individual differences on searching the web. In: *Proceedings of the 66th Annual meeting of the American society for information science and technology*, vol. 40, 2003, pp.240 – 246.

[2] GRANKA, L.; HEMBROOKE, H.; GAY, G.; FEUNSER, M. Eye-Tracking Analysis of User Behavior in WWW Search. *In: Proceedings of the 27th annual international conference on Research and development in information retrieval*, 2004, pp. 478-479. Available at: <http://www.cs.cornell.edu/People/tj/publications/granka_etal_04a.pdf>, [accessed 24.06.2006].

[3] RESNICK, M. L.; MALDONADO, C. A.; SANTOS, J. M.; LERGIER, R. Modeling On-line Search Behavior Using Alternative Output Structures. In Proceedings of the Human Factors and Ergonomics Society 45th Annual Conference,

Minneapolis, MN, USA, 2001, pp. 1166-1171.

[4] DUMAIS, S.; CUTRELL, E; CHEN, H. Optimising search by showing results. In: *Proceedings of CHI 2001*, 2001. Available at: <http://research.microsoft.com/~sdumais/chi2001.pdf>, [accessed 25.10.2006].

[5] CHEN, H.; DUMAIS, S. Bringing order to the Web: automatically categorising search results. In: *Proceedings of CHI-00, ACM International conference on human factors in computing systems, 2000, p.145-152.* Available at http://research.microsoft.com/~sdumais/chi00.pdf, [accessed 23.10.2006].

[6] ZAMIR, O.; ETZIONI, O. Grouper: a dynamic clustering interface to web search results. In: *Proceedings of the 8th International WWW conference. Toronto, Canada*, 1999, pp. 1361-1374.

[7] RELE, R.S.; DUCHOWSKI, A. T. Using eye tracking to evaluate alternative search results interfaces. In: *proceedings of the Human Factors and Ergonomics society, September 26-30 2005, Orlando, FL, HFES*, 2005. Available at:

<http://andrewd.ces.clemson.edu/research/vislab/docs/Final_HFES_Search.pdf>, [accessed 09.07.2006].

[8] PRATT, W.; FAGAN, L. The usefulness of dynamically categorizing search results. *Journal of the American Medical Informatics Association*, 7 (6), 2000, pp. 605-617.

[9] FRASER, B.; GLUCK, M. Usability of geospatial metadata or space-time matters. *Bulletin of the American Society for Information Science*, vol.25, no.6, August/September 1999. Available at: <http://www.asis.org/Bulletin/Aug-99/fraser_gluck.html>, [accessed 12.07.2006].

[10] SARACEVIC, T. Relevance: a review of and framework for thinking on the notion in information science. *Journal of the American Society for Information Science*, no.6, November-December 1975, pp.321-344.

[11] BARRY, C. User-defined relevance criteria: an exploratory study. *Journal of the American Society for Information Science and Technology*, vol.45, no.3, 1994, pp.149-159.

[12] TANG, R.; SOLOMON, P. Use of relevance criteria across stages of document evaluation: on the complementarity of experimental and naturalistic studies. *Journal of the American Society for Information Science and Technology*, vol.52, no.8, 2001, pp.676-685.

[13] DRORI, O. How to display search results in digital libraries : user study. *In proceedings of the New Developments in Digital Library, 3rd International workshop*, 2003. Available at: <shum.huji.ac.il/~of...drori042003c.pdf>, [accessed 24.07.2006].

[14] PAEK, T.; DUMAIS, S.; LOGAN, R. WaveLens: a new view onto Internet search results. *In: proceedings of CHI2004, April 24-29, 2004, Vienna, Austria,2004, pp.727-734*. Available at: <http://research.microsoft.com/~timpaek/Papers/chi2004.pdf>, [accessed 25.10.2006].

[15] RIEH, S. Y. Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology*, vol.53, no.2, 2002, pp.145 – 161.

[16] LIDDY, et al. MetaTest: evaluation of metadata from generation to use. In: *Proceedings of the 3rd Joint Conference on Digital Libraries*. IEEE., 2003, pp.398. Available at: <http://csdl2.computer.org/comp/proceedings/jcdl/2003/1939/00/19390398.pdf>, [accessed 18.07.2006].

[17] LIDDY, L.; FINNERAN, T. *Meta Test: automatic generation of metadata and preliminary evaluation of its utility in information seeking*. [power point presentation], [n.d]. Available at: <http://nsdl.comm.nsdlib.org/meeting/poster_docs/2003/1093_MetaTest.pdf?nsdl_annual_meeting=d0e759bf75a8ccda313c2639cb72be81>, [accessed 03.07.2006].

[18] DIEKEMA, A. R. Evaluating metadata from different perspectives. [power point presentation]. *In: Metadata tools for digital resources repositories, JCDL Workshop,* June 15, 2006. Available at <http://www.ils.unc.edu/mrc/jcdl2006/slides/diekema.pdf>, [accessed 12.07.2006].

[19] *FAILTE's SearchLT evaluation.* FAILTE, 2002. Available at: <http://www.failte.ac.uk/documents/eval_report.rtf>, [accessed 12.07.2006].

# Disclosing Freedom of Information Releases

*Ann Apps*

MIMAS, The University of Manchester, M13 9PL, UK
e-mail: ann.apps@manchester.ac.uk

## Abstract

The Freedom of Information (FOI) Acts passed in 2000 in England and Wales and in 2002 in Scotland require organisations, including UK Higher Education Institutions (HEI), to provide requested information within certain conditions. The JISC Information Governance Gateway (JIGG) project aims to provide a single online gateway into information and resources related to HEIs' compliance with information governance legislation, including FOI. One of the project's objectives is to provide dissemination of the FOI disclosure logs by a web search within the gateway and also using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). It is hoped this will assist HEI records management practitioners in sharing their experiences of dealing with FOI requests, and lead to future collaborations within a wider community. This paper describes the development of a JIGG FOI Application Profile as a 'template' for FOI disclosure log entries, and its subsequent translation into a practical application.

**Keywords:** Freedom of Information; information governance; records management; OAI-PMH; Dublin Core Application Profile

## 1 Introduction

The Freedom of Information (FOI) Acts passed in 2000 in England and Wales [1] and in 2002 in Scotland [2] require organisations to provide requested information to enquirers within a given timescale, unless the requested information is exempted under the legislation. The organisations covered by this requirement include UK Higher Education Institutions (HEI). To show just a few examples, people have asked The University of Manchester for information about the University's coat of arms, the amount of money taken in library fines, prospectuses supplied on recycled paper, and the awarding of honorary degrees to the Bee Gees.

HEIs are encouraged to publish disclosure logs that summarise the FOI requests they have received and the information they have released. Currently only a minority of HEIs maintain such public disclosure logs and in most cases these consist of simple lists on web pages, for example one page per year. Many other UK organisations do publish FOI disclosure logs but not in any consistent format or single place [3].

The JISC Information Governance Gateway (JIGG) project [4] aims to provide a single online gateway into information and resources related to HEIs' compliance with information governance legislation, including FOI, as well as data protection, environmental information, practical issues such as records management and related legislation such as copyright. The gateway also provides a private discussion area for HEI records management practitioners.

In addition to being a portal for relevant and up-to-date information about governance resources, JIGG publishes HEI Publication Schemes as defined under the FOI Acts, and Disclosure Logs where available. A project objective is to provide dissemination of the FOI disclosure logs [5] using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [6] as well as providing a web search within the gateway. Provision of a JIGG FOI OAI-PMH service will allow other applications regularly to gather new JIGG FOI disclosure log entries into their own databases. It is hoped this will assist HEI records management practitioners in sharing their experiences of dealing with FOI requests, and lead to future collaborations within a wider community.

## 2 Methodology

### 2.1 The JIGG FOI Domain Model

In order to disseminate the FOI disclosure log entries over OAI-PMH it was necessary to define a 'template' of appropriate data fields. This was developed by investigating the content of existing FOI disclosure logs, with subsequent agreement within the community, the ‘template’ being expected to conform to the requirements of the Office of the Information Commissioner, the Scottish Information Commissioner and the Department of Constitutional Affairs. Thus the main entity within the JIGG FOI domain, or application, model is an FOI disclosure log entry, comprising this identified set of properties. Additionally each disclosure log entry has an associated set of administrative metadata that describes information ‘about’ the disclosure log entry within the application.

A disclosure log entry is a ‘closed’ record of a request for, and the release of, information. It is not anticipated that there would be any changes to FOI disclosure log entries after they’ve been entered into the JIGG system, except for minor textual corrections.

## 2.2 The JIGG FOI Application Profile

The disclosure log entry and the administrative metadata are documented using a Dublin Core Application Profile, based on the European standard (CEN Workshop Agreement) Dublin Core Application Profile (DCAP) Guidelines [7]. Standard Dublin Core properties [8] are used where applicable. The Application Profile indicates the source Dublin Core definitions of and comments about these properties as well as the application specific variations. Additional JIGG specific properties have been introduced where there were no suitable standard properties. Each of these is defined in a JIGG FOI namespace, currently within a human readable ‘mini application profile’ with its URI (Uniform Resource Identifier, a unique persistent identifier within the global internet) grounded on its position in that document, and with an intention of persistence.

The DCAP Guidelines specify an application profile that captures a single entity. This corresponds to a single resource description within the Dublin Core Abstract Model (DCAM) [9], which specifies a flat set of properties for a single resource, with no provision for any composite properties according to any hierarchical model and syntax. Thus some extension of the DCAP has been made to capture both a disclosure log entry and its associated administrative metadata, which together make up a “description set” within the DCAM. Thus the DCAP is composite with a section for each “description”, preceded by a section that specifies these entities.

## 2.3 FOI Disclosure Log Entry Properties

Some of the properties within an FOI disclosure log entry are available for discovery purposes through the application. These properties capture the content of the request, when it was made, the HEI, and the relevant legislation and exemptions. All the properties provide documentation of the course of a request, such detail being potentially useful for processing future similar requests.

Table 1 lists the FOI disclosure log entry properties taken from namespaces Dublin Core (‘dc’ and ‘dcterms’) and JIGG (‘jigg’), the URIs of which are defined within the ‘dcxm:descriptionSet’ of Figure 2. Occurrence requirements are displayed as ‘Min’ and ‘Max’, which also implicitly indicate whether properties are mandatory or optional and whether they are repeatable.

An FOI disclosure log entry contains several properties that summarise in free text the information that was requested, the information that was released and how the request was processed. Each disclosure log entry has a title indicating very briefly the topic of information requested. A more detailed summary of the request is the free text value of a description property. Optional summaries of the information released and the process of answering the request are further text fields. The topic of a request may be indicated by terms taken from the JISC Function Activity Model Vocabulary [10]. Each of the textual fields can be tagged with a language code, included for possible future enhancement of the application.

The HEI to which the FOI request was made is captured as the publisher. The JIGG FOI data submission system aims to ensure consistency of institution names. The country where the HEI is based is captured with values taken from a JIGG-defined vocabulary, which currently contains only the four UK countries (England, Northern Ireland, Scotland and Wales). An optional local identifier may be included where it is used by the HEI to denote the log entry. Ideally this identifier should be a URI and mandatory. But it was felt that at this stage of the JIGG project, requiring a global identifier may present too high a barrier to inclusion of records from information governance practitioners who may not currently use a consistent identification system, and may not be conversant with URIs. HEIs may currently publish some details of their FOI requests in some way, possibly a web page describing several requests within some time period. Thus a link to this composite disclosure log is included. Some HEIs publish the full text of the information released, so the disclosure log entry has an optional, repeatable link to possibly several documents.

| Property | Definition (Summary) | Content / Vocabulary | Min | Max |
|---|---|---|---|---|
| dc:title | Title of log entry | text | 1 | 1 |
| dc:identifier | Local identifier | | 0 | 1 |
| dc:publisher | Organisation publishing log | text | 1 | 1 |
| jigg:country | Country of publishing organisation | England; Northern Ireland; Sotland; Wales | 1 | 1 |
| dcterms:isPartOf | Organisation's disclosure log | URI | 1 | 1 |
| jigg:dateReceived | Date FOI request received | W3CDTF | 1 | 1 |
| dc:description | Summary of information requested | text | 1 | 1 |
| dc:subject | Topic of request | Function Activity Model vocabulary (in jigg namespace) | 0 | unbounded |
| jigg:infoReleased | How much information released | no; partial; yes | 1 | 1 |
| jigg:legislation | Applicable legislation | Freedom of Information Act 2000;<br>Freedom of Information (Scotland) 2002;<br>Environmental Information Regulations 2004;<br>Environmental Information (Scotland) Regulations 2004 | 1 | 1 |
| jigg:exemptionsUsed | Exemptions used when processing request | Exemptions relevant to above applicable legislation taken from vocabulary in jigg namespace | 0 | unbounded |
| jigg:requestHistory | How request was processed | text | 0 | 1 |
| jigg:responseSummary | Summary of information released | text | 0 | 1 |
| dcterms:references | Full text response | URI | 0 | unbounded |

**Table 1: FOI Disclosure Log Entry Properties (jigg:foiLog)**

The date on which an FOI request was received is included within the disclosure log entry. It is probable that this date will be used for discovery, as well as providing a record of when the request occurred. Within the plethora of Dublin Core 'date' element refinements there was not one that exactly fitted the semantics of receipt date. It seemed a better option to define a JIGG-specific property rather than trying to shoehorn a definition into an inappropriate Dublin Core date property, or adopting a term from an obscure namespace. Theoretically it would seem appropriate to indicate closure in dealing with an FOI request by also capturing the completion date, especially as the FOI Acts specify time limits. However, in practice, practitioners were reluctant for this detail to be included, partly because of a belief that it would expose them to unwelcome levels of scrutiny. Because there are concerns about the initial level of engagement with the project by HEI FOI practitioners, it was decided to omit completion date from the set of properties in the Application Profile.

The FOI Acts define various exemptions under which it is permissible to refuse to provide information or to supply it only partially. There are some differences in the lists of exemptions or in their wording between the Freedom of Information Act 2000, which applies to England and Wales [11], and the Freedom of Information Act (Scotland) 2004 [12]. Another pair of differing exemptions lists apply to the Environmental Information Regulations 2004 [13] and the Environmental Information Regulations (Scotland) 2004 [14]. Thus it is beneficial to record under which legislation, of these four, an FOI request has been processed, which exemptions were relevant (a repeatable property), and how much information was released (possible values being 'no', 'partial', or 'yes'). Each of these properties has a JIGG-defined vocabulary. Although the exemptions associated with the various legislations are listed in several publicly available documents, there do not appear to be any existing formal vocabularies. Thus vocabularies have been defined within the JIGG FOI namespace for reference by property values within the JIGG FOI application.

## 2.4 FOI Disclosure Log Entry Administrative Metadata

Associated with an FOI disclosure log entry is a set of administrative metadata, listed in Table 2. This includes the URI of the publishing organisation, which is JIGG itself for the central JIGG FOI application, and the originating organisation, which is the URI corresponding to the HEI named as publisher within the log entry itself. A JIGG identifier, a URI, is assigned to each FOI disclosure log entry within the application.

There are various rights captured about the disclosure log entry. Copyright belongs to the publishing HEI. Creative Commons [15] rights cover subsequent use of the disclosure log entry, indicating that the information is freely available for non-commercial use, provided attribution of its provenance is maintained, but no derivatives may be made. This seemed an appropriate rights statement for information released from publicly funded HEIs. A further rights statement requires that this administrative metadata must always be retained with the disclosure log entry.

The date when a disclosure log entry was entered into the JIGG FOI database, or when it was last updated, is recorded as ‘dcterms:modified’, which “dumbs down” to ‘dc:date’ when the administrative metadata is supplied according to simple Dublin Core. This is the significant date used for the OAI-PMH application, which provides harvesting based on ‘last modification date’.

