Multilingual Educational Repositories: Breaking Down the Language Barrier

Anastasios Koutoumanos
Agro-Know Technologies
15235, Athens, Greece
Email: tkout@agroknow.gr

Vassilis Protonotarios
Agro-Know Technologies
15235, Athens, Greece
Email: vprot@agroknow.gr

Abstract—This paper presents a framework for introducing multilingualism into web portals, aiming to facilitate bringing down the language barriers, bridging the language divide between information seekers and the wealth of resources from around the world that are made available through such portals. The main research question addressed is the definition of the problem itself and the description of design affordances that shall allow web portals to meet different linguistic profiles, needs and expectations from users around the globe. The paper describes the methodology used for the analysis of user requirements of the international user base of a thematic web portal aiming to promote discovery of educational resources and concludes with some initial results, discussion of the findings and suggestions for future work.

Keywords—multilingual; multilingualism; translation; portal; organic; agriculture; education; linguistic; internationalisation;

I. INTRODUCTION

“In the galaxy of languages, every word is a star.” According to the United Nations Educational, Scientific and Cultural Organisation (UNESCO) [1], information and knowledge are key determinants of wealth creation, social transformation and human development. Language is the primary vector for communicating knowledge and traditions, thus the ability to use one’s language on the Internet could determine the extent to which one can participate in emerging knowledge societies. Developing policies to advance multilingualism, including in cyberspace, is essential towards reducing this “language divide”.

The research work presented in this paper involves the definition of existing linguistic barriers, the different linguistic requirements of various user categories and the possible ways to meet them when designing multilingual web portals. In fact, the objective of this paper is two-fold: to provide a generic, analytic framework for addressing the requirements related to multilingualism of web portals, as well as to present a case of needs analysis for an existing web portal, currently in operation since 2010. The Organic.Edunet portal (http://organic-edunet.eu) couples together a network of associated repositories of learning objects on Organic Agriculture and related topics, such as sustainable agriculture, agro-ecology, aquaculture, etc. From its initial conception, design and implementation, Organic.Edunet has adopted a federated, standards-based approach that facilitates the incremental growth of the network along with the integration of more sophisticated services for its end users [2]. The portal is currently available in sixteen languages, thus already covering the needs of users from around the globe. Moreover, the knowledge representations recorded in the federated repositories (ontologies and metadata schemata) have been manually translated by human experts to several languages. However, the portal’s features and the facilitated access to an ever-expanding, quality-controlled catalogue of resources result in user communities from even more countries becoming increasingly interested in using the portal.

This success has been the basis for an interesting question: What are the current language barriers for the international community of the portal’s users and which are the best approaches for identifying and validating the needs of those, in order to break down the language barriers and facilitate access of all to the wealth of resources. This challenge is almost in direct alignment with the “Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace”. Unanimously adopted by its Member States in October 2003, UNESCO’s recommendation advocates for a multilingual cyberspace by focusing its interventions in three areas: (1) including new languages on the Internet; (2) creating and disseminating content in local languages in cyberspace; and (3) providing multilingual access to digital resources.

The following parts of this paper provide an insight of our work for capturing, analysing, documenting and organising the user requirements of the international user community of Organic.Edunet, along with some initial findings. In particular, section 2 presents the background for the problem at hand and related work for multilingualism on the web, section 3 describes the overall methodology for the analysis of user requirements, the identification of stakeholders, their current approaches for creating and locating learning resources and the tools they use, section 4 documents some initial results, with details of the user sample and the audience and section 5 concludes with a discussion of the findings and indication for future work.

II. BACKGROUND

In computing, internationalisation and localisation are means of adapting computer software to different languages,
regional differences and technical requirements of a target market. Two commonly used “numeronyms”\(^1\) are used in short for both terms: i18n and L10n\(^2\). i18n is the process of designing a software application so that it can be adapted to various languages and regions without engineering changes. L10n is the process of adapting internationalised software for a specific region or language by adding locale-specific components and translating text. These definitions lead us to multilingualisation, or m17n: support of multiple languages by computer systems can be considered a continuum between L10n, through m17n, to i18n.

