IPA Cross-Border PROGRAMME
Greece - Albania
2007-2013

Balkaneana

Digital Libraries: issues and best practices

The Project is co-funded by the European Union and by National Funds of Greece & Albania
under the IPA Cross-Border PROGRAMME "Greece - Albania 2007-2013"

Introduction

- A *digital* or *electronic library* is a collection of *digital objects* along with the software for storing, managing and retrieving the objects.
- The digital objects may include text, audio and video.
- Digital libraries provide an effective means for easily accessing content from anywhere and anyplace.
- Digital libraries play an important role in preserving human knowledge.

Introduction

- A digital library is a type of *information retrieval system*.
- *Information retrieval* (IR) is the activity of obtaining information resources relevant to an ***information need*** from a *collection* of information resources.
- Besides the actual digital objects, a digital library maintain additional information along with each digital object, called *metadata*, to facilitate their management.

Metadata

Metadata include

(i) information describing the fields and content of the object (similar to catalogue records in traditional libraries)

(ii) information necessary for curating the object such as its access rights and

(iii) information about the objects structure (e.g., division in chapters).

Searching

- A user ***information need*** is expressed through a *search query*.
- Queries vary in complexity from simple keyword queries to semantics-based ones.
- For example, in a digital library, a query may refer to the ***descriptive metadata*** associated with a digital object, such as the title, author, publisher of the object, or, keywords associated with the content of the object.
- Some systems also allow ***full-text*** search.
- More advanced query may allow ***semantic-based search***, for example, by referring to similar objects or concepts related with the object

Important considerations

Goals:

- allow uniform access to multiple digital collections. To achieve this goal, various standards have been developed to support interoperability.
- provide user-friendly interfaces to both librarians and users.

Digital vs Physical Libraries (1)

- The content of a digital is available from *anyplace*, that is, there is *no physical boundary*. There is no need for a user to go to the library physically; people from any place in the world can gain access to the same information, as long as an Internet connection is available.

- The content of a digital library is available *anytime*, that is, there is *no time boundary*. The content of a digital library is continuously available to its users.

- Subject to copyright issues, a digital object may be *accessed simultaneously* by a number of institutions and patrons.

Digital vs Physical Libraries (2)

- In a traditional library, to locate a relevant book, a user has very limited choices: use the catalogue information, physically search in the bookshelves, or ask advice from a librarian. Digital libraries offer *information retrieval enhanced search*. A user may search the whole collection for all books containing a specific term, phrase, title or subject. Many systems also support semantic search, recommendations and user-friendly navigation interfaces.

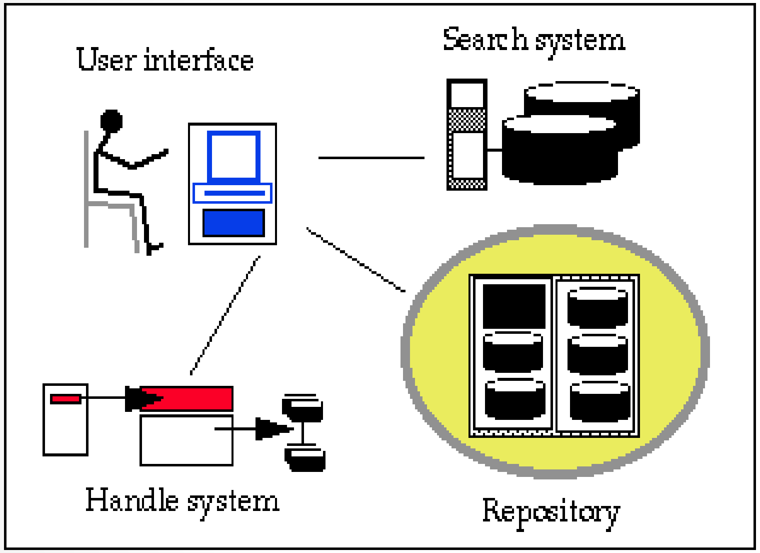
- Digitization allows the *preservation* of books and other materials for which the quality of their physical copies may deteriorate due to time and repeated use.

- Digitization may *enhance the quality and legibility* of images and text by restoring to some extent the original copies by, for example, removing visible flaws such as discoloration.