Finally there is a relation that ties the administrative metadata to the disclosure log entry. This relation is used when both entities appear within an XML document description set, with value an internal identifier of the disclosure log entry within that document. It is not used for OAI-PMH dissemination where the relation between the ‘about’ part of a record and the ‘metadata’ is implicit.

| **Property** | **Definition (Summary)** | **Content** |
|---|---|---|
| dc:identifier | Identifier of log entry within JIGG | URI |
| dc:creator | Originating organisation | URI |
| dc:publisher | Publisher of disclosure log entry | “http://www.jigg.ac.uk” |
| dcterms:modified | Date log entry added to JIGG repository | W3CDTF |
| dc:rights | Copyright over log entry | text |
| dc:rights | Creative Commons rights over reuse | “http://creativecommons.org/licenses/by-nc-nd/2.0/uk/” |
| dc:rights | Administrative metadata requirement | “The JIGG administrative metadata must always be retained with its associated disclosure log entry description.” |
| dc:relation | Link to related FOI disclosure log entry | Local identifier within an XML document |

**Table 2: FOI Disclosure Log Entry Administrative Metadata (jigg:foiAdmeta)**

## 2.5 The JIGG FOI XML Serialisation

Dissemination of records via OAI-PMH is by an XML serialisation of the data that is defined in the JIGG Application Profile [16] and conformant to an XML schema. Because the Application Profile is conformant to the Dublin Core Abstract Model (DCAM), it seemed appropriate to follow Dublin Core guidelines for the XML serialisation. A proposed ‘Dublin Core in XML’ [17] format that is consistent with the DCAM is under development by the Dublin Core Metadata Initiative. However the capability within this proposed XML format to support the full DCAM results in a rather verbose XML record for general usage. Although the XML data is intended for use by machines, which have no concerns about complexity apart from efficiency, there are also human considerations. A complex XML format requires more effort to both create and process, and so is consequently more error prone. Because of these concerns a restricted functionality version of Dublin Core in XML has also been suggested. The JIGG FOI XML schema follows this ‘Dublin Core in XML Minimal’ [18] as it is was proposed at the time of schema development.

```
@prefix dc: <http://purl.org/dc/elements/1.1/>
@prefix dcterms: <http://purl.org/dc/terms/>
@prefix jigg: <http://www.jigg.ac.uk/foi/terms/#>

DescriptionSet (
#FOILog
   Description ( DescriptionId (log-1234567-890123)
      Statement ( PropertyURI ( dc:title )
          ValueString ( "An Example FOI Disclosure Log"  Language ( en-GB ) ) )
      Statement ( PropertyURI ( dc:identifier )
          ValueString ( "02/2006" ) )
      Statement ( PropertyURI ( dc:publisher )
          ValueString ( "Somewhere University" ) )
      Statement ( PropertyURI ( jigg:country )
          VocabularyEncodingSchemeURI ( jigg:FoiCountryVocab )
          ValueString ( "England" ) )
      Statement ( PropertyURI ( dcterms:isPartOf )
          ValueURI ( <http://www.somewhere.ac.uk/foilogs/> ) )
      Statement ( PropertyURI ( jigg:dateReceived )
          ValueString ( "2006-08-01"  SyntaxEncodingScheme ( dcterms:W3CDTF ) ) )
      Statement ( PropertyURI ( dc:description )
          ValueString ( "Details of the information requested"  Language ( en-GB ) ) )
      Statement ( PropertyURI ( dc:subject )
          VocabularyEncodingSchemeURI ( jigg:FAMVocab )
          ValueString ( "D External Relations" ) )
      Statement ( PropertyURI ( jigg:infoReleased )
          VocabularyEncodingSchemeURI ( jigg:InfoReleasedVocab )
          ValueString ( "partial" ) )
      Statement ( PropertyURI ( jigg:legislation )
          VocabularyEncodingSchemeURI ( jigg:FoiLegislationVocab )
          ValueString ( "Freedom of Information Act 2000" ) )
      Statement ( PropertyURI ( jigg:exemptionsUsed )
          VocabularyEncodingSchemeURI ( jigg:FoiAct2000Vocab )
          ValueString ( "26 Defence" )
          ValueString ( "29 The Economy" ) )
      Statement ( PropertyURI ( jigg:requestHistory )
          ValueString ( "Details of processing the request"  Language ( en-GB ) ) )
      Statement ( PropertyURI ( jigg:responseSummary )
          ValueString ( "The answer" Language ( en-GB ) ) )
      Statement ( PropertyURI ( dcterms:references )
          ValueURI ( <http://www.somewhere.ac.uk/foilogs/02-2006.pdf> ) ) )
#FOIAdmeta
   Description ( ResourceURI( http://www.jigg.ac.uk/foi/ids/1234567-890123)
      Statement ( PropertyURI ( dc:identifier )
          ValueURI ( http://www.jigg.ac.uk/foi/ids/1234567-890123 ) )
      Statement ( PropertyURI ( dc:creator )
          ValueURI ( <http://somewhere.ac.uk> ) )
      Statement ( PropertyURI ( dc:publisher )
          ValueURI ( <http://www.jigg.ac.uk> ) )
      Statement ( PropertyURI (dcterms:modified )
          ValueString ( "2006-09-05"  SyntaxEncodingScheme ( dcterms:W3CDTF ) ) )
      Statement ( PropertyURI ( dc:rights )
          ValueString ( "Copyright Somewhere University 2006" ) )
      Statement ( PropertyURI ( dc:rights )
          ValueURI ( <http://creativecommons.org/licenses/by-nc-nd/2.0/uk/>  ) )
      Statement ( PropertyURI ( dc:rights )
          ValueString ( "The JIGG administrative metadata must always be retained with its
                         associated disclosure log entry description." ) )
      Statement ( PropertyURI ( dc:relation )
          DescriptionRef ( log-1234567-890123 ) ) ) )
```

**Figure 1: A DC-Text Example of an FOI Disclosure Log Entry**

As an interim stage, a DC-Text [19] hypothetical example was produced to illustrate conformance to the DCAM. DC-Text provides a formal but relatively syntax-free means to document a metadata description set that is ideal for the development and discussion stage. This DC-Text example, which informed the development of the JIGG FOI XML schema [20], is shown in Figure 1.

# 3 Results

## 3.1 Dissemination of FOI Disclosure Log Entries

FOI disclosure log entries stored in the JIGG FOI central database are disseminated over OAI-PMH according to an ‘oai_jiggfoi’ metadata format. The XML ‘metadata’ part of a ‘GetRecord’ response conforms to the JIGG FOI XML schema. Examples are shown in Figures 2 and 3, illustrating different styles of FOI records management by two HEIs, and the use of different properties taken from the Application Profile.

```
<metadata>
 <dcxm:descriptionSet xmlns:dcxm=“http://dublincore.org/xml/dc-xml-min/2006/09/18/”
  xmlns:dc=“http://purl.org/dc/elements/1.1/” xmlns:dcterms=“http://purl.org/dc/terms/”
  xmlns:jigg=“http://www.jigg.ac.uk/foi/terms/#” xmlns=“http://www.jigg.ac.uk/foi/”
  xmlns:xsi=“http://www.w3.org/2001/XMLSchema-instance”
  xsi:schemaLocation=“http://www.jigg.ac.uk/foi/ http://www.jigg.ac.uk/foi/schemas/2006/10/30/jiggfoi.xsd”>
  <jigg:foiLog>
   <dc:title>LJMU Review Magazine cost</dc:title>
   <dc:identifier>FOI 5/12</dc:identifier>
   <dc:publisher>Liverpool John Moores University</dc:publisher>
   <jigg:country dcxm:vocabEncSchemeURI=“http://www.jigg.ac.uk/foi/terms/#FoiCountryVocab”>
    England</jigg:country>
   <dcterms:isPartOf dcxm:valueURI=“http://www.ljmu.ac.uk/secretariat/75554.htm”/>
   <jigg:dateReceived dcxm:syntaxEncSchemeURI=“http://purl.org/dc/terms/W3CDTF”>
    2005-06-07</jigg:dateReceived>
   <dc:description>
    Costs of producing the LJMU Review Magazine and detailed accounts for the Marketing department for
     the previous 5 years
   </dc:description>
   <jigg:infoReleased dcxm:vocabEncSchemeURI=“http://www.jigg.ac.uk/foi/terms/#InfoReleasedVocab”>
    partial</jigg:infoReleased>
   <jigg:legislation dcxm:vocabEncSchemeURI=“http://www.jigg.ac.uk/foi/terms/#FoiLegislationVocab”>
    Freedom of Information Act 2000</jigg:legislation>
   <jigg:exemptionsUsed dcxm:vocabEncSchemeURI=“http://www.jigg.ac.uk/foi/terms/#FoiAct2000Vocab”>
    43 Commercial interests</jigg:exemptionsUsed>
   <jigg:requestHistory>
    LJMU disclosed information on the costs of producing the magazine, although not all the information
    requested was held. LJMU, after applying the public interest test, refused disclosure of the details
    accounts of the Marketing Department, citing Section 43 of the FOIA.
   </jigg:requestHistory>
  </jigg:foiLog>
 </dcxm:descriptionSet>
</metadata>
```

**Figure 2: An Example FOI Disclosure Log Entry from Liverpool John Moores University**

As required by OAI-PMH, records are also disseminated in simple Dublin Core for interoperability, informed by a mapping from the FOI Application Profile. The administrative metadata is disseminated in simple Dublin Core within an ‘about’ section of the ‘GetRecord’ response, an example being shown in Figure 4. Further ‘about’ sections detail metadata rights according to the OAI-PMH Guidelines for Conveying Rights, and the provenance of any records that have been harvested from elsewhere, conforming to the appropriate OAI-PMH Provenance schema.

```
<metadata>
 <dcxm:descriptionSet xmlns:dcxm=“http://dublincore.org/xml/dc-xml-min/2006/09/18/”
  xmlns:dc=“http://purl.org/dc/elements/1.1/” xmlns:dcterms=“http://purl.org/dc/terms/”
  xmlns:jigg=“http://www.jigg.ac.uk/foi/terms/#” xmlns=“http://www.jigg.ac.uk/foi/”
  xmlns:xsi=“http://www.w3.org/2001/XMLSchema-instance”
  xsi:schemaLocation=“http://www.jigg.ac.uk/foi/ http://www.jigg.ac.uk/foi/schemas/2006/10/30/jiggfoi.xsd”>
  <jigg:foiLog>
   <dc:title>Student accommodation landlords</dc:title>
   <dc:publisher>The University of Manchester</dc:publisher>
```

```
      <jigg:country dcxm:vocabEncSchemeURI="http://www.jigg.ac.uk/foi/terms/#FoiCountryVocab">
        England</jigg:country>
      <dcterms:isPartOf dcxm:valueURI="http://www.manchester.ac.uk/aboutus/documents/foi/disclosurelog/"/>
      <jigg:dateReceived dcxm:syntaxEncSchemeURI="http://purl.org/dc/terms/W3CDTF">
        2005-01-05</jigg:dateReceived>
      <dc:description>List of landlords on the student accommodation list</dc:description>
      <dc:subject dcxm:vocabEncSchemeURI="http://www.jigg.ac.uk/foi/terms/#FAMVocab">
        B Student Administration and Support</dc:subject>
      <jigg:infoReleased dcxm:vocabEncSchemeURI="http://www.jigg.ac.uk/foi/terms/#InfoReleasedVocab">
        yes</jigg:infoReleased>
      <jigg:legislation dcxm:vocabEncSchemeURI="http://www.jigg.ac.uk/foi/terms/#FoiLegislationVocab">
        Freedom of Information Act 2000</jigg:legislation>
      <dcterms:references
        dcxm:valueURI="http://www.manchester.ac.uk/medialibrary/foi/disclosures/studentadmin/landlords.pdf"/>
    </jigg:foiLog>
  </dcxm:descriptionSet>
</metadata>
```

**Figure 3: An Example FOI Disclosure Log Entry from The University of Manchester**

```
<about>
  <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
    xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
      http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
    <dc:creator>http://www.manchester.ac.uk</dc:creator>
    <dc:publisher>http://www.jigg.ac.uk</dc:publisher>
    <dc:date>2007-03-28</dc:date>
    <dc:rights>Copyright The University of Manchester 2005</dc:rights>
    <dc:rights>http://creativecommons.org/licenses/by-nc-nd/2.0/uk/</dc:rights>
    <dc:rights>The JIGG administrative metadata must always be retained with its associated log entry
      description.</dc:rights>
  </oai_dc:dc>
</about>
```

**Figure 4: Administrative Metadata for the FOI Disclosure Log Entry of Figure 3.**

## 3.2 Data Creation

Submission of FOI disclosure log entries into the JIGG central database was a significant factor for the project to consider. A Web-form based data Editor allows input of values of the various fields defined by the Application Profile. Use of a dedicated Editor ensures consistency of records, in particular with respect to the various vocabularies, and the generation of valid XML. The design of this system is based on that of the JISC Information Environment Service Registry (IESR) [21], as is the rest of the JIGG FOI application, thus reusing an established application developed as part of another JISC project.

A consideration is obviously the effort required to supply FOI disclosure log entries to JIGG, and the additional steps that may have to be included in existing workflows. It is hoped that use of the JIGG FOI data Editor for log entry submission will not be too onerous. It is thought that the majority of administrative operations within HEIs are based on Excel spreadsheets, thus necessitating 'copy and paste' into the JIGG submission form.

Where possible data fields are populated automatically, which also ensures consistency. For example, the publisher's name will be taken from an HEI's initial registration as a data contributor. Further data congruity is achieved by setting values from vocabulary term lists, such as exemptions, by selection menus within the Editor.

## 3.3 Data Harvesting

A future vision is automatic population of the central JIGG FOI database of disclosure log entries via OAI-PMH. If an HEI provided a harvest service onto their FOI disclosure logs, using the OAI-PMH standard and the 'oai_jiggfoi' metadata format, JIGG could gather and ingest them on a regular basis. Possibly an HEI could incorporate population of this OAI-PMH enabled database into the process of responding to FOI requests. If they were to publish their FOI disclosure log entries in this way, then submission to JIGG could become automatic.

# 4 Discussion

## 4.1 Incorporation into the JIGG Portal

The FOI disclosure logs application is just a part of the JIGG portal. The human user's view of the gateway is controlled by a Content Management System (CMS) to provide a consistent interface to all aspects, informed by a considered information architecture, and implemented by a JIGG-specific template and web style sheet. Thus it is necessary for the web search of the FOI disclosure logs, and their display to end users, to appear within the CMS, rather than as a separate, potentially inconsistent, application provided by the IESR-based implementation. This implies that the web interface to the FOI disclosure log entries will be provided by a server within the CMS. This server will maintain its data records by regularly gathering new records from the separate FOI disclosure logs application. OAI-PMH is the obvious choice for this data interchange, because of its capability for supplying new or changed records on a regular basis after an initial bulk data load.

## 4.2 Publication of FOI Disclosure Logs

So far this paper has focused mainly on the technical aspects of the JIGG FOI disclosure logs application. But there are, of course, social aspects. As yet few HEIs publish their disclosure logs. This reluctance may be simply because of insufficient staff resource. But there may be a lack of motivation because of a perception that there is no value in sharing this information. Or, further, there may be an active objection because of concerns about accountability.

The project hopes to encourage more HEIs to publish their FOI disclosure logs and to promote their publication in JIGG. One approach will be to hold workshops for HEI records management practitioners who are potential data contributors, to advertise and explain the facility. The JIGG project has engaged with, and has support from, a range of stakeholders, and has several UK regional Advisory Panels consisting of practitioners and representatives from relevant bodies. An 'Information Legislation and Management Survey' of HEIs [22], which portrayed their current handling of FOI requests was recently undertaken by JISC infoNet.

Hopefully, as JIGG is populated with a sizeable corpus of FOI disclosure log entries, the value of such a resource will become apparent. Publishing summaries of information released following FOI requests, and in some cases the full text of responses, will potentially reduce the number of requests for the same information. It will enable HEIs to share their experiences of responding to such requests. It should avoid 'reinventing the wheel' by individual HEIs as they consider aspects of legal compliance that apply to the whole sector. Currently they could potentially give differing responses. Thus the JIGG FOI database should both help to ensure a consistency of response to similar requests, and potentially reduce the resource requirements on records management staff. Essentially JIGG is providing a platform for accumulating and sharing 'frequently asked questions' and their answers.

# 5 Conclusions

At present these advantages of a central repository of HEI FOI disclosure logs are largely hypothetical. The JIGG project intends to provide the practical infrastructure to realise the vision as the project matures over the next eighteen months. But this does depend on engaging the participation of HEI records management practitioners.

One consideration, mentioned above, is obviously the effort required to supply FOI disclosure log entries to JIGG. The vision is a scenario where HEIs publish their FOI disclosure log entries in an OAI-PMH enabled database, incorporated into their business processes for dealing with FOI requests. JIGG would harvest these disclosure log entries into its central database on a regular basis. The use of OAI-PMH would remove the need for manual effort once the system is in place. But this scenario does imply knowledge of OAI-PMH and technical development capability by the HEI's administration department.

The experience of using an Application Profile within the JIGG project has shown it to be invaluable for developing and formally documenting a metadata schema. It proved to be an ideal format to assemble, communicate and discuss suitable properties during the process of gaining agreement, and for dissemination of the details to other interested parties. This was within a sector where there was not general awareness of metadata schemas and no previous knowledge of OAI-PMH. The Application Profile provides a clear specification even to those who are not conversant with metadata schemas. It affords a relatively 'syntax free' format understandable by non-technical people. The JIGG FOI Application Profile is a web document, so it includes hyperlinks between various sections and definitions, which hopefully enhance usability by readers. At the same time it is regarded as a formal specification with a persistent URI.

The JIGG project is utilising, and thus disseminating awareness of, OAI-PMH within a new sector. But the marriage of Open Archives and Freedom of Information seems apt. It is envisaged that sharing of FOI disclosure log entries may be broadened to other organisations beyond HEIs, if they adopt the JIGG FOI Application Profile. This interoperability is assisted by using a standards-based approach, in particular by employing OAI-PMH.

## Acknowledgements

The JIGG project is funded at The University of Manchester by the Joint Information Systems Committee (JISC) of the UK Higher and Further Education Councils, with technical development by MIMAS. The author wishes to acknowledge the contribution of project colleagues to the developments presented in this paper, in particular Steve Woods and Chris Watts previously at Liverpool John Moores University.