These definitions for i18n, L10n and m17n, although generic and concerning all kinds of software, have a slightly different perspective when referred to web sites, web content, web applications and web portals. It would be useful however to first discuss the definition of web portals. According to [3], a portal is a doorway, entrance or gate, especially one that is large and imposing. In fact, the more frequent usage for the term has been in science-fiction, and some “New Age” philosophies, for describing a gateway to another world of the past, present, or future, or to an expanded awareness, a matterless vortex used to travel between different dimensions, including the astral world. To a certain degree, the term has retained the meaning of its original usage when started to being used, in the late ’90s, during the internet frenzy, to describe a website that is considered as an entry point to other websites, often by being or providing access to a search engine [4]. In fact, according to Dewan et al [5] in order to attract traffic portals provide value added services to the users: they continuously scan the web for relevant and timely information, screen and prioritise the links, and provide a consistent interface. It is exactly this “added-value” that makes portals among the most successful web sites: the aggregation, annotation, selection, categorisation, quality control of resources scattered at different locations, all over the world. But it is exactly these characteristics and qualities that make support for m17n even more demanding for portals: interesting resources, being aggregated by international collections, addressed to an international audience, call for an extensive selection, categorisation, quality control of resources scattered at different locations, all over the world. But it is exactly these characteristics and qualities that make support for m17n even more demanding for portals: interesting resources, being aggregated by international collections, addressed to an international audience, call for an extensive selection, categorisation, quality control of resources scattered at different locations, all over the world.

This is, therefore, an interesting issue to challenge: how to prepare a web portal for to support multilingualism? Or, to put it in other words: what are the affordances that will make a web portal an efficient and effective instrument for information seekers around the globe? The problem, thus, becomes a problem of definition of the affordances. But, what is really an affordance? The word “affordance” itself was originally invented by the psychologist J. J. Gibson [6]. Gibson uses the term to refer to the actionable properties between the world and an actor (a person or animal). To Gibson, affordances are a relationship. They are a part of nature: they do not have to be visible, known, or desirable. Norman [7], however, suggests that when it comes to design, the term “perceived affordance” should have been used, since the focus is much more about what the user perceives than what is actually true. In any matter however, an affordance is a quality of an object, or an environment, which allows an individual to perform an action.

Having set the semantic context, we can move back to the challenge web portals and m17n, employing the “divide and conquer” paradigm [8], breaking down the problem into three pieces in order to easier address each one separately:

- a problem of collecting, aggregating and organising multilingual data (resources);
- a problem of adequately defining data structures for annotating those resources (metadata); and
- a problem of providing an effective, multilingual user interface for our global audience.

None of these three problems is easy to address. On-going research for almost the last fifteen years has provided some answers and technical solutions, but it has also raised more issues, making apparent more challenges, along with their complexity and interdependency. The objective of m17n is to naturally enable users to work in a culturally and linguistically familiar computer environment, which is thus easy to master. The users language thus becomes a working language of computer use. Software localisation in a poorly endowed language thus helps to enhance the prestige of that language in the eyes of users and, in particular, of its own speakers [9].

In particular, with regards to the Organic.Edunet web portal, new m17n features are expected to further facilitate users who search for educational content, making it easier for them to access all resources relevant to their topic of interest. Fragmentation of knowledge as a result of language barriers is expected to be significantly reduced. Also, the cost-effectiveness of the translation work will be increased by providing support in the process of labels, metadata and resource descriptions. Furthermore the generation of suggested descriptive metadata from text and translations is going to increase the completeness of metadata.

Since its official launch in January 2010, the Organic.Edunet portal has attracted more than 76,500 unique visitors from 184 different countries. This traffic towards the portal was expressed as more than 94,700 visits and 327,700 page-views, with an average time on portal of more than 2 1/2 minutes per visitor [10].

---

1 A numeronym is a number-based word (source: Wikipedia). Such words begin with the first letter, have number of letters in the middle, and end with the last letter.
2 The capital L in L10n helps to distinguish it from number one (1).
III. METHODOLOGY

Given the identified needs and linguistic background of the Organic.Edunet users, our main aim related to 
ml7m of the portal is in providing a wider and more efficient multilingual access and effective location of the learning resources available through the federated repositories. Our approach to this end is to analyse and re-engineer the related facilities and associated business models in order to integrate multilingual access, effective localisation and automatic translation services for the aggregated metadata descriptions and learning objects. The main targeted audiences of Organic.Edunet fall in the following categories:

1) School teachers involved in teaching of Organic Agriculture and related topics, either directly (e.g. through environmental education activities) or indirectly (e.g. through educational activities on cross-disciplinary topics like biology, chemistry, economics or history).
2) Academics and researchers involved into teaching, tutoring or researching related topics.
3) Learners and Practitioners, interested in locating useful material on the topics addressed by the portal.
4) Content providers and other stakeholders involved in the creation, production, organisation and/or publication of content around Organic Agriculture and related topics, including a wide variety of users such as academic and research institutions, EU-funded initiatives, public organisations, private not-for-profit organisations, agricultural libraries or publishers, existing portals and repositories, as well as end-users themselves (farmers and other communities of practice).
5) Last but not least, a main stakeholder is the community of developers, technology providers and integrators of machine translation tools, services and linguistic technology in general, as well as other related portals and/or online services that could benefit from the foreseeable new services of the Organic.Edunet portal.

In order to address all identified stakeholders and engage them in activities towards capturing the most relevant and important feedback, a series of events has been organised and carried out. All activities were performed within a carefully planned methodological framework and fall into three categories, which are described in the following paragraphs.

A. Consultation meetings with new content providers

The first type of activities concerned a number of stakeholder meetings in order to explore and document the requirements from the content providers that participate in the Organic.Edunet network of federated repositories. The content providers, being one of the main stakeholders, have been identified as a major source of feedback and input for requirements elicitation. These activities have provided valuable feedback through a number of events, both structured in the form of interviews and small-scale focus groups, as well as unstructured in the form of e-mail exchange and skype discussions. Moreover, end-users have been also consulted, since there has been an expressed interest for providing content by individuals, in the form of “user-generated” content. The main objectives for the consultation meetings have been: (1) study how the newly introduced multilingual components can be integrated in the existing content population procedures of all repositories, leading to a new, revised content population procedure, which will take into consideration the new components; (2) study how the revised content population procedure can be adopted by new content providers; and (3) study how the content population procedure can be adapted for user-generated content. Our team, being responsible for the organisation and coordination of the consultation meetings, prepared a description of the approach and the methodology that was made available online as a set of wiki pages. In fact, those wiki pages have been also used as a means of reporting the feedback from each event and have proved invaluable for updating the instructions and the relevant material as the work was progressing.

B. Consultation meetings with new user organisations

The second type of activities concerned the preparation, organisation and reporting on a number of focused meetings in order to explore and document the anticipated requirements from the user communities. The aim of the events with end-users, mainly in the form of workshops, has been to engage those users into participatory sessions, in order to collect feedback that will help us better understand the current shortcomings of the portal and the limitations that render it unusable for certain categories of people who would not manage to use it before specific support and motivation is provided. Therefore, the main objectives of the workshops have been to: (1) understand the “Language Barriers” between various categories of end-users and related resources; (2) briefly introduce Organic.Edunet and the associated content federated through various collections, along with the current search and browsing mechanisms; (3) validate already identified requirements and record any possible new, anticipated features; and (4) motivate people to follow the activities of our team and join the Organic.Edunet community.

C. Online survey of potential users

The last activity has been an online questionnaire-based survey, in order to reach as many related people as possible, among the identified stakeholders and safeguard that all possible feedback is collected and used for our requirements analysis. The survey actually managed to engage people from all over the world, with an interest in Organic Agriculture and related fields in general, and for educational resources more specifically, who wouldn’t have otherwise participated in the requirements elicitation and validation
activities. The survey was focused on the recognition of user needs expectations, relation to the thematic coverage of the resources that are federated through the Organic.Edunet platform, the current linguistic barriers of the targeted users for locating and accessing those resources and requirements concerning multilingualism regarding the portal’s user interface and the need for new features and services. The survey has been based on a questionnaire, which has been carefully designed in order to capture as more feedback as possible without imposing an excessive time requirement thus hindering completion.

The questionnaire has been created with the use of the Lime Survey open source tool and was made available online. It was initially designed during November 2011 and was pilot tested in order to review for possible shortcomings and revise to fully address the intended questions. After this revision, a translation process has been employed, through the usage of a combination of online tools, including wiki pages, automatic machine translation and, coordinated manual translation, in order to cope with the difficult task of the translation of the questionnaire into several languages. The targeted languages were selected according to the Organic.Edunet portals web analytics and the indicated traffic from different countries, as well as according to the communities that have already expressed interest or are considered important, mainly due to their user-size. The result of this process was more than satisfactory: the questionnaire was made available in sixteen languages. The availability of the questionnaire in many languages guarantee that we were able to get feedback from a more representative sample of users and identify differences between users with an advanced level of proficiency in English and those who have a low level or do not speak English at all.

