Preservation of Digital Publications – An OAIS Extension and Implementation

ACM Symposium on Document Engineering
Grenoble
November 20-22, 2003

Peter Rödig
Uwe M. Borghoff
Jan Scheffczyk
Lothar Schmitz
Institute for Software Technology
University of the Federal Armed Forces, Munich
peter.roedig@informatik.unibw-muenchen.de
Introduction to the (long-term) Preservation Problem - A layered Model

- Physical Damages
- Obsolescence
- Missing technical Knowledge
- Missing cultural Knowledge

- Media Layer
- Machine Interaction and Interpretation
- Human Interaction and Interpretation
- Interpretation of Intellectual content
OAIS Reference Model – Introduction

Open Archival Information System Reference Model (OAIS) [ISO 4721:2002 ]

CCSDS Consultative Committee for Space Data Systems

Core Features:

• A generic framework, including terminology and concepts, for describing and comparing architectures and operations of existing and future archives

• Independence from concrete preservation strategies

• Not a specification of a design or an implementation
OAIS Reference Model – Functional Elements

SIP: Submission Information Package
AIP: Archival Information Package
DIP: Dissemination Information Package
OAIS Reference Model + Extensions of OCLC/RLG WG – Static Elements (subset)

Information Object \{CI, PDI, PI, DI\}

Knowledge Base
  External to the archive

Data Object
  Stream of bits

Representation Information
  Makes data meaningful

Data Object Description

Environment Description
Why extending OAIS?

Aim

Support of Archiving Tasks in general and scientific Libraries

Considerations

- Numerous Publications delivered on Media
- Complexity and Heterogeneity of Publications
- Old Material (~ 30 years)
- Active Archives required
- Mixed Consumer Community (technical know how, support)
- Limited Influence over Producers
  - Market and technological Progress rule Media and Formats
  - Metadata are missing, or implicit, or not uniform
  - Black Boxes („autorun will do it for you“)
Focus of Discussion

Detachment Creating SIP Ingest

Creating CI Creating AIP Creating PDI

Extension to OAIS Archive

Implementation
Extending OAIS – The Detachment Process

Rational

• Increased Media Independence
  by reducing the number of interpreting instances (devices, software) above the media layer

• Improved Online Availability (Administration and Access)
  by storing digital publications on media supported by standard database management systems
  (see Implementation - The Database Approach)
Digital Publications – A Conceptual Model

Detaching and Creating SIPs and AIPs

Considerations

Role and Granularity of Digital Objects/Sub-Objects

• Identification (*Reference Information* and *PID*)
• Relations (*Context Information: Manifestation* and *References to Intellectual Content*)
  • Redundancies (semantic, technical) (reduction of environments, error handling)
  • Dependencies (semantic, technical) (partly migration, error handling)
• Preservation Metadata (Level for *PIDs* and *Representation Information*)
Digital Publications – A Conceptual Model

Scenario 1

External View

Information Entity

Digital Entity

Digital Entity

Nondigital Entity

Windows95

Windows95

MacOS

Scenario 2

Information Entity

Digital Entity

Windows95

MacOS

External View

Internal View
Digital Publications – A Conceptual Model

External View

• Relations between Semantic Elements and Medium Elements ?
• Intellectual Content (Mediumtype 1) = Intellectual Content (Mediumtype 2) ?
• Intellectual Content (Operating System 1) = Intellectual Content (Operating System 2) ?
• Look and Feel Part of Intellectual Content ? (*PDI: Intellectual content vs. Manifestation*)
• Any Shared Objects ? (e.g., datafiles used by different OS-Applications)
• Any References to Medium Elements ?

Not generally decidable (at least in practice)

=> Medium Elements are handled as identifiable Semantic Elements

Medium Elements: {Volume, Dimension} in UML-Model
Digital Publications – A Conceptual Model (ONGOING WORK)

Internal View

• Relations between Semantic Elements and Medium Sub-Elements?
• Intellectual Content (Medium Sub-Element 1) = Intellectual Content (Medium Sub-Element 2)?
• Intellectual Content (Operating System 1) = Intellectual Content (Operating System 2)?
• Look and Feel Part of Intellectual Content? (PDI: Intellectual Content vs. Manifestation)
• Any Shared Objects? (e.g., datafiles used by different OS-Applications)
• Any References to Medium Sub-Elements?

Note: Modern media internalize elements of the external view (hybrid)

Not generally decidable (at least in practice)

• What types of Medium Sub-Elements should be modelled as Semantic Elements?
• What is the right Level of Decomposition of Semantic Elements?
• How is the Influence of Preservation Strategies like Emulation, Migration, Virtual Machines?

Medium Sub-Elements: {Session, Track, Sector, Lead-In, ..., Byte}
Semantic Elements: {Volumedescriptor, Filesystem, File, File Component (hybrid file), TOC, ...}
Digital Publications – Detachment Process

Digital Entity

Information Entity

Digital Entity  Digital Entity  Nondigital Entity

Windows95  Windows95  MacOS

Images (in some cases different versions)

Detachment

Digital Entity

Information Entity

Digital Entity

Windows95 + MacOS

Detachment
Digital Publications – Detachment and Creation of SIPs and AIPs

Detachment Tool

Detachment Creating SIP Creating AIP

Media

SIP

Preservation Metadata: Mediatype, [Formatting, OS, FS]

SIP PI / IDs

Directory+ File List

(Error) Protocol

Image

AIP

Preservation Metadata: Mediatype, [Formatting, OS, FS]

AIP PI / IDs

Directory+ File List

(Error) Protocol

Image

Analyzing Tools

Preservation Metadata

Knowledge !?

(1) Decomposition (2) Completing PDI and Representation Information

Preservation Metadata

Preservation Metadata
Implementation of OAIS – The Database Approach

Core Feature of the Database Approach

The DBMS controls the complete AIP including *Content Information* (CI)

Rational

- Higher Level of Abstraction w.r.t. Storage Media and File Systems
- Tight Integration of OAIS's functional Elements *Data Management* and *Archival Storage*

Flexibility and Security

- Analysis of CI anytime
  
  (=> deferred Generation of Preservation Metadata or Discovery of Intellectual Content)
- Reorganisation of Information Packages including Identification
- Consistency between and within Information Packages
- Support of semi-structured documents and mixed Collections (XML)
Lessons Learned
from our conceptual work and experiments

Detachment
• Quality, heterogeneity and obsolescence of media require a restart
• Quality of tools is not sufficient

DBMS
• Database paradigm is a central issue (alternative solutions in future?)
• Storage abstraction is not perfect
• Schema changes need long-term considerations
• DBMS vs. file system is still an issue
• Identification of digital objects is critical

In general
A knowledgebase is necessary (for archiving and/or consumer community)
Thank You

merci beaucoup