## Notes and References

[1] *The Freedom of Information Act 2000*. London : The Stationery Office Ltd., 2000. Retrieved, March 29, 2007, from http://www.opsi.gov.uk/acts/acts2000/20000036.htm

[2] *The Freedom of Information (Scotland) Act 2002*. Scotland : The Stationery Office Ltd., 2002. Retrieved, March 29, 2007, from http://www.hmso.gov.uk/legislation/Scotland/acts2002/20020013.htm

[3] WOOD, S. Freedom of Information & Open Government Blog Disclosure Log Index. 2007. Retrieved, March 29, 2007, from http://www.foi-directory.org/

[4] JIGG. JISC Information Governance Gateway. 2007. Retrieved, March 29, 2007, from http://www.jigg.ac.uk/

[5] APPS, A. JISC Information Governance Gateway: FOI Disclosure Logs. 2007. Retrieved, March 29, 2007, from http://www.jigg.ac.uk/foi/

[6] LAGOZE, C.; VAN de SOMPEL, H.; NELSON, M.; WARNER, S. The Open Archives Protocol for Metadata Harvesting. 2004. Retrieved, March 29, 2007, from http://www.openarchives.org/OAI/openarchivesprotocol.html

[7] *CWA 14855: Dublin Core Application Profile Guidelines*. Brussels : CEN/ISSS, 2003. Retrieved, March 29, 2007, from ftp://ftp.cenorm.be/PUBLIC/CWAs/e-Europe/MMI-DC/cwa14855-00-2003-Nov.pdf

[8] *DCMI Metadata Terms*. Dublin Core Metadata Initiative, 2006. Retrieved, March 29, 2007, from http://www.dublincore.org/documents/dcmi-terms/

[9] POWELL, A.; NILSSON, M.; NAEVE, A.; JOHNSTON, P. DCMI Abstract Model. 2005. Retrieved, March 29, 2007, from http://www.dublincore.org/documents/abstract-model/

[10] JISC. Function Activity Model. 2006. Retrieved, March 29, 2007, from http://www.jisc.ac.uk/whatwedo/themes/eadministration/recordsman_home/srl_structure.aspx

[11] *FOI Full Exemptions Guidance*. Department for Constitutional Affairs, 2006. Retrieved, March 29, 2007, from http://www.dca.gov.uk/foi/guidance/exguide/intro/annex_a.htm

[12] JISC, Freedom of Information (Scotland) Act 2002: Implementation and Practice: Exemptions. 2003. Retrieved, March 29, 2007, from http://www.jisc.ac.uk/publications/publications/pub_ib_fois.aspx#08exemptions

[13] *The Environmental Information Regulations 2004: 12, Exceptions to the duty to disclose environmental information*. London : The Stationery Office Ltd., 2004. Retrieved, March 29, 2007, from http://www.opsi.gov.uk/si/si2004/20043391.htm#12

[14] *The Environmental Information (Scotland) Regulations 2004: 10, Exceptions from duty to make environmental information available*. Scotland : The Stationery Office Ltd., 2004. Retrieved, March 29, 2007, from http://www.hmso.gov.uk/legislation/scotland/ssi2004/20040520.htm#10

[15] Creative Commons. Retrieved, March 29, 2007, from http://creativecommons.org/

[16] APPS, A.; WATTS, C.; WOOD, S. JISC Information Governance Gateway Freedom of Information Disclosure Logs Application Profile. 2006. Retrieved, March 29, 2007, from http://www.jigg.ac.uk/foi/profile/

[17] JOHNSTON, P.; POWELL, A. Expressing Dublin Core metadata using XML. 2006. Retrieved, March 29, 2007, from http://www.dublincore.org/documents/dc-xml/

[18] JOHNSTON, P.; POWELL, A. Expressing Dublin Core metadata using XML (DC-XML-Min). 2006. Retrieved, March 29, 2007, from http://dublincore.org/architecturewiki/DCXMLRevision/DCXMLMGuidelines/2006-09-18

[19] JOHNSTON, P.; POWELL, A. DC-Text: a simple text-based format for DC metadata. *DC2006: Proceedings of the International Conference on Dublin Core and Metadata Applications, 3-6 October 2006, Manzanillo, Mexico*. Mexico : University of Colima, 2006, p. 24-30.

[20] JIGG FOI XML Schema. Retrieved, March 29, 2007, from http://www.jigg.ac.uk/foi/schemas/xsd/jiggfoi.xsd

[21] APPS, A. Disseminating Service Registry Records. *ELPUB2006: Proceedings of the Tenth International Conference on Electronic Publishing, 14-16 June 2006, Bansko, Bulgaria*. Sofia : FOI-COMMERCE, 2006, p. 37-47.

[22] JISC infoNet. Information Legislation & Management Survey 2006 – Results. 2007. Retrieved, March 29, 2007, from http://www.jiscinfonet.ac.uk/foi-survey/2006/results

# EPrints 3.0: New Capabilities for Maturing Repositories

*Leslie A. Carr*

School of Electronics and Computer Science, University of Southampton, Southampton, United Kingdom
e-mail: lac@ecs.soton.ac.uk

## Abstract

There are now a large number of repositories in the world, contributing a significant amount of content to the world's scholars and scientists. The landscape has changed since the emergence of the Open Archiving Initiative: as well as Open Access, we have seen Multimedia Scholarly Collections, Teaching, Scientific Data, Preservation, Research Management and Assessment emerging as key drivers for repository adoption and development across the world. A new version of the EPrints repository software has been developed to address the growing demands on repositories to accommodate a wider variety of digital objects and metadata, to integrate with a wider range of services and applications and to support higher deposit rates to serve the needs of the whole institution. Already described by one reviewer as "a significant milestone towards ideal repository software", EPrints 3.0 provides new features for well-managed, high quality, high value repositories.

**Keywords:** repositories; EPrints; digital libraries

## 1 Introduction

There are now a large number of repositories in the world, contributing a significant amount of content to the world's scholars and scientists. The EPrints platform was developed as an outcome of the Open Archiving Initiative in 1999. This movement is still gaining momentum, as there are many institutions across the world that have yet to commit to providing a service for the dissemination and curation of their scientific and scholarly materials.

The landscape has changed since 1999: as well as Open Access, we have seen Multimedia Scholarly Collections, Teaching, Scientific Data, Preservation, Research Management and Assessment emerging as key drivers for repository adoption and development across the world.

EPrints has responded to the diversification of the repository ecology thanks to the flexibility and adaptability of its software platform (designed for easy customisation at every level) and its metadata and data management facilities (not limited to one schema or standard profile). Examples of EPrints repositories that support each of the agendas listed above can be seen at the Exemplars pages[1] at the EPrints web site.

In developing the latest version of EPrints (version 3.0) we responded to the community's requests for increased interoperability with other services and applications, improved user experience of depositing & managing a wider range of digital objects and better open source development opportunities. EPrints v3 is a more effective and efficient platform to underpin any kind of repository, and has recently been described as "a significant milestone towards ideal repository software" (Ariadne 50, January 2007).

All repositories aim to provide high quality information services to enable (or improve) administrative, scholarly or research tasks based on high quality records of research/scholarship. These tasks impose high standards of metadata and data curation – acquisition, transformation, maintenance and dissemination. Consequently repository managers face two challenges with their faculty staff: firstly in encouraging them to regularly deposit their research outputs or artefacts, and secondly in remedying deficiencies in the metadata provided.

The new EPrints software addresses the apparently irreconcilable issues of low-impact-deposit and high-metadata-quality by making data entry easier to get right. Firstly, a range of importers allow existing objects to be imported from other services or data sources. Secondly, when editing newly imported (or freshly created) objects, intelligent metadata assistance is provided for key information items. Auto-completion on the author names field means that a complete creator (surname, forenames, title and email address or staff ID) may be

[1] http://www.eprints.org/software/examples

entered in as little as three keystrokes and a menu choice. Not only does this lead to less effort for the depositor, it also means that author names are complete and consistently referenced throughout the repository. Other assistance means that the official journal name will always be used together with its ISSN and publisher and that the project identifier assigned by the item's funding agency can be relied on and that duplicate deposits are avoided.

Beyond these specific user interface features, the aim of EPrints is to fulfill three key objectives:

1. The *High Quality Repository*: a repository where the object metadata and object data is complete, correct and consistent – a repository where data entry is easy, errors and omissions are minimised at source and an ongoing quality management process is facilitated;

2. The *High Value Repository*: a repository whose items can be used and reused in many contexts for many tasks – not only dissemination and information discovery but also administrative reporting, bibliography management, CVs, institutional portals, the Semantic Web and desktop and webtop applications (such as Microsoft Office and Google Maps);

3. The *Well Managed Repository*: a repository that supports efficient curation and reporting of the objects that it manages, that supports effective monitoring and feedback of the workflow processes and operators who enact the workflow, and whose configuration and management is possible by the librarian manager rather than the technical system administrator.

As Open Access mandates from funding agencies and Research Assessment activities (both institutional and national) impose external deadlines with real financial consequences for failure to deliver then significant challenges emerge for staffing the repository processes that are needed to satisfy them, EPrints provides the solution for an enterprise repository that collects scholarly and scientific materials from thousands of its faculty staff and researchers without imposing onerous demands on individual researchers, library staff or information services.

# DCMI-Tools: Ontologies for Digital Application Description

*Jane Greenberg*[1]*; Thomas Severiens*[2]

[1] School of Information and Library Science, University of North Carolina in Chapel Hill, USA
e-mail: janeg@ils.unc.edu
[2] Institut für wissenschaftliche Information e.V., Universität Osnabrück, Germany
e-mail: severiens@mathematik.uni-osnabrueck.de

## Abstract

The growth in electronic and digital publishing on the World Wide Web has led to the development of a wide range of tools for generating metadata. As a result, it can be difficult to select the appropriate type of application and the best metadata tool to support a project's metadata needs. The Dublin Core Tools (DCMI Tools) Community recognizes this need and is developing an application profile and a taxonomy of tool functionalities for describing metadata applications. The community will use the application profile and the taxonomy to standardize access to information on metadata via the DCMI Tools and Software program. This paper reports on the DCMI Tool Community's activities to develop an application profile for describing the wide range of applications (algorithms; metadata templates, editors, and generators; and other software) fitting this rubric. The paper begins with an introduction to metadata application challenges, and introduces the DCMI Tools Community in order to provide important historical context. Next, the paper reviews the concept of application profile and emphasizes the importance of this approach for describing metadata tools. The paper reviews procedures to develop the application profile and presents the DCMI Tools application profile. The paper also presents a metadata tool functionality taxonomy (to be used with the application profile), a glossary (to assist people in learning about metadata tools), and the DCMI Tool Community's implementation plans. The final part of the paper presents several conclusions and highlights next steps.

**Keywords:** metadata tools; application profile; DCMI Tools Community

## 1 Introduction

Today's metadata tool environment includes offerings ranging from algorithms that plug in to various multi-functional software applications to fully developed tools specifically labeled as metadata editors, templates, and generators. Included in the mix are many software applications, such as word processing and publication software (e.g., Microsoft's WORD and Acrobats Adobe) and MP3 software that increasingly include functionalities supporting metadata generation. Tools in this category often include templates for storing summary metadata, such as "keywords" or "author name" or a brief "description". This type of software generally automatically generates a range of metadata, such as "date created", "date last modified", "size" and "format" [1]. There is also an evolution of blog software and social software (e.g., Flickr or Del.icio.us) supporting similar metadata generation, including tags. Metadata generated with any of these applications (designated metadata tools, software applications, and social software) can be harvested by metadata tools to create coherent or more substantial metadata records, which can be ported into a metadata repository to support resource discovery and other desired metadata functionalities [2].

Although these developments are exciting, they have complicated our view of the metadata tool landscape. That is the wide range and diversity of applications can make it difficult to select the appropriate type of application and the best metadata application to support a project's metadata needs. Should a digital library project invest in a fully functional off-the-shelf metadata generation application? What open source algorithms might be accessible that could be integrated with an institutions existing software suite to satisfy metadata needs? Catalogers, metadata professionals, information architects, and project managers are constantly asking these and other questions to determine which applications will suit their needs. Their inquiry is made difficult because of the absence of a single place providing unified and consistent descriptions of metadata tools.

The Dublin Core Tools (DCMI Tools) Community, a part of the Dublin Core Metadata Initiative (DCMI), is addressing this challenge [3, 4]. For the last several years this community has provided a Web page with access information and brief descriptions of applications supporting the generation of Dublin Core metadata records. As the metadata tool community has grown to include both developers and users, so too has the need to provide unified and collective information about metadata applications. The need expands beyond applications supporting Dublin Core metadata, to tools supporting metadata creation following:

- Standard schemes beyond the Dublin Core (e.g., ONIX or the EAD).
- Content value standards (e.g., *Library of Congress Classification* system) and authority files.
- Encoding schemes to standardize the use of content value standards even further (e.g., W3C Date Time Format standard).

The DCMI Tools Community is addressing this need via the development of an application profile and a taxonomy of tool functionalities—both of which can be used for describing metadata applications generally accessible for digital library and related initiatives.

This paper reports on the DCMI Tools activities to develop an application profile for describing the wide range of applications (algorithms; metadata templates, editors, and generators; and other software) fitting this rubric. The following sections of this paper are ordered as follows: section 2 introduces the DCMI Tools Community and provides some historical context; section 3 reviews the concept "application profile" and emphasizes why this approach supports a unified description of metadata tools; section 4 presents procedures to develop the current DCMI Tools application profile; section 4 presents the DCMI Tools application profile, section 5 presents a taxonomy of tool functionalities for classifying metadata applications, a glossary containing terminology that is important for the metadata tool community, and application profile implementation steps; section 6 includes several conclusions and next steps.

## 2 The DCMI Tools Community

The DCMI Tools Community is a "forum for individuals and organizations involved in the development and usage of tools and applications based on Dublin Core Metadata or other metadata standards that interoperate with and enhance functionality of the Dublin Core" [5]. The DCMI Tools community was initially a working group and was initiated at the 1999 Dublin Core conference in Frankfurt, Germany. The founding chairs were Roland Schwänzl (Osnabrück University) and Harry Wagner (OCLC). The working group initially focused on RDF-Tools and XML-Schema, as well as on DAML+OIL (which since that time has developed as SOAP web-services). At the outset, the DCMI Tools WG recognized the metadata community's need to access information about metadata applications. The Tools WG, therefore, took up the initiative of documenting and making accessible basic and important information about metadata applications via the DCMI Website through the "Tools and Software" program [6].

Although no formal descriptive standard was created to describe the applications, a broad taxonomy was developed to classify the range of applications being represented. Metadata tools being currently represented via DCMI's Tools and Software program are classed accordingly: Utilities, Creating Metadata (Templates), Tools for the Creation/Change of Templates, Automatic Extraction/Gathering of Metadata, Automatic Production of Metadata, Conversion Between Metadata Formats, Integrated (Tool) Environments, Application Profiles (Examples and Tools), and Metadata Search Engines. Details given for the tools represented via this site range from brief abstracts to more descriptive accounts documenting the metadata elements and schemes a tool supports.

During the Dublin Core 2006 conference in Manzanillo, Colima, Mexico, the DCMI Tools working group was transformed to what the DCMI refers to as a community [5]. The goal of a DCMI community is to facilitate the "exchange of information, general discussion within a specific area of interest" [3]. This change was very timely for the DCMI Tools WG, which had a year earlier Madrid, Spain, revised their charge to develop as a forum for two classes of users: tool developers and individuals interested in using tools. The DCMI Tools working group sponsored a workshop at the 2006 Joint Conference on Digital Libraries (JCDL), bringing together these users into a single community [7]. These developments, and the growing interest in metadata tools well beyond the immediate DCMI community, have motivated the reevaluation of the current classification of tools represented via the DCMI Tools and Software program [6]. This work has been a major focus of the DCMI Tools community via the last year, through a task group comprised of the DCMI Tools community co-chairs, with input from other members of the DCMI tools community. Our process of revision has required the creation of an application profile. The next section of this paper defines what an application profiles is, and why we selected this approach.

## 3 Application Profiles: A Practical Approach for Describing Metadata Tools

An application profile is a declaration of the metadata terms an organization, information resource, application, or user community uses in its metadata. In a broader sense, an application profile includes the set of metadata elements, policies, and guidelines defined for a particular application or implementation. The elements may be from one or more element sets, thus allowing a given application to meet its functional requirements by using metadata elements from several element sets including locally defined sets. For example, a given application might choose a specific subset of the Dublin Core elements that meets its needs, or may include elements from the Dublin Core, another element set, and several locally defined elements, all combined in a single schema, as for example the "Dublin Core Collection Description Application Profile" [8] does. An application profile is not considered complete without documentation that defines the policies and best practices appropriate to the application.

Application profiles are created for practical reasons. *First*, it makes no sense to reinvent the wheel. Why should a project create a metadata scheme from scratch, when there is already a scheme, or a series of schemes, that have already defined needed metadata elements, including implementation and use requirements? *Second*, it is recognized that often, a single scheme may not fully satisfy the needs of an individual initiative. For example, the Dublin Core metadata scheme is very useful for supporting resource discovery of digital resources in a digital library, although the elements do not adequately document and help manage resource preservation. A digital library wanting to facilitate the functions of both "resource discovery" and "preservation" might create an application profile, integrating elements from both the Dublin Core metadata standard and the PREMIS metadata standard [9]. *Third*, an application profile, pulling together elements from other schemes, facilitates greater interoperability on the World Wide Web. The reasons stated here, the growth in the availability of metadata tools, and the expansion of the DCMI Tools community, have motivated the development of the DCMI Tools application profile for representing algorithms, code pieces, and software tools. The next section of this paper presents a more detailed description on the methodology used generating the application profile.

## 4 Procedures for Developing the DCMI Tools Application Profile

The objective of our application profile activity is to describe algorithms, crosswalks, software, software-tools, and utilities collected in www.dublincore.org/tools/ in a coherent way. In moving forward, we have aimed to achieve best practices, resulting from discussion with participants from various fields. Our procedure has included the following steps:

1. An assessment of all elements in the available from the Dublin Core (ISO 15836-2003), DCTERMS (www.dublincore.org/documents/dcmi-terms/), and DOAP (Description of A Project) (usefulinc.com/doap/) and their applicability to the DCMI Tools community's goals. We selected these three schemes to cover the obvious needs for describing those applications being collected in our initial repository;

2. An initial three level ranking of each element's usefulness to our goals, with level *one* being necessary, *two* being potentially valuable *three* being not germane;

3. The composition of a DCMI Tools application profile, which included all level one ranked elements, and slightly over half of the level two items;

4. The development of a taxonomy of metadata tool functionalities—to be used with the application profile and for classifying metadata tools;

5. The development of a glossary to aide with tool classification and to facilitate communication among the metadata tool user community.