- In a traditional library, the storage space is limited, whereas in a digital library the collections can grow *without a space limit*.

IPA Cross-Border PROGRAMME
Greece - Albania
2007-2013

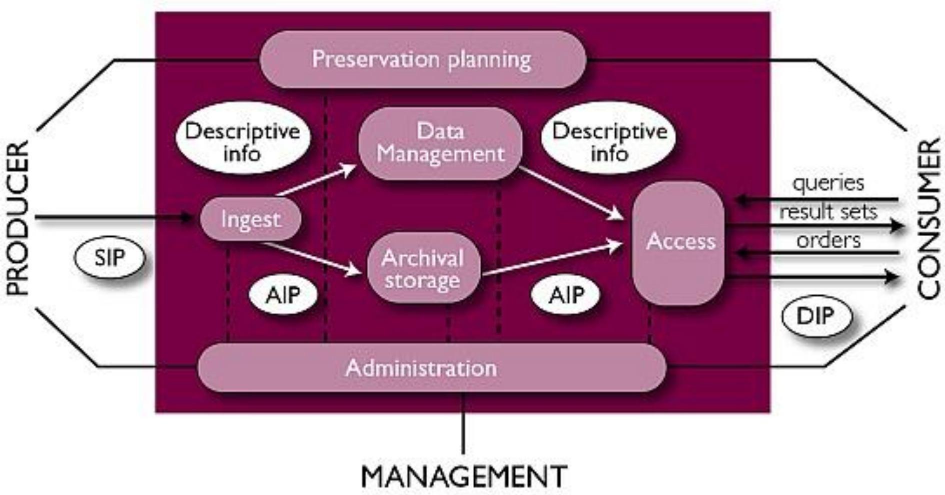
Architecture of a DL



•A *repository* that stores the digital objects

•A *handle system* that manage the digital objects
•A *search system* that supports searching for digital objects of interest
•*User interfaces that* provide easy access to the system.

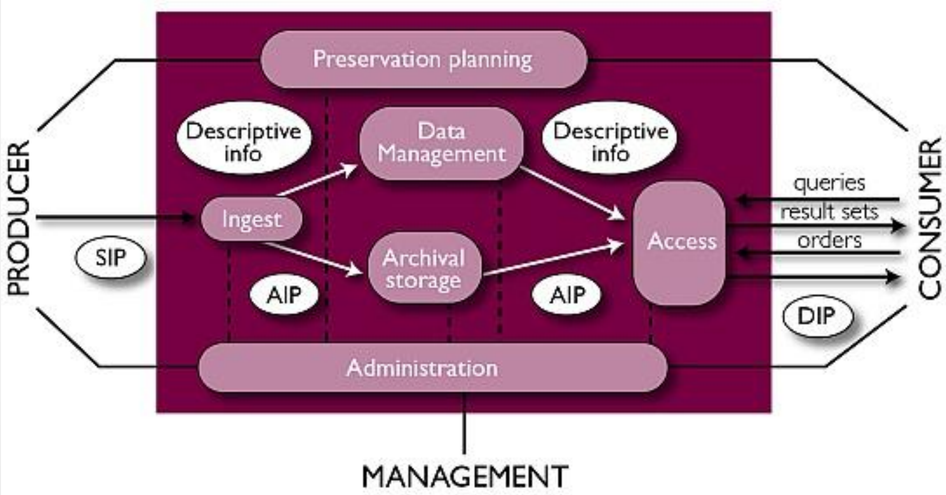
Reference Model



- A useful categorization and reference model for digital preservation systems is provided by the *Open Archival Information System (OAIS)*. OAIS is widely adopted as a conceptual model and framework for a digital repository and archival system.
- There is a number of functions that an archival digital repository must perform. OAIS defines them as Ingest, Access, Administration, Data Management, Preservation Planning and Archival Storage.
- OAIS also defines the structure of the various information packages that are necessary for the management of data.

http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284

Reference Model



- The *Submission Information Package (SIP)* is the content and metadata received from an information producer by a preservation repository.
- An *Archival Information Package (AIP)* is the set of content and metadata managed by a preservation repository, and organized in a way that allows the repository to perform preservation services.
- The *Dissemination Information Package (DIP)* is distributed to a consumer by the repository in response to a request, and may contain content spanning multiple AIPs.