The next step of this procedure has been to integrate all translations in the Lime Survey tool and re-test all the language versions of the questionnaires. This activity took part in January 2012 and the final versions of the questionnaire were then pilot tested by a small set of end users during the first half of February. The survey was then officially launched in February 19, 2012, when a total of 925 responses, of which 834 were fully completed, were received. The survey was focused on the recognition of user needs expectations, relation to the thematic coverage of the resources that are available through the portal. For the sake of estimate, we assume that this targeted population is approximately 5,000 people, including individuals from all targeted stakeholders and from all countries. It is interesting to note that this number is almost equal to the number of people that have registered a user account during the first two years of the web portal’s normal operation. Another important concern about the expected participants and the associated selection of the sample has been that of the focus of our survey: indeed, the focus being on language barriers and multilingualism, we have emphasised in the inclusion in our sample of potential users not only from countries that are currently covered (at least to some extend) by the portal and available resources, but rather aim to include people who are majorly impacted by the language barriers and the lack of support and content suitable for their linguistic profiles.

The large number of targeted users meant that we had a large sample that could give us a solid picture of the users’ requirements. At the start of the survey a goal of a minimum of 500 replies was set (10% of the targeted users), a sample for statistical analysis of the results. The overall feedback exceeded our expectation, reaching by mid-May a total of 925 responses, of which 834 were fully completed, thus providing an apparently proper basis for the analysis that follows. The feedback was supported by the distribution of the participants, their native language being among an impressive list of 24 different languages.

IV. RESULTS

While at the time of writing of this paper the online survey is still in operation, it is already providing interesting feedback and responses to our questions. The diverse profile of the respondents fits the wide range of different categories of Organic.Edunet’s members. This fact also allows us space for comparisons between different user categories, thus facilitating understanding of the different needs of each group. Fig. 2 presents the details about the linguistic profile of the respondents. It is obvious that a vast majority has a competency in English, totalling to a 82.35% (5.88% mother tongue + 76.47% second or third language competency). On the other hand, only a very small minority of 10.29% has no other competency other than their mother tongue. If we subtract the English respondents among those, this relative frequency becomes even lower, adding up to a 6.62%. This fact can be an indication that our portal’s members are people who feel competent in using English and, in general, have a linguistic profile of an above average competency level of that expected from the portal’s intended audience. This indication could be supportive of the hypothesis that people who face language barriers are not using the portal and cannot benefit from its services and learning resources.

It is also interesting to note the main difference between the users that use the English version of the portal and the
ones that use the one on their own language; the former also search in English while the latest do not. This also supports the claim made regarding teachers and researchers. We can also observe another major difference by taking a look at the replies in Q3.3 “Is it easy to locate available content in other preferred languages?” As researchers are more used to using English this process seems significantly easier to them whistle it is much more difficult than the average for teachers.

The survey revealed several more interesting findings, some of which had been anticipated, while others were not expected. A first important finding for our work was the existence of two distinct groups of users - one using the English version of the portal, searching in English and being constitute mostly by researchers and the other using the version of the portal in their own language and preferring to search using terms in their mother tongue. It is interesting to notice here that a reason why researchers seem to be more accustomed to searching in English might be the fact that most of the content available through Organic.Edunet is actually in English. As a result only researchers with a substantial knowledge of this language can benefit from using it, and thus, are potential users interested in the portal’s services and resources.

Another major finding is that a lot of users prefer navigating using the English version of the portal mostly, even though there’s a version in their mother tongue, for the following three reasons: (a) availability of resources; (b) personal preferences / habits; and (c) quality of translation. Reasons (a) and (c) along with the fact that most users reported that sometimes they have to guess an English term to find the content they need are clear marks that a better approach towards multilingualism is needed, especially in allocating content in different languages. It is worth noting here that despite the existence of mistakes the quality of the translations available was marked as being very high.

Finally, users have reported that they would mostly appreciate services that would allow automatic translation of their search terms to other languages (e.g. English), the existence of a multilingual Organic Lexicon (with terminology translated to all languages) and being able to use the Semantic search mechanism in their mother language. On the other hand, most users would not be very interested in a feature to rate translated material and for others to rate their translations, image based search mechanisms and a feature for the translation of comments made by users. These results are in alignment with the findings of the learning analytics research that is trying to investigate issues related to the use of multiple languages, reported in [11].