## 5 DCMI Tools Application Profile

The DCMI Tools application profile contains 17 elements, drawing from the Dublin Core, the DCTerms, and DOAP schemes. Nine of these elements contain qualifiers. Qualifiers can refine the meaning of an element, indicate where the value associated with an element came from, or the content formatting of an element (e.g, the format of year-month-date: YYYY-MM-DD). Table 1 presents an overview of our application profile, including examples for two applications, DC-dot and Picard Tagger.

| Name-space | Element | Qualifiers | Example DC-dot | Example Picard Tagger |
|---|---|---|---|---|
| dc | contributor | doap:maintainer<br>doap:developer<br>doap:documenter<br>doap:translator<br>doap:tester | Rachel Heery | developer: LukasLalinsky<br>developer: RobertKaye |
| dc | creator | | Andy Powell | |
| dc | date | dcterms:created<br>dcterms:dateCopyrighted<br>dcterms:modified<br>dcterms:issued | Created: 7 July 1997 | issued: 2006-06-25 |
| dc | description | | Extracts and validates metadata from HTML resources and MS Office files. The generated metadata can be edited using the form provided and converted to various other formats (USMARC, SOIF, IAFA/ROADS, TEI headers, GILS, IMS or RDF) if required. | PicardTagger allows you to automatically look up the releases/tracks in your music collection and then write clean metadata tags (ID3 tags, Vorbis comment fields, etc.) to your files. It also allow syou to specify how and where to write cleanly tagged files to your hard drive. |
| dc | identifer | doap:repository | http://www.ukoln.ac.uk/metadata/dcdot/ | http://musicbrainz.org/doc/PicardTagger<br>repository:<br>http://svn.musicbrainz.org/picard |
| dc | language | | en-us, en-GB | |
| dc | publisher | | | |
| dc | relation | dcterms:hasPart<br>dcterms:hasVersion<br>dcterms:isPartOf<br>dcterms:isReplacedBy<br>dcterms:isRequiredBy<br>dcterms:isVersionOf<br>dcterms:replaces<br>dcterms:requires<br>doap: release | requires: Libwww-perl, soif.pl, Jon Knight's MARC module | requires:<br>PyQt4<br>Mutagen (1.7)<br>python-musicbrainz2<br>isPartOf:<br>https://musicbrainz.helixcommunity.org/<br>release: 0.7.1 |
| dc | rights | dcterms:accessRights<br>dcterms:license | accessRights: open source<br>license:<br>http://www.gnu.org/copyleft/gpl.html | accessRights: open source<br>license:<br>http://www.gnu.org/copyleft/gpl.html |
| dc | rightsHolder | | | |
| dc | source | dcterms:URI | | Workman,<br>http://musicbrainz.org/doc/Workman |
| dc | title | dcterms:alternative | DC-dot | Picard Tagger |
| dc | type | dcterms:dataset<br>dcterms:InteractiveResource<br>dcterms:service<br>dcterms:software | dcterms:InteractiveResource | dcterms:software |
| dcterms | audience | dctools:developer<br>dctools:users | | dctools:users<br>dctools:developer |
| doap | location | | Bath, UK | |
| doap | programming-language | | Perl | Python |
| doap | operating-system | | | |

**Table 1: DCMI Tools Application Profile**

## 6 A Taxonomy of Metadata Tool Functionalities

The application profile can be implemented within a semantic web framework, with automatic processes and requires the use taxonomy terms wherever possible. This will improve the representation of objects described, allowing for fairly complete the metadata descriptions. The most important part of the application profile is the classification of objects by genre, represented in our taxonomy.

Every object described may be in one or more of the following classes, which allows for sorting of tools by functionalities:

- Conversion
- Crosswalk
- Metadata Creation
- Metadata Encoding
- Metadata Extraction
- Metadata Generation
- Metadata Harvesting
- Metadata Templates
- Search Engines
- Translation
- Transliteration
- Validation

We will extend these classes as new types of software are developed. Classes not filled with latest software will be deleted, and the list will be revised as needed to allow for appropriate growth. We see this lists as being organic—in order to meet the needs of the tools community over time.

Some still open questions remain as part of our work in developing the profile. For example, location information requires additional attention. The most useful and precise approach is to give geographical coordinates, so a service can link to map serves. An alternative approach is to use a controlled vocabulary for geographic names. In this case, it would be desirable to allow for access and linking via international names (e.g., "Wien" (German version) versus "Vienna" (English version) versus "Wenen" (Dutch version). For the agent roles in the application profile we tried to use the roles defined in DOAP namespace (usefulinc.com/doap/) mostly reused from the foam-project results:

- developer
- documenter
- maintainer
- tester
- translator

To re-use the collected information in multiple frameworks, it will be requested to clearly define all vocabulary used. For use in semantic web framework this will be offered as RDFS, for human readability we restrict to textual representation in this article.

To assist with our work and further bring the metadata tool user community together, we have also developed a Glossary. This is presented in Table 2. The glossary is a new development produced by the DCMI Tools Community, and will be enhanced and modified as we continue our work.

**Algorithm**
a finite set of well-defined instructions for accomplishing some task which, given an initial state, will terminate in a defined end-state. (Wikipedia)

**Application Profile**
an assemblage of metadata elements selected from one or more metadata schemas and combined in a compound schema. Application profiles provide the means to express principles of modularity and extensibility. The purpose of an application profile is to adapt or combine existing schemas into a package that is tailored to the functional requirements of a particular application, while retaining interoperability with the original base schemas. Part of such an adaptation may include the elaboration of local metadata elements that have importance in a given community or organization, but which are not expected to be important in a wider context. (Duval)

**Conversion**
can refer to either

- conversion between schemas
- conversion of encoding (x/html to xml)

**Crosswalk**
a semantic mapping of metadata elements across metadata schema specifications. Crosswalks permit searching across multiple databases that use different schemas (Greenberg)

**Metadata**
An item of metadata may describe an individual data item or a collection of data items. Metadata is used to facilitate the understanding, use and management of data. (Wikipedia)

**Metadata Creation**
creation of metadata can be either

- by professional metadata creators; these include catalogers, indexers, and database administrators
- by technical metadata creators; these include webmasters, data in-putters, paraprofessionals, encoders and other persons who create metadata and may have had basic training but not professional level training
- by content creators; people who create the intellectual content of an object and the metadata for that object
- by community / subject enthusiasts; people who have not had any formal metadata-creation training but have special subject knowledge and want to assist with documentation (Greenberg)

**Metadata Encoding**
the syntax or prescribed order for the elements contained in the metadata description (NISO)

**Metadata Extraction**
synonym to Metadata Harvesting

**Metadata Generation**
the act of creating or producing metadata. Metadata can be generated by people, tools and processes (Greenberg)

**Metadata Harvesting**
a technique for extracting metadata from individual repositories and collecting it in a central catalog (NISO)

**Metadata Template**
Metadata format designed for some specific use or subject. (Severiens)

**Namespace**
In XML, a namespace is a collection of names, identified by a URI reference, that are used in XML documents as element types and attribute names. In order for XML documents to be able to use elements and attributes that have the same name but come from different sources, there must be a way to differentiate between the markup elements that come from the different sources. (Webopedia.com)

**Schema**
In general terms, any organization, coding, outline or plan of concepts. In terms of metadata, a systematic, orderly combination of elements or terms. In terms of DCMI term declarations represented in XML or RDF schema language, schemas are machine-processable specifications which define the structure and syntax of metadata specifications in a formal schema language. In terms of an encoding scheme, is a set of rules for encoding information that supports a specific community of users. See also Encoding scheme. (DCMI)

**Search Engine**
A utility capable of returning references to relevant information resources in response to a query. (DCMI)

**Software**
consisting of programs, enables a computer to perform specific tasks (Wikipedia)

**Software-Tool**
small piece of software, designed for developmental and laboratorial use (Severiens)

**Translation**
the interpretation of the meaning of a text in one language and the production, in another language, of an equivalent text that communicates the same message. Translation between may also convert meaning between semantics or schemes. (Wikipedia, Severiens)

**Transliteration**
Conversion of names or text not written in the roman alphabet to roman-alphabet form. (AACR Glossary)

**Utility**
software program that functions for a particular purpose. (Wikipedia)

**Validation**

- validating that syntax of element contents is correct (e.g. YYYY-MM-DD)
- validating the encoding (e.g., XML)

**Table 2: DCMI Tools Glossary**

The database, from which www.dublincore.org/tools is being generated, contains the following structure:

- Title: corresponding to the dc.title field in the app. profile.
- URL: corresponding to the dc.identifier field in the app. profile.
- Description: corresponding to the dc.description field in the app. profile.
- Classification: used to sort the service into the different classes.
- Free/commertial: this field is corresponding to the dc.rights qualifier dcterms:accessRights
- Online/download/webservice: corresponding to the dc.type field in the app. Profile and its qualifiers dcterms.InteractiveResource / dcterms.software / dcterms.service, a tag for dcterms.dataset may be added, if an entry is being included into the database.
- Country: corresponding to the field doap.laocation.
- Comment: This field allows some free text comments.
- Provider: corresponding to the dc.publisher field in the app. profile

Based on application profile developments, our plan is to add the following fields to the database:

- Information on the contributors, which can be
  - developers
  - documenters
  - maintainers
  - testers
  - translators
- Information on the creator(s)
- Information on the dates associated with the object, like
  - the date of its creation,
  - date of its latest modification,
  - date it was issued,
  - or the date of its copyright notice
- Information on the language of the object
- Information on the relations of the object to other objects in the database
- Information on the license like
  - a link to the licence text,
  - information on the licence holder,
  - while the date of the licence was already given with the dates above.
- Information on the source, if they differ from the compiled resource
- Information on the used programming language, if a source is available
- Information on the operating systems requested for running the software, if its not an webservice or online service.

## 7 Conclusions and Next Steps

The experience of creating the DCMI Tools application profile has been fruitful and resulted in an application profile that is ready for implementation. The DCMI Tools Community will be meeting at the DCMI-2007 Conference in Singapore this August to update members on this work. Prior to this conference, we will be testing the application profile and revising the DC Tools and Software program [6]. Our implementation will allow us to evaluate the overall effectiveness of the DCMI Tools application profile and identify areas requiring attention and revision. We will use our time in Singapore to share our findings and discuss any other outstanding issues, such as integrating location vocabulary from doap:location field. We will then begin to work on a collection and maintenance policy plan for keeping the DC Tools program up-to-date.

## Acknowledgements

We would like to acknowledge Tessa Sullivan, doctoral students, School of Information and Library Science, University of North Carolina at Chapel Hill, USA, for her work on creating a table for the DC Tools Application Profile.

## Notes and References

[1] GREENBERG, J.; SPURGIN, K;.CRYSTAL, A. Functionalities for Automatic-Metadata Generation Applications: A Survey of Metadata Experts' Opinions. International Journal of Metadata, Semantics, and Ontologies, (2006) 1(1): 3-20.

[2] GREENBERG, J. Understanding Metadata and Metadata Schemes. Cataloging & Classification Quarterly, (2005) 40(3/4): 17-36.

[3] DCMI Tools Community: http://dublincore.org/groups/tools/.

[4] Dublin Core Metadata Initiative: http://dublincore.org/.

[5] DEKKERS, M. Operational aspects of DCMI Work structure: Communities and Task. Date Issued: 2006-12-18: Groups: http://dublincore.org/documents/workstructure/.

[6] Tools and Software: http://dublincore.org/tools/.

[7] GREENBERG, J.; SEVERIENS, T., Metadata Tools for Digital Resource Repositories, D-Lib Magazine, (2006), Volume 12, Number 7/8, DOI:10.1045/july2006-greenberg.

[8] Dublin Core Collection Description Application Profile: http://www.ukoln.ac.uk/metadata/dcmi/collection-application-profile/.

[9] PREMIS: Data Dictionary for Preservation Metadata: http://www.oclc.org/research/projects/pmwg/premis-final.pdf.

# DRIVER – Supporting Institutional Repositories in Europe

*Mary L. Robinson*[1]*; Wolfram Horstmann*[2]

[1] SHERPA, Information Services, University of Nottingham
Greenfield Medical Centre, Medial School, QMC, Nottingham, UK.
e-mail: mary.robinson@nottingham.ac.uk;
[2] Research and Development Department, SUB Göttingen
Universität Göttingen, Göttingen, Germany
e-mail: whorstmann@sub.uni-goettingen.de

## Abstract

This workshop will provide an analysis of the current state of development of institutional repositories across Europe, how this compares to initiatives in the rest of the world and will explain how the DRIVER [1] project will promote and support the development of an integrated European repository network. The workshop will provide information on DRIVER technological developments and services and on the DRIVER test bed of repositories being used to test DRIVER software and services. The success of the DRIVER project depends, not just on the technical integration and enhancement of a European repository network, but also on the involvement and participation of all those actively involved in European research or in its publication, dissemination or access. Hence DRIVER has an active advocacy and community building programme to address and support key stakeholder groups in Europe. DRIVER draws on existing services within the DRIVER partnership such as OpenDOAR [2] and SHERPA/RoMEO [3] as well as developing new services such as the Mentor service [4]. This workshop will be of value to all involved in European research and for those keen to play a role in its future development. It will be of particular interest to those involved in the development of individual repositories, those co-ordinating national repository networks and those interested in the implications of a European repository network for European research. The workshop will provide a unique opportunity to learn about the DRIVER project, to meet DRIVER representatives, share best practice and discuss the current trends in the development and future of institutional repository networks.

**Keywords:** open access; repositories; European research

## 1 Introduction

The current system of academic publication developed as a means to disseminate the findings of research. However, this system can impede the very process it was set up to serve, with access to articles being limited by publishers to only those who can afford to subscribe.

Open access digital repositories provide a means whereby the traditional publishing model can co-exist with the needs of authors and their readers, as well as with the demands of research funders for research impact and hence, value for money. Subject to copyright, authors can deposit copies of their finished articles in open access repositories, in addition to publishing them in research journals.

The recent study of scientific publication markets in Europe funded by the European Commission [5] strongly recommends the development of a European policy mandating open access to EC-funded research. In addition, it recommends an exploration of interoperability issues and how open access repositories can be implemented Europe-wide.

DRIVER (Digital Repository Infrastructure Vision for European Research) is an EU-funded project with 10 international partners and reflects the growing awareness in Europe surrounding Open Access. DRIVER sets out to build a testbed for a future knowledge infrastructure of the European Research Area. It aims to deliver any form of scientific output, including scientific/technical reports, working papers, pre-prints, articles and original research data to the various user groups. The testbed is based on existing nationally organized digital repository infrastructures. Other work includes the support of new European repositories.

## 2 Objectives

The five objectives of DRIVER are:

1. To organise and build a virtual, European scale network of existing institutional repositories;
2. To assess and implement state-of-the-art technology, which manages the physically distributed repositories as one large scale virtual content resource;
3. To assess and implement a number of fundamental user services;
4. To identify, implement and promote a relevant set of standards;
5. To prepare the future expansion and upgrade of the DR infrastructure across Europe and to ensure widest possible involvement and exploitation by users.

## 3 Discussion

Thus far DRIVER has conducted focused research studies including an inventory of the type and level of OAI compliant digital repository activities in the EU [6], to facilitate the iterative development of DRIVER and is developing the necessary infrastructure middleware and user guidelines to meet the DRIVER objectives. The project is now actively advocating repository development - creating an informed and active environment for repository infrastructure development in EU countries with focused activities, information and contextualized support.

## Acknowledgements

The authors would like to thank Bill Hubbard and Sophia Jones of SHERPA for their assistance in the preparation of this paper.

## References

[1] DRIVER (Digital Repository Infrastructure Vision for European Research),
http://www.driver-support.eu

[2] http://www.opendoar.org

[3] http://www.sherpa.ac.uk/romeo.php

[4] http://www.driver-support.eu/en/community/mentor.html

[5] European Commission. Study on the economic and technical evolution of the scientific publication markets in Europe, Jan. 2006.
http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf

[6] DRIVER. Inventory study into the present type and level of OAI compliant Digital Repository activities in the EU, Apr. 2007.

# Cultural Content Management at a New Level: Publishing Theater and Opera Details by Means of Open Technologies from the Web 2.0

*Markus W. Schranz*[1,2]

[1]Distributed Systems Group, Institute of Information Systems, Vienna University of Technology
Argentinierstrasse 8/184-1, A 1040 Vienna, Austria
e-mail: schranz@infosys.tuwien.ac.at
[2]Research Development, Culturall Handelsges.m.b.H.
Graf-Starhemberg-Gasse 37/4, A-1040 Vienna, Austria
e-mail: markus.schranz@culturall.com

## Abstract

Creating Internet services for a specific auditorium involves technical, organizational and sociological challenges to developers, processes and used technologies. Authoring and distributing recent cultural news, opera schedules, notes to theater visitors or even cultural service maintenance can be handled by modern electronic publishing systems, ideally finding user-friendly solutions to the above-mentioned challenges. Although the Web has greatly improved its level of interactivity within the first ten years of existence, a significant gain in usability and value has been reached by introducing concepts summarized in the term Web 2.0. In this paper we provide an example of on-the-edge technology in managing and publishing cultural contents for Internet services, focusing on content management and workflow management features rom within the entire software framework concept. We demonstrate, that modern Web 2.0 technologies are well suited to increase the quality of electronic publishing for both the consumers as well as the producers of the content and the service itself by providing rich user experience enhancements at different levels of the content and service management process.

**Keywords:** content management; cultural content; Web 2.0; technical framework

## 1 Introduction

Most up-to-date software technologies and new applications in the Internet have initiated a change in the perception of what was generally understood as the „Internet". The first decade of the World Wide Web was dominated by a strict role distribution between few content providers and experts that used complex technical mechanisms and powerful tools to publish „centralized" data to the mass of content consumers, far away from open content interchange. As stated in [1] the shift to a new class of Internet applications has brought innovation in both technical and practical use of software applications for the Internet. Without being able to fasten it down to a single event or technology we sense a significant change in how applications provide information management, distribution and communication control to end users.

Increasingly, local proprietary solutions have been exchanged by open network services, desktop software is extending to integrated network solutions, programs and applications are serviced and updated in an open and self-contained way, single services are becoming ready for exchange and interoperability and even technical laymen are ready to use modern services to share information and facilities in an easy way. Despite the criticism in terms of the technological step contributed [2], Web 2.0 has become a synonym for openness, innovative technologies and user-friendly applications to integrate the abilities of all web users. Consequently and as described in the following sections also for dedicated services and technologies for specific user communities, media and content, the innovative character of the Web 2.0 can be identified in the area of electronic publishing.

Following thorough research basics and experiences from Web application engineering methodologies [3], the logically consecutive step is to construct user-oriented modern applications for specific application domains. Public contents in the area of science, education, news [4] or culture have been utilizing modern technologies to be distributed in a wide-spread manner. Resulting services are exemplary for innovation in the creation and consumption of Internet technologies for the discussed areas, e.g. online current contents for the Vienna Opera House [5]. Modern concepts like agile programming [6] and the utilization of dynamic languages, which were smiled at by software engineers a few years ago, have been the recent choice of innovators to create open services and access to digital publishing and content distribution.