Metadata for libraries (1)

The MARC standards are a set of digital formats for the description of items catalogued by libraries developed in the 1960s by computer scientist Henriette Avramd along with Library of Congress.

The MARC standards define three aspects of a MARC record:

(i) *The field designations within each record*: each field provides particular information about the item described by the record including the author, title, publisher, date, language and media type.

A simple three-digit numeric code (from 001-999) is used to identify each field in the record, for example field 260 corresponds to the publisher. Fields above 008 are further divided into subfields using a single letter or number designation. For example field 260, , is further divided into subfield 'a' for the place of publication, 'b' for the name of the publisher, and 'c' for the date of publication.

Metadata for libraries (1)

(ii) *The structure of the record*: several records are typically stored and transmitted as binary files.

To define the structure of each record, either the ISO 2709 standard, or XLM (MARCXML) is used. The ISO 1709 standard includes a marker to indicate where each record begins and ends, as well as a set of characters at the beginning of each record that provide a directory for locating the fields and subfields within the record. In the MARCXML schema the fields remain the same, but are expressed in the record in XML markup. Libraries typically expose their records as MARCXML via a web service, often following the SRU or OAI-PMH standards.

(ii) *The actual content of the record itself*: The actual content is usually defined by standards outside of MARC, except for a handful of fixed fields such as Resource Description and Access that defines how the physical characteristics of books and other items should be expressed. The Library of Congress Subject Headings (LCSH) are a list of authorized subject terms used to describe the main subject content of the work.

Metadata for digital libraries (1)

A useful typology for digital library metadata includes [Ga08]:

- *Descriptive metadata*: similar to the tradition catalogue record, that is, information on the fields and contents of the item to facilitate searching for an item.
- *Administrative metadata*: information necessary to curate the digital item, that includes, for example
  - *Technical metadata*: the necessary technical information (for example, file formats) to allow the host system to store and process the item
  - *Rights management*: specification of the rights held in the item and information necessary to restrict its delivery only to those entitled to access it
  - *Digital provenance*: information on the creation and subsequent treatment of the digital item.
- *Structural metadata*: information necessary to record the internal structure of an item so that it can be rendered to the user in its correct form (for instance, a book must be delivered in its page order). This type of metadata is necessary since a digital object may often be composed of multiple files (e.g., images of a digitized book).

Metadata for digital libraries (2)

- For digital libraries, as part of the MARC Standards, the ***Metadata Encoding and Transmission Standard (METS)*** is a metadata standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library.
- METS is intended to promote the preservation and interoperability between digital libraries.
- METS is being developed as an initiative of the Digital Library Federation (DLF).
- Depending on its use, a METS document could be used in the role of Submission Information Package (SIP), Archival Information Package (AIP), or Dissemination Information Package (DIP) within the Open Archival Information System (OAIS) Reference Model.

Metadata for digital libraries (3)

- *Header metsHdr*: information about the METS document itself, such as its creator, editor, etc.

- *Descriptive Metadata dmdSec*: descriptive metadata (such as information about the author, title, etc)

- *Administrative Metadata amdSec*: how files were created and stored, intellectual property rights, metadata regarding the original source object from which the digital library object derives, and information regarding the provenance of files comprising the digital library object (such as master/derivative relationships, migrations, and transformations).

```
<mets>
   <metsHdr/>
  <dmdSec/>
  <amdSec/>
  <fileSec/>
  <structMap>
  <structLink>
   <behaviorSec>
</mets>
```

Metadata for digital libraries (4)

- *File Section fileSec*: Lists all files containing content which comprise the electronic versions of the digital object.

- *Structural Map structMap:* Outlines a hierarchical structure for the digital library object, and links the elements of that structure to associated content files and metadata. The Structural Map is the only section required for all METS documents.

- *Structural Links structLink:* Allows METS creators to record the existence of hyperlinks between nodes in the Structural Map. This is of particular value in using METS to archive Websites.

- *Behavioral behaviorSec:* Used to associate executable behaviors with content in the METS object. Each behavior has a mechanism element identifying a module of executable code that implements behaviors defined abstractly by its interface definition.

```
<mets>
   <metsHdr/>
  <dmdSec/>
  <amdSec/>
  <fileSec/>
  <structMap>
  <structLink>
   <behaviorSec>
</mets>
```

Ontologies

An *ontology* defines a set of representational primitives with which to model a domain of knowledge or discourse [Gr09].