V. DISCUSSION AND FUTURE WORK

Quite a large number of active members of the portal, as well as respondents to the survey would chose to use the original language version of a web site, even when this is not in their native language. In many cases, users who are not apparently native English speakers, originating for example from India or Germany, would prefer to use the English version of the portal. It is certainly an issue that needs further investigation, certainly related to common errors that are introduced by the L10n of the portal and the inefficiency of the i18n mechanisms. A feature that would allow for easy reporting on such errors and issues, capturing users’ feedback without asking for extensive effort or time from their behalf could certainly help towards addressing this issue.

Another interesting finding is that services that ask for increased user interaction, especially the ones with no-immediate benefit for the users, are among the ones with the lower score in terms of anticipated importance. This is evident with the responses for “A feature to rate translations made by other users and for others to rate your translations”, and “A feature to participate in the translation of metadata of existing resources”. This finding could provide serious motivation for trying to introduce new-comers to the benefits of full engagement and active participation in the portal, as registered members. Only then would services “that allow automatic translation of comments made by other users” or “translation of descriptive tags from other users to your native language” would become more apparent and important. Related to multilingualism, a strong incentive for user registration would be an intuitive mechanism for storing linguistic preferences that would combine information and feedback collected from the interactions of the user with the portal. This collected feedback could be temporarily stored at the user-side, as a cookie, but could be also form the basis for a permanent user profile, that would persuade users to take the extra step and register with the portal. Another apparent motivation for user registration could be closely related to the search mechanism, allowing registered users to store and recall saved searches, either in terms of search terms and queries or, in a later stage, in terms of full support for management of the results-set (i.e. marking of results as relevant or irrelevant, for further investigation, etc.) Last, but not least, access to features such as machine translation services for translation of documents or suggestions for user-generated content should be only available for registered users.
<table>
<thead>
<tr>
<th>Mother Tongue</th>
<th>Rel. Freq.</th>
<th>None</th>
<th>English</th>
<th>French</th>
<th>German</th>
<th>Italian</th>
<th>Spanish</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>1.47%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>10.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>1.47%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Czech</td>
<td>0.74%</td>
<td>100.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>English</td>
<td>5.86%</td>
<td>62.5%</td>
<td>0.0%</td>
<td>25.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>25.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Estonian</td>
<td>16.28%</td>
<td>0.0%</td>
<td>95.5%</td>
<td>0.0%</td>
<td>2.3%</td>
<td>0.0%</td>
<td>4.5%</td>
<td>50.0%</td>
</tr>
<tr>
<td>Finnish</td>
<td>0.74%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>French</td>
<td>2.21%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>33.3%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>German</td>
<td>7.35%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>40.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>20.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Greek</td>
<td>16.28%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>18.2%</td>
<td>0.0%</td>
<td>9.1%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Hindi</td>
<td>2.21%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Italian</td>
<td>0.74%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Korean</td>
<td>0.74%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Portuguese</td>
<td>1.47%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>50.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>50.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Romanian</td>
<td>5.86%</td>
<td>25.0%</td>
<td>75.0%</td>
<td>12.5%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>12.5%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Russian</td>
<td>15.44%</td>
<td>23.8%</td>
<td>28.6%</td>
<td>4.8%</td>
<td>9.5%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Serbian</td>
<td>1.47%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>50.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Spanish</td>
<td>13.24%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>36.7%</td>
<td>11.1%</td>
<td>5.6%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Turkish</td>
<td>6.62%</td>
<td>11.1%</td>
<td>88.9%</td>
<td>0.0%</td>
<td>11.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

For example 76.47% of our sample has some competency in English

For example 100% of the person whose mother tongue is Arabic have a competency in English and 62.5% of those whose mother tongue is English respondent no other language competency

Figure 2. Linguistic profile of survey's sample

ACKNOWLEDGEMENTS

This paper includes research results from work that has been funded with support of the European Commission, and more specifically the project CIP-ICT-PSP-270999 “Organic.Lingua: Demonstrating the potential of a multilingual Web portal for Sustainable Agricultural & Environmental Education” of the ICT Policy Support Programme. The authors want to express their gratitude towards all partners of the Organic.Lingua project, for their support in all work reported in this paper, including the translation of the questionnaire and its promotion among local user communities.

REFERENCES