In the following sections we focus on the innovative way to electronically publish and distribute cultural contents by facilitating modern technologies, integrating strong support of the collective intelligence as a basis of the Web 2.0. Since data and information are treated as the most central good, we describe the management, contribution and provision (publishing and distribution) of cultural contents such as theater programs and schedules, the ticket management and innovative clearing and control services for several theaters and opera houses in Europe, as well as the workflow coordination of the software engineers, working on the electronic publishing services that are based on modern standards and open content exchange technology. The paper outlines Web 2.0 technologies for the discussed application domain in section 2, specific content and service management approaches in section 3 and gives a brief summary of our results in the conclusion.

## 2 Modern Web 2.0 Technologies for Cultural Content Presentation

The term Web 2.0 is describing rather vaguely an updated perception and utilization of the World Wide Web. Renowned experts in the area of software development and Internet technologies criticize the Web 2.0 to offer little innovation in terms of technical development. For the use in the specific application domain cultural content management and publication we focus on the organizational view of Web 2.0: users create and manage contents in an increasing amount on their own. User-oriented Web interfaces facilitate simple theater and opera detail publishing and modern technologies support the content management, asset handling and schedule interchange based on open methodologies and standards such as RSS.

Cultural content as discussed in this paper include opera and theater programs, event and performance details in text and multimedia presentation, event schedules, ticket management and presentation, event access management and reservation services for single theatres, multi-client cultural organizations or several independent theatres and opera houses in Europe. As demonstrators we explain case studies of the modern implementations for the Hamburgische Staatsoper, Vienna State Opera and Symphonic Orchestra of Bern. These and a dozen more cultural providers are managed by Culturall, a technology innovator strongly supporting the research on Web 2.0 application technology for cultural content management and publishing.

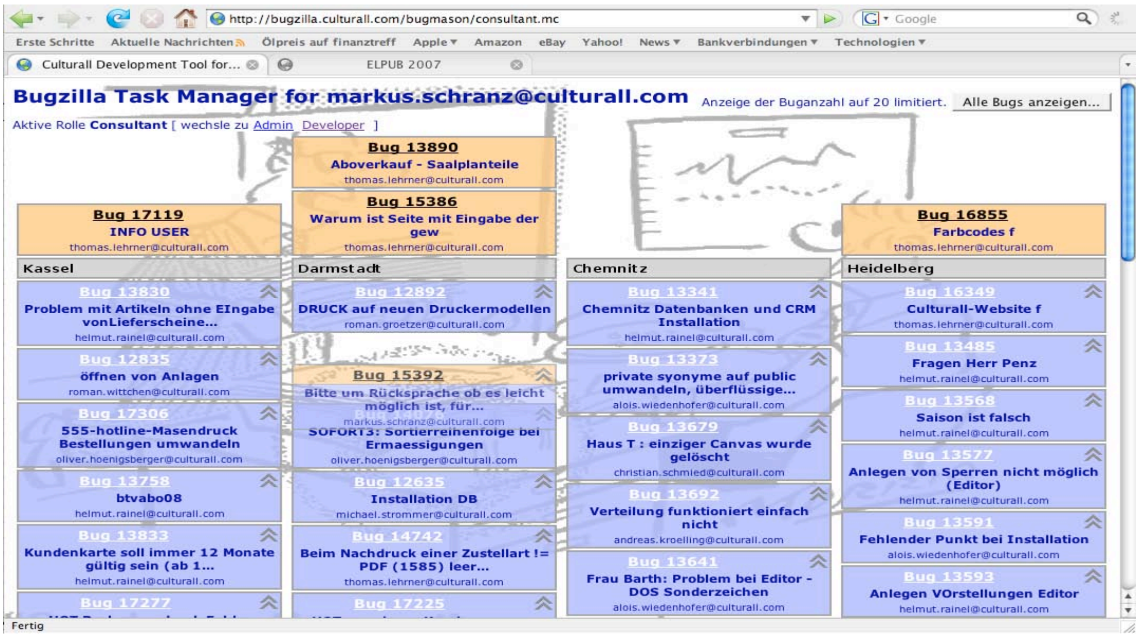
Most stimulative to the introduction of Web 2.0 technology in the cultural content management domain has been the important principle of supporting a Rich User Experience in the dedicated applications. The goal of software following the RUE principle is the creation of graphical interfaces that allow a handling that is comparable to that of local/desktop software implementations. Specific details of cultural content assets have to be exchanged frequently between user client and the providing server, thus interrupting the flow of the users visit by well-known brakes in between single Web pages. Based on the herein introduced approach we have developed a prototype that handles theater event descriptions, date scheduling, seat reservation and personalized ticket management has been integrated in a smooth and user-friendly way, following the basic principles of Web 2.0. Since all details of the concepts and a full description of the prototype is out of scope of this paper, we demonstrate the task management feature exemplarily to proof the usability and beneficial effects of Web 2.0 technologies in the application domain electronic pubishing.

Besides the technological advantages, the rich utilization of Web 2.0 technologies supports the researchers in extending the sense of community within the multiple users of the theater content management services. Since the information and the data/assets itself denote the highest value of the cultural services, content provision, management and publishing has to be simplified to a maximum extent. With the participation of all users, including the administrators, the theater experts, the content visitors, and the ticket bookers, the information is shared in a technologically well supported open way. Modern user-friendly content management interfaces as well as standardized open content exchange interfaces, such as RSS feeds for event schedules and ticket reservation assets for third party providers underline the open approach of the theater content management service in use.

## 3 Content Management and Digital Asset Management in Commercial Services

Based on the innovative application development methodologies outlined in the principles of Web 2.0 the content management and publishing services researched and developed for the cultural application domain includes mainly web-based services and database integration. Web software has been developed using dynamic development languages, which are well accepted in the Web 2.0 Lightweight Programming Models and the Agile Programming paradigms [7]. Based on open interfaces like AJAX [8] the web services have been developed and are currently under research investigations in terms of user-acceptance and scalability checks in current field studies. Particularly, the user interaction at the workflow management component of the cultural content publishing services is shown in Figure 1. Herein, a content developer can move a particular task (bug 15392) via Web 2.0 technology b simply dragging and dropping the item on the canvas. The service will not reload the page but instead send the alternated order via a ajax call to the server, where the new order is stored persistently in the database. With similar features, the visitors can pick theater and opera house seating, comment on published texts, etc.

The software implemented with ajax technology, object-oriented web application servers like Mason [9] provide access via user interfaces beyond device borders, so cross-platform and cross-device applications allow an open access to cultural contents via desktop computers, notebooks, handhelds and mobile phones. Furthermore the utilization of dynamic development languages like perl and java underline the principle of overcoming the software lifecycle in Web 2.0 applications. Instead of delivering version after version of a desktop application the cultural content management application is provided as a service that is under permanent development. Following this principles allow easier software maintenance and wide-spread service availability for a great mass of users. Providing services for a wide user group and offering open interfaces such AJAX/XML links and RSS feeds enables a distributed service enrichment similar to open source development. Specific features contributed by domain experts are demonstrated on the publicly available services [5].



Bugzilla Task Manager for markus.schranz@culturall.com Anzeige der Buganzahl auf 20 limitiert.

Aktive Rolle Consultant [ wechsle zu Admin Developer ]

Bug 17119
INFO USER
thomas.lehrner@culturall.com

Kassel

Bug 13830
Problem mit Artikeln ohne EIngabe vonLieferscheine...
helmut.rainel@culturall.com

Bug 12835
öffnen von Anlagen
roman.wittchen@culturall.com

Bug 17306
555-hotline-Masendruck Bestellungen umwandeln
oliver.hoenigsberger@culturall.com

Bug 13758
btvabo08
helmut.rainel@culturall.com

Bug 13833
Kundenkarte soll immer 12 Monate gültig sein (ab 1...
helmut.rainel@culturall.com

Bug 17277

Bug 13890
Aboverkauf - Saalplanteile
thomas.lehrner@culturall.com

Bug 15386
Warum ist Seite mit Eingabe der gew
thomas.lehrner@culturall.com

Darmstadt

Bug 12892
DRUCK auf neuen Druckermodellen
roman.groetzer@culturall.com

Bug 15392
Bitte um Rücksprache ob es leicht möglich ist, für...
markus.schranz@culturall.com

SOFORT3: Sortierreihenfolge bei Ermaessigungen
oliver.hoenigsberger@culturall.com

Bug 12635
Installation DB
michael.strommer@culturall.com

Bug 14742
Beim Nachdruck einer Zustellart != PDF (1585) leer...
thomas.lehrner@culturall.com

Bug 17225

Chemnitz

Bug 13341
Chemnitz Datenbanken und CRM Installation
helmut.rainel@culturall.com

Bug 13373
private syonyme auf public umwandeln, überflüssige...
alois.wiedenhofer@culturall.com

Bug 13679
Haus T : einziger Canvas wurde gelöscht
christian.schmied@culturall.com

Bug 13692
Verteilung funktioniert einfach nicht
andreas.kroelling@culturall.com

Bug 13641
Frau Barth: Problem bei Editor - DOS Sonderzeichen
alois.wiedenhofer@culturall.com

Bug 16855
Farbcodes f
thomas.lehrner@culturall.com

Heidelberg

Bug 16349
Culturall-Website f
thomas.lehrner@culturall.com

Bug 13485
Fragen Herr Penz
helmut.rainel@culturall.com

Bug 13568
Saison ist falsch
helmut.rainel@culturall.com

Bug 13577
Anlegen von Sperren nicht möglich (Editor)
helmut.rainel@culturall.com

Bug 13591
Fehlender Punkt bei Installation
alois.wiedenhofer@culturall.com

Bug 13593
Anlegen VOrstellungen Editor
helmut.rainel@culturall.com

**Figure 1: Task Management feature of the Web 2.0 cultural content management service prototype**

# 4 Conclusion

In this paper we describe the research and development of a content management and publishing service for a specific application domain: theater and opera content management based on innovative Web 2.0 concepts and technologies. The shift to a new class of Internet applications has brought innovation in both technical and practical use of software applications for the Internet.

Beyond the thoroughly discussed area of technical innovation contributed by the Web 2.0 we especially extend the organizational principles and sociological aspects of this new direction in software development for the Internet. As a demonstrator application we have researched, conceptualized and implemented a framework for cultural content and asset management, including the publishing and distribution of theater event details, ticket assets and reservation management services, and a highly sophisticated workflow management service for the software development process. Modern principles such as the rich user experience for GUIs, end user integration for content management and open interfaces to exchange domain specific contents are well adaptable for the application domain of cultural content providers. Modern application frameworks can be well applied to other domains which we investigate in future work.

## References

[1] O'REILLY, T., What Is Web 2.0?, O'Reilly Network, http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[2] BERNERS-LEE, T., developerWorks Interview podcast series, http://www-128.ibm.com/developerworks/podcast/dwi/cm-int082206.txt, July 28, 2006

[3] KAPPEL G, Pröll B, Reich S. Web Engineering. The Discipline of Systematic Development of Web Applications, Wiley & Sons, 12 May 2006

[4] SCHRANZ M., Pushing the quality level in networked news business: semantic-based content retrieval and composition in international news publishing, proceedings of the Elecronic Publishing conference, Bansko, Bulgaria, June 2006

[5] CULTURALL, Opera2007 - Vienna State Opera, https://www1.culturall.com/mirror_ctn/plsql/ctn_suche.spielplan, 2005

[6] THOMAS D., Heinemaier-Hansson D, Breedt L, Agile Web Development, Pragmatic Programmers, Dezember 2006

[7] FOWLER, Martin, The New Methodology, http://www.martinfowler.com/articles/newMethodology.html, 2005

[8] W3SCHOOLS, AJAX Tutorial, http://www.w3schools.com/ajax/default.asp, 2006

[9] ROLSKY, D. et al., Embedding Perl in HTML with Mason, O’Reilly, Oktober 2002

# Ontologies at Work:
# Publishing Multilingual Recreational Routes Using Ontologies

*Bert Paepen*

Centre for Usability Research, Katholieke Universiteit Leuven
Parkstraat 45 bus 3605, B-3000 Leuven, Belgium
e-mail: *bert.paepen@soc.kuleuven.be*

## Abstract

Even though there is nothing new about the idea, ontologies are a hot topic. Built for many reasons and appliances, the use of ontologies in real-life applications remains limited. The WalkOnWeb project has developed ontologies in the area of recreational routing and applied them in a real application. This demonstration will show these applications and explain how they use ontologies. With the "Walk Planner" hikers can plan their trip by looking for trails, creating new routes and getting detailed information in print, web or mobile format. Authors can create and describe routes using the "Authoring Tool". By creating ontologies and using them in these applications the WalkOnWeb project has developed a system to publish electronic routes in a flexible and personalized way.

**Keywords:** ontology; navigation; XML; SVG

## 1 Introduction

Ontologies are well-structured representations of knowledge in a certain domain. They consist of concepts and relations between them, described in a computer readable form while still being descriptive for humans. Applied in many areas, their use often remains theoretical, so that the practical utility of ontologies in real applications remains unclear. The European research project WalkOnWeb [1] has tried to break with this tradition by applying ontologies in a real-life application area: outdoor navigation.

## 2 Publishing Recreational Routes Electronically

One of the problems the WalkOnWeb application is trying to overcome is the lack of flexibility offered to outdoor enthusiasts by traditional publications. Hiking guidebooks for example describe a route in only one direction, one language and from a fixed starting point. When switching to an electronic publishing paradigm a hiker can expect more flexibility: she should be able to get information about itineraries via the Internet, whichever country she is visiting. She expects to get information in her own language, to choose her preferred starting point and walking direction, and maybe even combine parts of existing routes to a new, personalized route.

For publishers this type of requirements poses huge challenges when publishing their material in an electronic form. In theory electronic publishing should be more cost effective, avoiding high fixed costs for printing books. In practice however the issue is not that simple. First, users expect to get up to date information, forcing publishers to continuously provide updates. Second, multilingual publishing involves high translation costs. Finally, the material used for paper publications does not support the type of flexibility expected in electronic publishing, both in terms of content and technology.

## 3 Information Model

Taking these considerations into account WalkOnWeb has defined a new information model for flexible electronic publishing of recreational routes. A walk ontology was developed for this purpose. Using an innovative software engineering process this ontology was then converted automatically to Java business objects and mapped to a relational database. This paved the way for practical application development.

The figure below depicts the information model developed during the project. Using topographic map data as a basis an author creates a networks of paths on which hiking is possible, nice, safe, legal, etc. The author then enriches the map data further by linking all kinds of information to a route: points of interest, pictures, texts, and others. The *walk ontology* includes concepts for many of these information items. For example: practical info could be “keep dogs on leash”, “hunting season” or “dangerous crossing”.

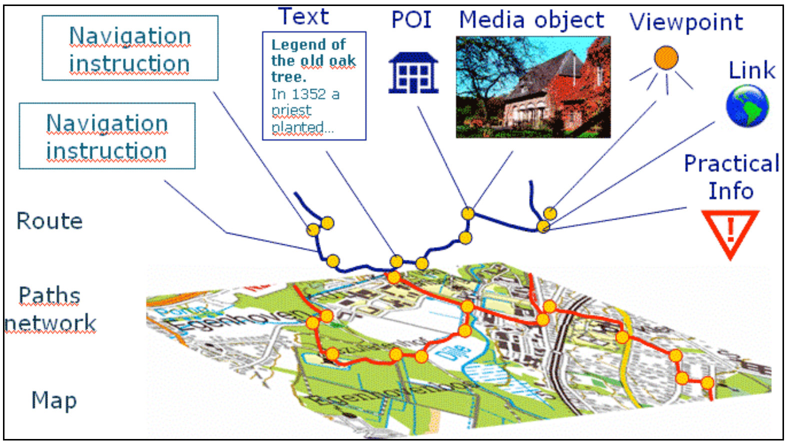
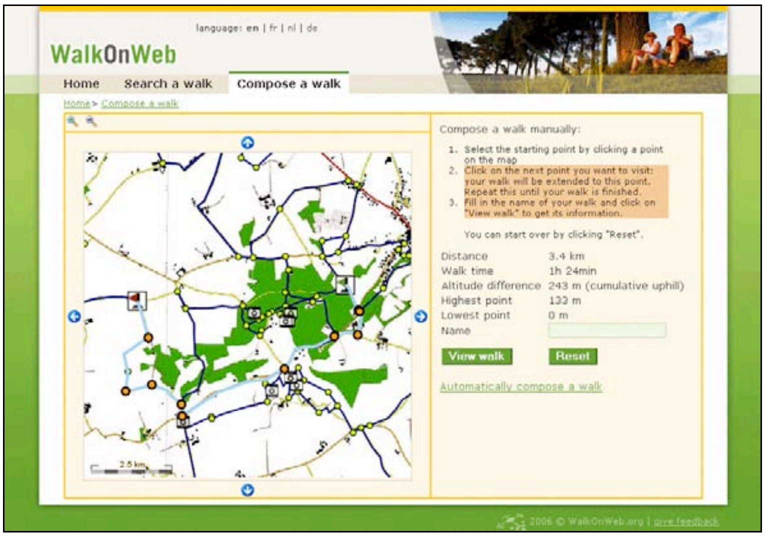


**Figure 1: WalkOnWeb information model**

In addition a navigation ontology was created to allow a new way of describing navigation instructions. Taking into account costs involved in traditional multilingual publishing this approach allows authors to describe a route in a language independent way, using a set of predefined ontology concepts (“building blocks”). We have first developed this ontology in theory (described in [2]) and then brought it to practice in two applications: the Walk Planner and the Authoring Tool.

## 4 Real-life Applications

On the "Walk Planner" website hikers can search for hiking trails using criteria like duration, difficulty, child friendliness and geographic location. All walks are shown on a geographic map using SVG. Hikers can also compose their personal walk based on parts of existing trails (see Figure 2, where the user has composed a walk from the green flag up to the red flag). Finally they can export detailed information for a walk to paper, electronic document or mobile device. This publication happens on the fly, based on user preferences: language, type of information, starting point and walking direction. For navigation instructions this means that the system generates a readable text from the navigation ontology concepts that the author has selected.


language: en | fr | nl | de
WalkOnWeb
Home
Search a walk
Compose a walk
Home > Compose a walk
View walk
Reset
Automatically compose a walk
2006 © WalkOnWeb.org | give feedback

Compose a walk manually:

1. Select the starting point by clicking a point on the map
2. Click on the next point you want to visit: your walk will be extended to this point. Repeat this until your walk is finished.
3. Fill in the name of your walk and click on "View walk" to get its information.

You can start over by clicking "Reset".

| | |
|---|---|
| Distance | 3.4 km |
| Walk time | 1h 24min |
| Altitude difference | 243 m (cumulative uphill) |
| Highest point | 133 m |
| Lowest point | 0 m |
| Name | |

**Figure 2: Walk Planner: compose a walk**

With the "Authoring Tool" application authors are able to create and describe routes in a language independent way. This means that they apply the model depicted in Figure 1 and use ontology concepts for describing the details of the itinerary.

## Notes and References

[1] PAEPEN, B et al. *WalkOnWeb project website*. www.walkonweb.org Leuven, 2006.

[2] PAEPEN, B; ENGELEN, J. *Using a Walk Ontology for Capturing Language Independent Navigation Instructions.* In ELPUB2006. Digital Spectrum: Integrating Technology and Culture - Proceedings of the 10th International Conference on Electronic Publishing held in Bansko, Bulgaria 14-16 June 2006 / Edited by: Bob Martens, Milena Dobreva. ISBN 978-954-16-0040-5, 2006, pp. 187-196

# The PURE Institutional Repository: Ingestion, Storage, Preservation, Exhibition and Reporting

*Bo Alroe*

Atira A/S and 10 Danish Universities, Niels Jernes Vej 10, NOVI, DK-9220 Aalborg OE, Denmark
e-mail: ba@atira.dk

## Abstract

Jointly developed over 5 years by Atira A/S and a number of university libraries, the commercial repository system PURE is a tool in the research administration and dissemination effort at 11 Danish and Swedish universities. Also the hospital sector, the pharmaceutical industry and other research institutions use PURE.

**Keywords:** meta-data models; long-term preservation; OAI-PMH; research portal; bibliometric analysis

## 1 Introduction

PURE is a commercial modular standard software platform for building institutional repositories at universities and other research institutes. The term CRIS [1] also applies. It supports the publication registration process in full - from ingestion and storage to publishing and reporting. PURE is a J2EE application running on most server operation systems and SQL environments and in connection with DSpace and FEDORA. This document offers a complete overview of PURE by addressing four basic repository issues: Data modeling, data ingestion, data storage and data exhibition.

## 2 Application Features and Architecture

### 2.1 Data Modeling

Apart from publications, the following content types can also be handled in PURE: Person, Organization, Project, Student Thesis, Activity, News Clip [2] and Clinical Trial. To some extent, these types are related to separate modules of the application, which are available as integral extensions to the basic module. The separate modules are: Reports, External Publications, Student Thesis, Bibliometry, News and Clinical Trials.

The terminology above is from the PURE meta-data model, which can be delivered with PURE - either as is or adapted to the individual research institution. But any meta-data model can be implemented in PURE; part of the application architecture is development frameworks that facilitate such implementations. During the last 5 years, 4 universities have specified their own meta-data model and have had it implemented.

Currently, the PURE meta-data model is the only default meta-data model available. However, a default implementation of the CERIF2006 [3] meta-data model is currently being planned. It is expected to be completed by the end of 2007. A key specification with both meta-data models mentioned is the use of relations objects and many-to-many relations between all primary content types.

### 2.2 Data Ingestion

PURE's user interface is browser-based. Both Firefox and Internet Explorer are supported on Windows and Mac, though only in relatively recent versions. Users are assigned roles, which define two things: First, what functionality a user have access to. The individual users interface is adapted accordingly. Next, what the user's rights are. About 10 pre-defined roles comes with PURE and custom roles can be defined and added.

Workflows in PURE allow several users to participate in different parts of the same work process - for example the process of registering a publication. Workflows allow different users to take part in a registration process by modifying, enriching, validating manually entered data.

Data from existing systems can be used in PURE by means of one or more dynamic integrations. Usually, data for Person-objects - e.g. a person's name, title, room number, employment number, direct telephone number, e-mail address, etc. - already exists in one or more systems, and using such systems as a data source for PURE will be desirable in many situations. Organizational data would be a typical example, too, originating from systems such as LDAPs or Active Directories. Integration between PURE and any number of local sources is possible. Also user authentication and Single Sign-On integration is possible.

Imports are different from dynamic integration in that they deal with historic data and are carried out only once per data set. Publication registrations from an old repository would be a good example. To facilitate import of such data, the PURE XML Archive format (PXA) was specified. PXA files are created outside of PURE and imported via an XML I/O unit. Relations between a) publications to be imported and b) organizations and persons already in PURE is possible, depending on available data.

## 2.3 Data Storage

Data access is encapsulated using an object/relational persistence and query service. This allows the use of most SQL-database environments with PURE. Microsoft SQL-Server, Oracle and PostgreSQL are the most used SQL-environments among current PURE-users. PURE also maintains an index to support searches.

A so-called file connector interfaces PURE with the OS file system. The connector stores full-text files to the OS file-system while still retaining relations to the relevant meta-data objects. Further, two additional connectors are available for storing to a DSpace environment and for storing to a FEDORA environment.

## 2.4 Data Exhibition

PURE's two web services are a Document/Literal web service, which will make XML easily available for re-styling, and an RPC/Encoded web service, which allows requisition of data from PURE as whole objects. In both cases rich libraries of methods are available. Parameters such as date-range or organization can be added to each methods. Further, PURE has a portal-framework called PUREportal, which is an internal framework for building customized websites for exhibiting data from PURE. The framework itself comes with each PURE license. Together, the two web services and the PUREportal framework are how PURE exhibits data to websites.

Two more services help exhibit data from PURE. One is an OAI-PMH data providing services, the other is a Z39.50 based service. Different formats ca be defined under OAI-PMH; Dublin Core and DDF-MXD is supplied by default. Z39.50 was added to PURE because many library systems interfaces nicely with it. SRU/SRW will be implemented as demand rises. In addition, Reference Manager exporting capabilities allows export of any data set from PURE in native RefMan format, saving double work in some cases.

Finally, the report generator in PURE is a reporting and statistics tool, that will respond to all data in the entire PURE repository. A number of standard reports are supplied and available in three categories: Lists, Analyses and Bibliometrics. To run these standard reports, only a time interval and an organization must be chosen. To that, each standard report can be customized. Finally, custom reports can be build from scratch.

# Notes and References

[1] See http://www.eurocris.org

[2] Clips from media where researchers are mentioned. Such clips are usually supplied as an XML feed from a 3rd party supplier. Upon import to PURE, clips can be related to the appropriate researchers.

[3] See http://www.eurocris.org:8080/lenya/euroCRIS/live/

# Access to Free e-journals via Library Portals: The Experience of the Shahid Chamran Ahwaz University in Iran as a Case Study

*Amir Reza Asnafi*

School of Education & Psychology, Department of Library & Information Science,
Shahid Chamran University
e-mail: aasnafi@gmail.com

## Abstract

Journals as one of the most important information carriers are useful resources for libraries and information centers. As publishers more fully actualize the e-journal, it soon will be as insufficient to offer only print journals as it is to provide only print abstracts and indexes. Using journals especially scholarly journals, scientists can contact together in scientific communities. Impact of information technologies on journals has changed the format of these resources into electronic and has facilitated information storage and retrieval. Free e-journals are alternatives for non free e-journals and are useful for libraries that can not afford expensive costs to provide subscription-based e-journals. If libraries really support Free e-journals, then one would assume that we would be pretty aggressive about making free content readily available to our users through library systems and access points (e.g. ILS, knowledge base, web site). Finding free and non free e-journals in the web environment is more difficult than finding print journals, since they rarely are found in bibliographic resources, so their locating and retrieval is difficult. Librarians are responsible for information organization and retrieval and they must corporate in designing search engines and web pages to offer electronic articles that are needed for users. Portals are one of the tools that can be used for accessibility of free and non free e-journals to users. Portals as website are windows to World Wide Web and often have a search engine, links to useful pages, news or other services. In this article various literature and experiences about access to electronic journals via web pages has been reviewed. We decided to create a special portal of free e-journals for postgraduate students, masters & researchers of Shahid Chamran Ahwaz University to use of these resources. So, we provided a list of Shahid Chamran Ahwaz University courses and on the basis of this list, we selected free e-journals of each course via Directory of Open Access Journals (DOAJ)[1]. DOAJ has become "the" Open Access journal site for libraries because it is of a manageable size, Many librarians may think it is comprehensive, It is well organized and easy to harvest. In this survey, Researcher could extract 198 journals from this website that all of them were peer reviewed. Since we wanted to design a portal for free e-journals, we added selected journals to webpage that was linked to Shahid Chamran Ahwaz University Central Library website. This webpage was called free e-journals portal. All students and Masters could access to their needed articles freely. In this article we will discuss about importance of open access or free e-journals and their role in scholarly communication. Finally we will offer free e-journals portal of Shahid Chamran Ahwaz University Central Library as a tool for access to open access articles.

**Keywords**: open access; library portals; DOAJ; central libraries

## 1 Introduction

Using journals especially scholarly journals, scientists can contact together in scientific communities. Impact of information technologies on journals has changed the format of these resources into electronic and has facilitated information storage and retrieval.Free electronic jounrals are one of the types of electronic journals. These journals are accessible freely via Internet for users. Now, free electronic journals are the main part of scientific resources.

## 2 Research Aim

The major aim of this research, is designing a special portal for free electronic journals for Shahid Chamran University Ahwaz University on the basis of the attitudes graduate students of this university about these journals.

[1] http://www.doaj.org

# 3 Methodology

Data collecting tools were literature review, Checklist, questionnaire and Yahoo search engine. For data statistical analysis, descriptive statistics, Chi-Square, Anova One Way and Scheffe test were used. Results of checklist analysis, that was distributed among Iranian experts in Library & Information Science field, indicated 36 criteria can be used for selecting free electronic journals via Internet. By these criteria 198 journals, were selected from Directory of Open Access Journals.

# 4 Findings

This research indicated that graduate students of Shahid Chamran university of Ahvaz have little familiarity with free electronic journals of their special course and their use of these journals is in low level. On the basis of the attitudes graduate students of Shahid Chamran university of Ahvaz, evaluating criteria of free electronic journals quality and needed features for designing a special portal for free electronic journals were gained. In secondary findings part of this research, Chi-Square Test cleared that there is no significant difference among using full time and part time graduate students of Shahid Chamran university of Ahvaz of free electronic journals. In this research, by Webometrics method, highly cited free electronic journals were assigned. By this method, 63 highly cited free electronic journals were determined. Finally, by Microsoft Frontpage, that is a special software for designing web pages, primary version of special portal of free electronic journals for Shahid Chamran university of Ahvaz was designed and created. Free electronic journals of each university course will be accessible from this portal. Figure 1 shows portal of free e-journals in Shahid Chamran Ahwaz University. Its address is: http://www.cua.ac.ir/lib/central-library3/fire-home.htm

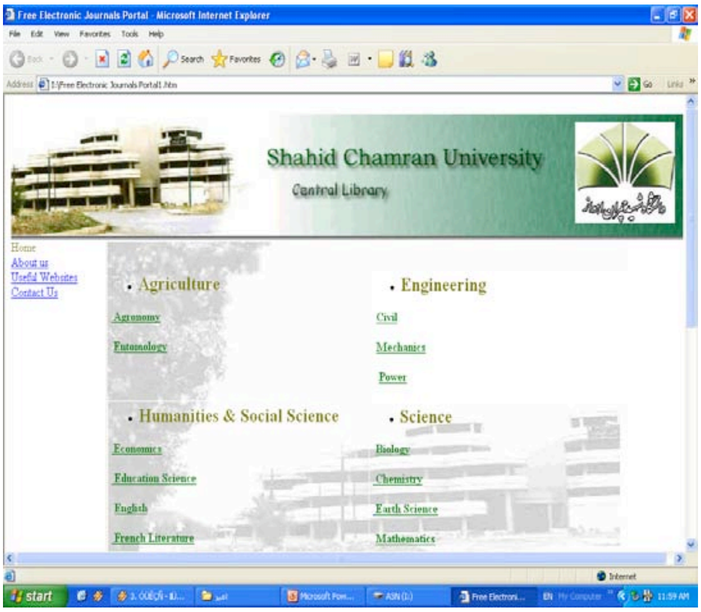


**Figure 1: Portal of Free e-journals**

## Notes and References

[1] ANDERSON, R. 2004. Open access in the real world: confronting economic and legal reality. *College and Research Library News* 64(4). Available at: http://dlist.sir.arizona.edu/archive/00000351/

[2] SADEH, T.; WALKER, J. 2003.”Library Portal : Toward the semantic Web”. *New Library World*. 104(1185), pp. 19-11

[3] RICH, L. A.; RABINE, J. L.1999.”How libraries are providing access to electronic serials: A survey of academic library websites”. *Serials Review*. Vol.25, No.2, pp. 35-46.

# Digital Archives at the University of Pisa

*Cinzia Bucchioni; Zanetta Pistelli; Barbara Pistoia*

Sistema Bibliotecario, Università di Pisa, Lungarno Pacinotti 44, Pisa, Italy
e-mail: bucchioni@angl.unipi.it; z.pistelli@bibant.unipi.it; b.pistoia@ing.unipi.it

## Abstract

At the end of the '90s, the Library System Centre of the University of Pisa began to create a system of digital archives in order to enhance and promote the Institution activities regarding teaching, research and administration: 1. ETD (Electronic Thesis and Dissertations); 2. UnipiEprints.

**Keywords:** digital archives; open archives

## 1 ETD (http://etd.adm.unipi.it/)

Due to the lack of a national system for thesis management and access in Italy, in the case of both paper and digital theses, the need arose for the University of Pisa to create a local system. This led to the choice of the NDLTD platform, an open source system developed at the Virginia Polytechnic Institute and State University, as it is a widespread international system specific to theses. The writing and discussion of a thesis is a significant moment in a student's career, and in the Italian academic context involves a series of administrative issues and many laborious bureaucratic processes. Therefore, the ideal system would need to implement a single work-flow, integrating the entire process of presenting, revising, discussing, cataloguing, giving access and preserving an academic and doctoral thesis: with this aim, a number of important software developments were necessary:

- The introduction of managing different deadlines: date for official deposit of the final version; date for public discussion; date for "last content revision" (the author can enter minor corrections within up to 48 hours before the discussion: immediate notification is automatically sent to the academic supervisors);

- In conformity with Italian copyright law and academic practices, the author can decide whether his/her thesis is open access (full text, only parts of the text), or only metadata are available: the system enables the author to change this option when required

- the most important development (in progress) regards interoperability between ETD and the other software systems used at the University of Pisa:

  - ESSE3, the administrative system for student records management
  - Aleph, the bibliographic system

Thus, when a student accesses ETD for the first time, the system retrieves all relative personal and academic data from ESSE3; moreover when the thesis is deposited, a UNIMARC record is created and sent to the Aleph catalogue.

The project entered the production stage in 2006. It has been recently registered in the Open Archives Initiative (OAI) register, and currently contains almost 3300 theses.

The digital deposit of theses is not compulsory: it is a shared decision of the academic supervisors and the student. The project group has been working on widespread promotion of the system in the academic community, with very different feedback from the various disciplinary communities: the STM community makes wide use of the system (to date around 60% of theses are native digital), while the humanities community appears to be more sceptical (around 10% of theses are native digital).

## 2 UnipiEprints (http://eprints.adm.unipi.it/)

The University of Pisa officially signed the Berlin Declaration concerning open access to knowledge in May 2005: this prompted the project to create a system of institutional archives devoted to scientific and educational documents produced at our University. Eprints, the open source software of the OA Initiative was chosen, in order to be included in the OAI services. No serious customization was necessary, with the exception of translating the interface pages and adding the discipline categories specific to the Italian academic context (degree courses, departments, research teams, etc.).

The top-down start-up has found the librarian community much more aware of access and publishing issues than the academic community; nevertheless on personal or Faculty web pages freely available and interesting documents can still be found, often used for teaching; naturally these are in no way systematically organised, with no possibility of easy retrieval or permanence.

Since our institutional open archives aim not only to meet the global demand for new scholarly communication models, but also to meet the more specific needs of our academic community, an architecture with double archives has been chosen:

- UnipiEPrints, institutional archive devoted to research works of teachers and researchers and to institutional documentation (in the production stage since December 2006);

- UnipiEPrints Didattica, a repository for the different typologies of educational documents (production phase scheduled for April 2007; access will be restricted to the University network - some software development is required).

We are now entering the promotion phase: a first official presentation of the system will take place at the University Senate; a series of meeting and seminars in the faculties and departments will follow, aimed at explaining how UniEprints works and making academic authors aware of the economic models of scholarly communication, the policies of publishers, and the issue of copyright. We imagine a promotion model strongly based on libraries and librarian support. We have already observed that the main questions posed by Professors concern academic evaluation of works, with a number of distinctions:

- The STM community appears to be more interested in the assessment of quality, impact factor and referee etc.;

- The Humanities community also publish in monographs or in journals which are not included in bibliomethrics databases and sometimes have distribution problems; their main concern regards certainty of the publication status, above all in a situation where recent changes in legal deposit law are still awaiting case law and practices;

- We have found allies in a research group of the Political Science Faculty: their research field regards the social and philosophical aspects of knowledge production and communication, and the group has developed their own OAI disciplinary repository, also based on eprints software. This group is willing to interoperate and to support promotion.

The project group proposal to the academic government, which has shown interest, is to connect the EPrint system with "Anagrafe della ricerca", i.e. the official data which Italian Universities have to record as a base for fund distribution: this combination would significantly enhance use of the system.

# A Survey on magiran.com: A Database for the Magazines of Iran

*Mortaza Kokabi*

Department of Library & Information Science, Shaheed Chamran University, Ahwaz, Iran
e-mail: Kokabi80@yahoo.com

## Abstract

This paper present the design and function of magiran.com, a databse of periodicals published in Iran. It also attempts to answer the following questions: How many of the total periodicals published in Iran are covered by magiran? What is the subject coverage of the periodicals covered? Which subjects seem to have been given importance among the periodicals covered? How many of the periodicals are available full text? What is the subject coverage of the periodicals available full text? What are the languages of the periodicals covered? How many of the periodicals accredited by MSRT are found in magiran? What is the subject coverage of the accredited periodicals? Which subjects seem to have been given importance among the accredited periodicals covered? How is the general structure of the site in terms of colors, icons, pull-down windows, and so on?

**Keywords**: Iranian periodicals; subject coverage; user feedback

Although there have been some sporadic activities to index and abstract Iranian periodicals by some organizations responsible for the press in Iran, there has not been any complete source, employing the Internet and indicating the outcomes of the activities of people involved in country's press. The *ftāb Software Company* (ftābsoft.com) sponsored the design and development of magiran website [1], a database of periodicals of the country, simultaneous with the Press Festival held in May 2001 in Iran. The purpose of designing the site is producing an effective source for Iran's periodicals in the Internet.

The site has been able to cover and present services related to more than 1300 periodicals in publication, authorized by the Ministry of Culture and Islamic Guidance (MCIG). The site claims that it is used by more than 15000 users inside and outside of Iran. The free services offered by magiran are of two kinds: General and special. The general ones, besides the ones mentioned above include: the allocation of special address such as http://www.magiran.com/YOURMAGAZINE to provide quick access to periodicals information, the allocation of an email address by POP3 service with at least 10 Mb capacity such as: *YOURMAGAZINENAME@magiran.com*, the inclusion of subscription rates and forms for each periodical to subscribe from inside or outside of Iran, informing the users of the publication of new issue of each periodical, and the inclusion of the full text of new issues and the last ten issues of some periodicals. Free special services include: providing special pages for each periodical. There are some Extra services that are based on request and payment. Another useful service of the site is its acting as a dealer.

The subject directory of periodicals is a list that is continued via a link in another page. On the same page, the periodicals are presented according to a subject directory, the Ministry of Sciences, Researches, and Technology (MSRT)-accredited periodicals given importance by being at the top. On the same page, the periodicals can be searched through a search box, the search can be limited by some options, and the newspapers covered are also shown. There's a "sending message, viewpoint, and suggestions" page, on the same page, as "Introduce the site to your friends" and "Report the problems with the information" options. The site is totally independent and private and has no connection to any governmental or non-governmental institution or organization.

This paper tries to find answers for the following questions: How many of the total periodicals published in Iran are covered by magiran? According to the latest statistics belonging to 2005 [2], the total number of periodicals published daily, monthly, bi-monthly and quarterly in Iran is 1832. Thus the site covers approximately 70% of the periodicals authorized by MCIG; 2. What is the subject coverage of the periodicals covered? Literature, Art, Technical and Engineering, Society and culture, Areas and ethnicities, Industries, Agriculture, Information, computer and internet, General, Basic sciences, Islamic sciences, Humanities, Groups, Ecology, Commerce and economics, Health and treatment, Sport and entertainment, Education and research, Associations and NGOs; 3. Which subjects seem to have been given importance among the periodicals covered? The periodicals covered by Magiran are categorized in 20 subject groups, each subdivided in turn into some sub-categories with each of which the total number of periodicals in that sub-category is given. Of these subcategories, "Industries" with 31 sub-categories has the highest rank but "Health and treatment" with 192 titles has the highest rank in respect to the total number of periodicals in sub-categories; 4. How many of the periodicals are available full text? 92 titles; 5. What is the subject coverage of the periodicals available full text? No information on the subject coverage of these full-text periodicals is given; 6. What are the languages of the periodicals covered? No specific information on language coverage is given in the site, but periodicals are mostly in Farsi, the official language of Iran, and some are in English. For some periodicals, only the abstracts are given in English. Some periodicals are also bilingual, or in fact bi-dialectal, such Farsi-Kurdish; 7. How many of the periodicals accredited by MSRT are found in magiran? 130 titles (10%); 8. What is the subject coverage of the accredited periodicals? Agriculture, Medicine, Basic sciences, Art and Architecture, Humanities, Technical and Engineering; 9. Which subjects seem to have been given importance among the accredited periodicals covered? "Medicine" with 65 titles seems to be the most important subject; the "Agriculture" and "Humanities" both with 20 titles is the second; 10. How is the general structure of the site in terms of colors, icons, pull-down windows, and so on? The site was matched against some criteria [3], and the results are as follows: The site address and domain and the keywords in the site name are visible; the name of the site is short and informative; the text is readable but the fonts could be better; the illustrations, though not very frequent, are attractive; no site map; writing and grammar are acceptable; there's the possibility of navigation through the site; the name of the designer company is seen, but not the site administrator; the last date of updating is not seen; no FAQ provision; the introduction of new periodicals exists; no "help" provision but "about us". The site takes between one and two minutes to load, much longer than the standard 8 seconds.

Suggestions to improve the site are as follows: more beautiful Farsi fonts could be applied; and the inclusion of more illustrations; the site map; the total periodicals authorized by MCIG; more full-text periodicals; the last date of updating; FAQ provision; and "help" provision seem necessary.

## Notes and References

[1] http://www.magiran.com

[2] http://www.sci.org.ir/portal/faces/public/sci/sci.gozide

[3] SABERI, M. *A comparison between the content and construct features of the central libraries' homepages of US, Canada, and Australia with those of Iran along with a survey on the viewpoints of users and experts to present an optimal model*. MLib. Dissertation, School of Education and Psychology, Shaheed Chamran University (in Farsi)

# Developing National Open Access Policies: An Ukrainian Case Study

*Iryna Kuchma*

Social Capital and Academic Publications Program, International Renaissance Foundation
46 Artema str., Kyiv, 04053, Ukraine
e-mail: kuchma@irf.kiev.ua

## Abstract

Since January 2007 Ukraine has a law mandating open access to publicly funded researches. It was widely supported by most of the Parliament members. And it is already the second parliamentary inquiry mandating the Cabinet of Ministers to take actions on creating favourable conditions for developing open access repositories in archives, libraries, museums, scientific and research institutions with open access condition to state funded researches. And for the second time the implementation of this law was interrupted by the political crises. Grass root initiatives of Ukrainian Universities and libraries as well as the political support from the principle legislative body in the country have still not resulted into a single well-functioning institutional/national repository. The poster highlights the developments that have taken place, actions for the years to come and recommendations for the countries that are in circumstances that can be compared to Ukraine.

**Keywords:** open access; mandating policy; publicly funded researches; institutional repository

## 1 Introduction and Developments

Mandating open access to publicly funded research in Ukraine was a movement launched by the scholars publishing their articles in open access journals, innovative librarians and University administrations. This movement was co-ordinated by International Renaissance Foundation (IRF, Soros Foundation in Ukraine), which since 2004 organised a number of awareness raising campaigns in mass media and regional seminars for the academic community. National Academy of Sciences (NAS) and International Researches and Exchanges Board (IREX) supported open access ideas and joined the movement.

The first public statement on open access policies in Ukraine was drafted during the international Open Access Scholarly Communication Workshop hosted by the National University Kyiv-Mohyla Academy (NAUKMA) and organised by IRF, Open Society Institute, NAS and International Association of Academies of Sciences on February 17-19, 2005. 140 researchers, administrators, librarians, information managers from higher educational institutions and scientific research laboratories involved in e-journal publishing and institutional repository development from 17 countries signed the Recommendations for Ukrainian authorities to ensure: the right of individuals and the public to access information and knowledge and to guarantee that intellectual property regimes are not the obstacles to the public access to knowledge, to encourage research and higher educational institutions to practice open access and to put an open access condition to state funded researches (except reasonable exceptions) and to provide state financing and technical assistance to research and higher educational institutions to set up and maintain open access repositories.

These Recommendations were endorsed by Ukrainian Vice Prime Minister. And on September 21, 2005, the Recommendations were presented at the first Parliamentary hearings on Developing information society in Ukraine. In December 2005 these hearings resulted into the Parliamentary Inquiry on Harmonisation of Governmental Educational Policies re open access movement [1]. Open access was one of the priorities in developing information society in Ukraine. The Cabinet of Ministers was responsible for creating favourable conditions for developing open access repositories in archives, libraries, museums and other cultural institutions and the Ministry of Education and Science of Ukraine – for encouraging development of open access resources in science, technology and education with open access condition to state funded research. Beginning of 2006 was also the time of parliamentary elections campaign, when the "old" Cabinet of Ministers didn’t feel any responsibility to start new activities like open access projects. And later on two "new" Cabinet of Ministers were busy trying to cope with political crises in summer and autumn 2006.

In September 2006 representatives of Parliamentary Committee on Science and Education, State Fund for Fundamental Researches, Scientific and Publishing Council of NAS, Ministry of Science and Education of

Ukraine, National Library of Ukraine after V.Vernadsky, State Department of Intellectual Property, Kyiv public administration, Association "Informatio-Consortium", Institute of Social Development and IRF created a working group on developing open access policies in Ukraine and pushing the Cabinet of Ministers to implement the resolution of Ukrainian Parliament on Open Access.

In November 2006 State Fund for Fundamental Researches commissioned IRF to develop an Open Access Policy for their grantees reporting publicly funded research. The goal was to require electronic copies of any research papers supported in whole or in part by Government funding to be deposited into an institutional digital repository immediately upon acceptance for publication.

Both initiatives turned previous parliamentary resolution into the law mandating open access to publicly funded research [2]. According to the law there should be six months of transition period (completed by July 2007). But the following political crises withdrew the attention of the Cabinet of Ministers from immediate implementation of this law.

Since October 2005 a grassroots initiative of the academic community undertook a project to create a network of open access repositories in Ukraine. Nine Ukrainian Universities reported this decision at the national conference for university and regional universal scientific libraries INFORMATIO 2005. The project has been implemented by Association "Informatio-Consortium", Scientific Library of National University Kyiv Mohyla Academy, Lviv Catholic University and Centre for the Humanities of Lviv National University after I.Franko. All these projects still lack financing and skilled staff. So far only pilot institutional repositories have been created.

Governmental institutions are still the unique donors of research and development in Ukraine. This is why a law mandating open access to publicly funded research plays a crucial role in open access initiatives. Delays with implementation of this law cause delays in the development of open access institutional repositories.

Nevertheless we will continue financial and expert support to Ukrainian network of open access institutional repositories encouraging Universities and research institutions to sign the Berlin Declaration and introduce self-archingl policies, develop model open access institutional repositories and providing training for the interested organisations. At the policy level we will keep pushing the implementing of the law of Ukraine mandating open access to publicly funded research. IRF implements open access projects in cooperation with the Information Program of the Open Society Institute and the Electronic Information for Libraries Consortia (eIFL).

## 2 Recommendations

Recommendations for countries that are in circumstances that can be compared to Ukraine: 1) alliances are crucial and local partners needed; 2) targeted web-sites and workshops proved to be useful tools for awareness raising and lobbying; 3) support from mass media is important to create public awareness.

## Notes and References

[1] Decree of the Parliament of Ukraine "On Recommendations of parliamentary hearings on developing information society in Ukraine: http://zakon.rada.gov.ua/cgi-bin/laws/main.cgi?nreg=3175%2D15

[2] The Law of Ukraine on the principles of developing information society in Ukraine for 2007-20015 (Закон України "Про Основні засади розвитку інформаційного суспільства в Україні на 2007-2015 роки") at www.rada.gov.ua

# Digitization of Scientific Journals in Serbia

*Žarko Mijajlović*[1], *Zoran Ognjanović*[2], *Aleksandar Pejović*[3]

[1] Faculty of Mathematics, University of Belgrade, Belgrade, Serbia
e-mail: zarkom@matf.bg.ac.yu
[2] Mathematical Institute SANU, Belgrade, Serbia
e-mail: zorano@mi.sanu.ac.yu
[3] School of Electrical Engineering, University of Belgrade, Belgrade, Serbia
e-mail: pejovica2@gmail.com

## Abstract

A digitization project in progress carried out by the Mathematical Institute of the Serbian Academy of Sciences, Belgrade (http://www.mi.sanu.ac.yu) and the Faculty of Mathematics, Belgrade (http://www.matf.bg.ac.yu) is described. The projects aim is to build a database and an electronic presentation of digitized scientific books and journals printed in Serbia, particularly in mathematical sciences (mathematics, mechanics, astronomy, computer science and physics) and to make them searchable and available in the full-text mode on the Internet.

**Keywords:** scientific journals; retro digitization; digital archives

# 1 Introduction

There is a number of mathematical and mathematically-related books and journals printed in Serbia. Some of the published works have been digitally created (usually in TeX and its versions) and more-or-less they are accessible using ordinary web browsers, while the others, mostly older papers, have been born in print and are harder to obtain. Thus, we have also started the process of retro digitization of these printed works to produce their digital images that can be read or printed. The main achievement of the project in last two years is complete retro digitization of two leading Serbian mathematical journals: *Publications de l'Institut Mathematique* and *Publications of the Faculty of Electrical Engineering – Series Mathematics and Physics*. Besides these two completely digitized journals, five Serbian journals in mathematics, teaching and computer science are stored in the database. Some of them, for example *NCD Review*, a journal on digitization technologies, founded in 2002, are presented completely, while the others are partially digitized, mainly from the beginning of nineties of the last century. All together, there are more than 4000 digitized articles having about 30000 pages. Digitized items are displayed at the virtual libraries: http://alas.matf.bg.ac.yu/biblioteka/home.jsp, http://publication.mi.sanu.ac.yu, http://pefmath2.etf.bg.ac.yu/ and http://elib.mi.sanu.ac.yu/pages/browse_journals.php.

# 2 Serbian Mathematical Journals in Digital Archives

The journal *Publications Publications de l'Institut Mathematique* is the oldest Serbian scientific journal in the field of mathematics established in the year 1932 under the name *Publications Mathmatiques de l'Université de Belgrade*. It was founded with the help of two foundations of the Belgrade University foundations. Seven tomes were published until the World War II, the eighth tome was lost in the German bombing of Belgrade in April 6. 1941. Immediately after the founding of the Mathematical Institute in 1946, the publication of the journal was restarted in 1947 under the new name *Publications de l'Institut Mathematique*. More then 2000 articles were published in 102 volumes until these days. The scope of the journal in the beginning was broader, not only in mathematics, but articles referring to mechanics and astronomy were published in it as well. Most prominent Serbian and Yugoslav scientist in these fields published in the journal, including Đ. Kurepa, J. Karamata, M. Petrović, M. Milanković, A. Bilimović, J. Plemelj, S. Mardešić, and others. Some of the leading world mathematicians published in Publications as well: H. Lebesgues, P. Montel, P. Erdös, W. Sierpinski, S. Shelah, and others. Most papers in the journal are in English, but there are papers written in Russian, French, and German as well. The second journal, *Publications of the Faculty of Electrical Engineering – Series Mathematics and Physics*, was founded in the year 1956. In the beginning, each contribution appeared separately bound and numbered consecutively, several times a year. Since 1959, the issues have been appearing collected in one or more volumes per year. In the first years, the journal had contributions from different fields apart from Mathematics: Physics, Mechanics, and Electrical Engineering. Papers were written in the Serbian, French, Russian, German and English. In the course of time, the journal focused almost exclusively on Mathematics, especially convexity, functional equations and differential equations, and English language became dominant. The digitized version of the journal contains about 1000 papers. Both journals are reviewed and indexed in: *Mathematical Reviews* (MR), *Zentralblatt Math* (ZBL) and Russian *Mathematical Surveys*.

## 3 Digital Objects and Metadata

The digital object in the virtual library usually consists of several components: digitized image of the manuscript, some graphic components, and metadata. We developed a particular data base and Internet oriented software for handling digitized journals. It relies on three types of metadata: descriptive, structural and administrative. Special data and services important for papers published in scientific journals were also included: keywords, scientific classification of AMS (American Mathematical Society classification), numbers of reviewer reports in MR and ZBL, DOI numbers, and statistics of accessing and downloading papers. Descriptive metadata follow data contained in librarian printed catalogs, i.e., they obey librarian standards. One problem was that old issues do not have standard descriptive tags such as ISSN numbers so to classify them we needed particular solutions. Structural data explain how the components of the digitized object are interconnected. Administrative data describe exactly how an item is preserved: resolution, rate of compression, file type containing the digitized image, etc. The success of digital preservation efforts will rest to a significant degree on the scope and reliability of the metadata records. For example, metadata made possible the asset-management systems that back up and periodically duplicate digital records. Cataloging information enable one to locate what they are looking for in the library. Metadata help to make various internet presentations. Therefore full repository system required tens of metadata elements for each digitized item. Building such database systems and populating them is very labor-intensive and expensive. Creating the table of keywords and assigning them to articles was particularly complicated and time-consuming job since it could be done only by scientists. Some trade off needed to be found. For resolving these issues, cooperation between institutions working in the field of digitization was very important, in particular exchange and agreement of metadata formats. Particular attention was given to standards. Scanning was performed in 300dpi, in tiff format. Papers were converted into pdf format and in this form they are accessible on Internet. As curiosity, let us mention that all issues of *Publications Publications de l'Institut Mathematique* between 1980 and 1990 are retyped in TeX, and all issues of this journal since 1980 are accessible in dvi format, as well.

## 4 Implementation

We decided to develop our own software instead of using commercial, or open source software. We decided so, since we wanted to lower the development and maintain cost, then because of future upgrading and integration it in larger information systems, such as virtual libraries of wider scope. The software supports all usual functions, browsing, searching under various criteria, examining and downloading papers. Since papers were written in several languages, we decided to keep the multilingual feature. Therefore, we have chosen MySQL server for a database as it supports UTF-8 encoding. The multilingual support is embedded into the model of data, so information related to the corresponding languages are saved. JAVA programming language is used in developing a web application for administering and searching the database, especially advanced features like JAVA beans and strucs, which enable a high performance web application. Other technologies include PHP and Apache as a web server.

## 5 Conclusion

Digital archive of Serbian scientific journals will contribute significantly to the widespread accessibility of articles printed in these journals, particularly since they are obtainable on the web free of charge. One of the consequences will be the rise of scientific impacts of these journals and articles printed in them. A further plan assumes that our Virtual library will include once editions of all important Serbian scientific journals. It is difficult to estimate when this task will be finished, but a decade, we believe, is a good guess.

## References

[1] MIJAJLOVIĆ, Ž., *On some undertakings in the field of digitization in the last decade*, NCD Review, 2002, http://elib.mi.sanu.ac.yu/pages/browse_article.php?cs=000001&rd=0000003.

[2] ARMS, W Y., *Digital Libraries*, MIT Press, 2001.

# DRIVER - Digital Repository Infrastructure Vision for European Research

*Mary L. Robinson*

SHERPA, Information Services, University of Nottingham
Greenfield Medical Centre, Medial School, QMC, Nottingham, UK.
e-mail: mary.robinson@nottingham.ac.uk

## Abstract

The current system of academic publication developed as a means to disseminate the findings of research. However, this system can impede the very process it was set up to serve, with access to articles being limited by publishers to only those who can afford to subscribe. This poster will explain the vision behind DRIVER and will describe how the various aspects of the project tie in together to form the knowledge infrastructure of the European Research Area. The poster will focus on the key aspects of the DRIVER project and the questions and needs that each addresses. The key aspects which will be addressed include: DRIVER technical developments and advice, the DRIVER Support website [1], community development, up-to-date news, and the benefits for various stakeholders. DRIVER is an ambitious and important project that will yield valuable results for individual researchers, the publishing community, funding agencies and the European Research Community as a whole.

**Keywords:** open access; repositories; European Research

## 1 Introduction

Open access digital repositories provide a means whereby the traditional publishing model can co-exist with the needs of authors and their readers, as well as with the demands of research funders for research impact and hence, value for money. Subject to copyright, authors can deposit copies of their finished articles in open access repositories, in addition to publishing them in research journals.

The recent study of scientific publication markets in Europe funded by the European Commission [2] strongly recommends the development of a European policy mandating open access to EC-funded research. In addition, it recommends an exploration of interoperability issues and how open access repositories can be implemented Europe-wide.

DRIVER- Digital Repository Infrastructure Vision for European Research- is an EU-funded project with 10 international partners and reflects the growing awareness in Europe surrounding Open Access. DRIVER sets out to build a testbed for a future knowledge infrastructure of the European Research Area. It aims to deliver any form of scientific output, including scientific/technical reports, working papers, pre-prints, articles and original research data to the various user groups. The testbed is based on existing nationally organized digital repository infrastructures. Other work includes the support of new European repositories and an active advocacy and community building programme to address and support key stakeholder groups in Europe.

## 2 Objectives

The five objectives of DRIVER are:

1. To organise and build a virtual, European scale network of existing institutional repositories;
2. To assess and implement state-of-the-art technology, which manages the physically distributed repositories as one large scale virtual content resource;
3. To assess and implement a number of fundamental user services;
4. To identify, implement and promote a relevant set of standards;

5. To prepare the future expansion and upgrade of the DR infrastructure across Europe and to ensure widest possible involvement and exploitation by users.

## 3 Discussion

This poster will provide key information on the DRIVER project including the ten project partners and the DRIVER Support website and logo. The poster will identify the various aspects of the DRIVER project and the questions and needs that each addresses. The poster will address the following: DRIVER technical developments and advice, the DRIVER Support website, community development, Up-to-date news, and the benefits for various stakeholders.

DRIVER has conducted focused research studies including an inventory of the type and level of OAI compliant digital repository activities in the EU [3], to facilitate the iterative development of DRIVER and is developing the necessary infrastructure middleware and user guidelines to meet the DRIVER objectives. The project is now actively advocating repository development - creating an informed and active environment for repository infrastructure development in EU countries with focused activities, information and contextualized support.

## References

[1] DRIVER (Digital Repository Infrastructure Vision for European Research), http://www.driver-support.eu

[2] European Commission. Study on the economic and technical evolution of the scientific publication markets in Europe, Jan. 2006, http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf

[3] DRIVER. Inventory study into the present type and level of OAI compliant Digital Repository activities in the EU, Apr. 2007.

# The Inclusion of Open Access Journals in Academic Libraries: A Case Study of Bioline International

*Jen Sweezie*[1]*; Nadia Caidi*[2]*; Leslie Chan*[1]

[1] Bioline International, University of Toronto at Scarborough, 1265 Military Trail, Scarborough, ON M1C 1A4
e-mail: {sweezie; chan}@utsc.utoronto.ca
[2] Faculty of Information Studies, University of Toronto, Toronto, ON, M5S 3G6
e-mail: caidi@fis.utoronto.ca

## Abstract

Specialized open access digital collections contain a wealth of valuable resources. However, major academic and research libraries do not always provide access to them, and thus do not benefit from these unique resources. This case study of one such digital collection, Bioline International, surveys 76 academic libraries in Canada and the United States to determine how often libraries are linking to the collection. A follow-up questionnaire was sent to librarians at the surveyed institutions to determine their opinions about the use of open access journals. The findings suggest issues of poor adoption rates of open access journals, as well as some reasons why such journals may not be actively adopted.

**Keywords:** open access; journal collection; developing countries; bioline international; digital library

## 1 Introduction

Librarians and information professionals continue to struggle with a growing number of available publications and limited budgets, making the selection of library resources difficult. Open access (OA), and in particular, specialized OA collections provide libraries with access to a broad range of high quality academic research and offer the possibility to help alleviate what is referred to as the serials crisis [1]. This study examines why such valuable resources are often not incorporated into library collections. A case study of Bioline International (BI) is used to illustrate this problem.

*Bioline International* (http://www.bioline.org.br/) is a specialized OA collection that offers open access to over 50 bioscience journals published in developing countries. The BI website provides free access to regional journals in environmental and agricultural sciences, heath and medicine, that may be difficult to obtain elsewhere [2].

## 2 Methodology

Between October 2005 and April 2006, an exploratory study of Canadian and American academic libraries was carried out to determine how many BI journals were included in the libraries' collections. E-journal or e-resource sections of each library website were searched and a checklist was completed, indicating the presence of BI journal titles. 76 (46 Canadian and 30 American) libraries were surveyed for BI's journals. A list of all academic libraries in Canada was generated using Yahoo categories, resulting in the survey of 46 libraries. Due to the large number of academic universities in the United States, a sample of libraries was necessary. A ranking of health science libraries by number of total electronic materials from the Association of Research Libraries (ARL)[3] was obtained. 30 American libraries were selected from the table. Effort was made to select libraries from the top, middle, and bottom of the ranking list, in an effort to ensure that the libraries selected represent a variety of different institutions (both private and public) across the United States.

After the e-journal lists were surveyed, a questionnaire was e-mailed to librarians identified during the library investigation process. Efforts were made to send the questionnaire to librarians working in collection development or on electronic resources lists. Where such a contact was not available, the questionnaire was sent to the head librarian. 76 questionnaires were emailed and 17 responses were received. Librarians were asked about their opinions surrounding the use of open access journals in library collections, as well as about institutional policies and decision making surrounding open access journals.

## 3 Results

Preliminary findings indicate that in Canada only 33% of research libraries (15 /46 libraries examined) were linking to 50% or more of the titles available from BI (more than 29 out of the 59 available titles). In the United States, 56.7 % (17/30 libraries studied) of the research libraries surveyed were linking to 50% or more of BI's 59 journal titles. 89% (41/46 libraries studied) of Canadian libraries offered at least one BI journal title through their library collection. In the United States, 96.7% (29/30 of libraries studied) of libraries offered at least one BI journal title of the 59 available through their library collection.

Libraries were considered to be aware of the BI project if they linked to one or more Bioline journals, even if they linked to the publishers (or other website) directly, and not to BI. Journals participating in the BI project actively promote their inclusion in Bioline through their websites and on the covers of their printed journals, thus knowledge of the BI project is assumed. In both countries, most libraries were aware of BI titles, either through the BI website, through another project or the publisher's website. However, inclusion of all 59 BI journal titles was still relatively low (under 11%) for both countries.

The results of the questionnaire sent to collections development and electronic resource librarians indicated that a number of reasons may contribute to the low level of appearance of BI journals in the library collection. Lack of inclusion in major commercial or open access databases and indexes, the length of time a journal has been publishing, lack of librarian time to seek out and catalogue (or even keep up with) new titles, institutional policy and reputation or perceived credibility of OA journals were all cited as factors in why such OA journals are or are not included in library lists.

## 4 Conclusion

The results suggest that despite the open access nature of BI and the range of its offerings, librarians are not making effective use of the BI collection. Projects such as BI must address the concerns and needs of librarians in order to improve the rate at which journals are added into library collections. Though librarians are often aware of open access journals, time constraints may be considered one of the largest barriers to adopting these journals. The unique situation of open access journals is that they do not often have the large scale budgets of large, mainstream publications, making it difficult to develop compliance with a number of web standards being used by libraries today. Standards such as OpenURL and metadata linking protocols can greatly increase the ease of adoption of these journals in libraries, and improve the chances of having such journals eventually indexed in major databases. These protocols allow librarians to easily link new journal material into their catalogues, and in some cases, some commerical applications are already adding some OA journal titles as options in their databases – librarians merely have to toggle them on or off. Issues of sustainability and journal quality will likely improve gradually as more and more libraries link to and promote OA journals, and journals are included in more databases and indexes. Projects such as BI can provide a sustainable platform by working independently from their individual journals – ensuring web access that is reliable – while also working for the collection as a whole in developing protocols and linking systems that the journals may not be able to generate on their own.

Ongoing study into the trends of inclusion of OA journals in library collections lends itself to a number of conclusions about the barriers to open access publications, how they could be better promoted, and how librarians can be encouraged to make use of these valuable resources.

## References

[1] OJALA, M. (2005) Open access: open sesame or opening Pandora's Box. EContent 28, 6, 31-32, 34-35.

[2] BIOLINE INTERNATIONAL (BI). "About Bioline" Retrieved Jan 24, 2006 from http://www.bioline.org.br/info?id=bioline&doc=about

[3] YOUNG, M.; KYRILLIDOU, M. (2005) ARL Academic Health Sciences Library Statistics 2003-2004. 29. Retrieved Jan 20, 2006 from http://www.arl.org/stats/pubpdf/med04.pdf

# Developments in Publishing: The Potential of Digital Publishing

*Xuemei Tian*

School of Business Information Technology, RMIT University
GPO Box 2476V, Melbourne 3000, Victoria, Australia
e-mail: Xuemei.Tian@rmit.edu.au

## Abstract

This research aims to identify issues associated with the impact of digital technology on the publishing industry with a specific focus on aspects of the sustainability of existing business models in Australia. Based on the case studies, interviews and Australian-wide online surveys, the research presents a review of the traditional business models in book publishing for investigating their effectiveness in a digital environment. It speculates on how and what should be considered for constructing new business models in digital publishing.

**Keywords:** digital publishing; business models; print-on-demand; publishers; Australia

## 1 Introduction

This poster session emerges from an Australian Government funded digital publishing research project. The research looks at the impact of technology and market change on traditional publishing practices, and the increasing imperative to digital publishing. For our purposes digital publishing is defined as publishing dependent upon the World Wide Web as its communication channel, producing digital content based on either domestic or global platforms, published and distributed online, with provision for the establishment of digital database facilities for future re-use. The process allows for links to e-commerce, for example, facilitating online payment, with all procedures in the process digitised. Based on customer requirements, the product (information) can be produced and provided in various formats, such as online, web, TV, CD Rom and if necessary, paper (Liu and Rao, 2005). Additionally, Print-on-Demand (PoD) and Video-on-Demand (VoD) are elements of digital publishing. There is a general consensus that the digital publishing production and supply chain incorporates authors, publishers, technology providers, databases, web distributors and end-users.

The Australian publishing industry has maintained a significant presence within the manufacturing and distribution sector over many years (IBISWorld, 2006). Over the past decade the publishing industry has undergone tremendous changes, including publishing markets, an increase in digital content formats, changes to distribution channels and supply chains. At the same time, revenues from online business have grown dramatically in recent years. Statistics reveal that major players in the Australian publishing industry have increased their online business revenues by about 30% from 2001 to 2005 (IBISWorld, 2006). In the process, publishers have been forced to re-evaluate their resources and capabilities, design new business strategies and re-engineer their business processes to take advantage of the potential of rapidly developing technology. This has led to the development and emergence of new supply chains in the publishing industry, and a need for new business models (IBISWorld, 2006)

This research investigates the practical implications of digitization for book publishing in Australia, focusing on aspects of re-engineering existing business models to maximize benefits to both the company and their customer base. Although not for profit publishers are already utilizing the Open Access system, this does not feature as a major element in this research.

## 2 Project Progress and Initial Findings

At the time of writing, the researchers have completed a national online survey of Australian publishers and are progressing through a series of eight case studies. The indications so far are that the Australian publishing industry has adopted a somewhat conservative approach to the challenges and opportunities presented by digital publishing. The majority of publishers believe however, that in the foreseeable future, there could be major changes in the industry. As publishing businesses have varying objectives, the pace of change from traditional to digital publishing will also vary, depending on their market and client base, as well as on take-up of technology. It seems clear however, that any reluctance on the part of publishers to embrace technological and other changes could be detrimental to their future.

## 3 Major Research Focus: Business Models

The major thrust of the research has ultimately been towards identifying the implications of digitization and related organizational and market changes for business models in the publishing industry. Our survey results indicate that subscription-based and content creation models maintain popularity, frequently in the context of niche markets. Faced with different markets and strategies challenging, businesses need models that fit with their particular circumstances. Rather than have recourse to, in this project we adhered to Weill and Vitale's (2001) approach to business models. They define business models as a description of the roles and relationships among a firm's customers, allies and suppliers that identifies the major flows of product, information and money and the major benefits to participants (Weill and Vitale 2001). Our case studies provide validity to this definition. To surface our perceptions of business models, we will display several business models at this poster session seeking feedback from conference participants. Figure 1 is a constructive example of a potential future business model for a book publisher, drawn directly from our research.

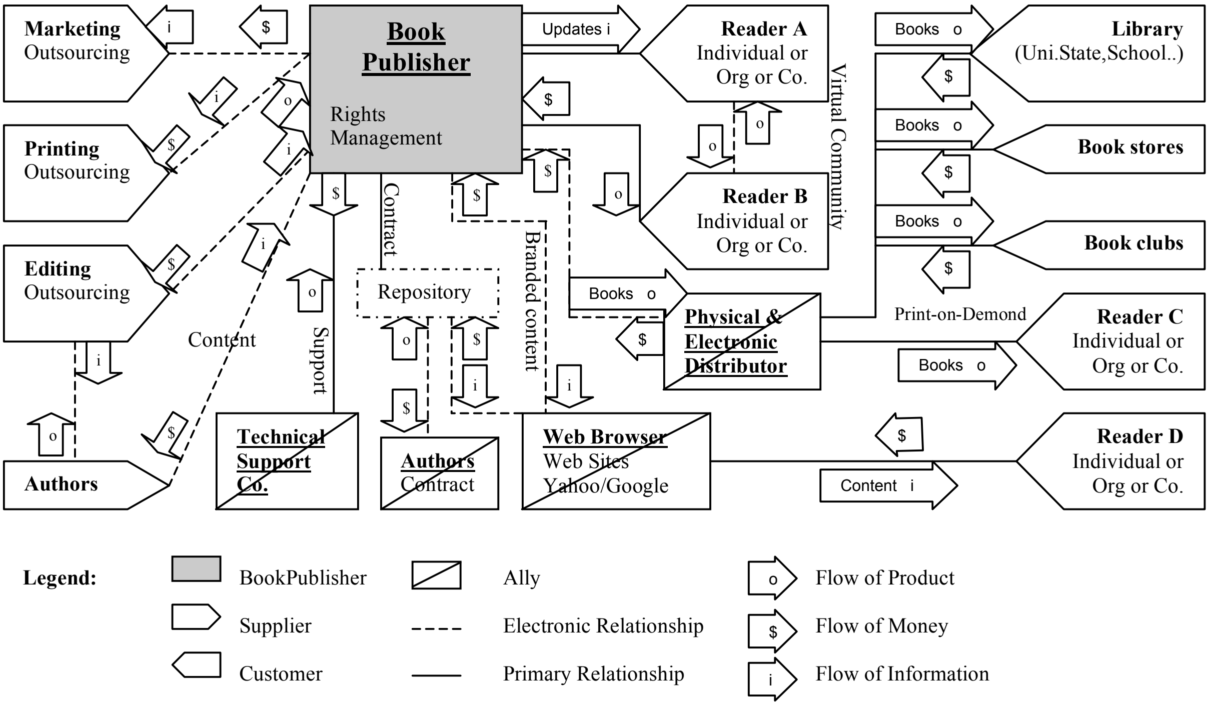


**Figure 1: A constructive business model for a book publisher**

## 4 Preliminary Conclusion

Although the research is being conducted in Australia, it has drawn widely in methodological terms from an international context. As such, it is likely that the findings, including those relating to alternative forms of business models, will be of wider relevance. Based on the models presented at this poster session we hope to add to our understanding and produce more refined models. At this stage preliminary conclusions are as follows:

- Currently book publishing models appear to be very familiar (i.e. largely traditional) with adding on digital elements.
- In 5 to 7 years time book publishing models could look vary different.

## References

[1] Ibis World Industry Report. *Book and other publishing in Australia*. Sydney, Ibis World Pty, 2006.

[2] LIU, M. L.; RAO, B. H. *Operational concepts and changes of academic journals in digital age*, ChongQing University Publisher, 2005.

[3] WEILL, P.; VITALE, M. *Place to space*: *Migrating to E-Business Models*, Boston, Harvard Business School Press, 2001.

# Index of Authors

# Index of Keywords