A domain is a specific subject area or area of knowledge, such as, for example medicine, tool manufacturing, real estate, automobile repair, financial management, etc.

Due to their independence from lower level representations, ontologies are used for integrating heterogeneous databases, enabling interoperability among different systems, and specifying interfaces to independent, knowledge-based services

The representational primitives of an ontology are typically:

- Classes (or sets),
- Attributes (or properties), and
- Relationships (or relations among class members).

Search (1)

- An important part of any library is the ability to locate works of interest. In traditional libraries, this ability depends on how well the items are catalogued and the knowledge of the librarians. Digital libraries offer to each users a variety of means for locating the information of interest to them.

There are two basic types of search:

(i) Text-based (lexicographic) search: Classic text-based search uses keywords as query terms.

These keywords may appear in the ***descriptive metadata*** describing the object (such as their title or authors). In case in which the target object is a digitized text document, the keyword may also appear anywhere in the document (***full-text search***).

At a pre-processing step, the digital objects and their metadata are analysed and an index is built on top of them. During search, the index of keywords is used to identify matching digital objects.

Search (2)

(ii) **Semantic search:** In this case, there is a metadata description of the digital objects using for example RDF.

Traditional semantic search uses a semantic query language (e.g., SPARQL, which is the query language for RDF). Each RDF object contains references to its concepts defined in one or more ontologies.

At a pre-processing step, the digital objects are analysed for concepts, as defined in the ontologies. Then, indexes are built based on both the textual content, the metadata and the concepts that the ontologies contain.

During search, the semantic query is executed on the ontology index to identify the containing concepts and the digital object index to identify the objects referring to these concepts.

Digital vs Paper Reading

- Despite the advances in digital reading, paper still plays an important role in reading.
- Material in printed form is more resilient to the physical elements, such as water and wind, is relatively cheap to replace and easy to annotate.
- However, there is an increasing trend towards digital reading. One of the most important reasons is search.
- Digital books and articles allow a user to search and locate the occurrence of any word and phrase in the text without any effort.
- Furthermore, digital reading offer the possibility of continuous access from a variety of devices.

Reading Activities

Reading to identify (ID)	Glancing at a document only to identify which document it is.
Skimming (SK)	Reading rapidly to get a rough idea of what is written and decide on whether there is something to be read in detail later
Reading own text to remind (REM)	Reading specifically to remind oneself of what to do next (e.g., shopping or to-do lists)
Reading to search, answer questions (SAQ)	Reading to search for specific information (e.g., to answer a question, make a decision, etc.); goal-directed reading
Reading to self-inform (SI)	Reading for attaining general knowledge without any specific goal to which to apply the information
Reading to learn (LE)	Reading to relate or so as to be able to apply Information later on; refer to both reading to review basic concepts and to reading of a more reflective nature
Reading for cross-referencing (CR)	Cross-referencing documents from a single or multiple sources to integrate information
Reading to edit or critically review text (ED)	Reading to monitor the quality of what is written including one's own text
Reading to support listening (LI)	Reading to support listening to someone else talk
Reading to support discussion (DI)	Reading during a discussion to establish a common frame of reference and focus for the discussion

Writing Activities - Annotations and notes

Creation	Writing to create a new document or to modify an existing one
Note-taking	Writing in an abbreviated manner to serve an temporary purpose other than document creation
Annotations	Marking or writing on top of an existing document to relate the marks or notes with their surrounding context
Form-filling	Filling in structured forms or writing in a pre-specified way
Updating	Updating calendars or schedules

Mayor Digital Libraries

- ***Europeana*** is an internet portal that acts as an interface to millions of books, paintings, films, museum objects and archival records that have been digitised throughout Europe. http://www.europeana.eu
- ***Google Books*** is a service provided by Google that allows users to search the full texts of a large number of books and articles (over 25 million of book titles as of October 2015). The books and magazines were scanned, converted to text using OCR (optical character recognition) and stored in a digital database. https://books.google.com
- ***The Internet Archive*** is a non-profit library that provides free public access to petabytes collections of digitized materials, including web sites, software applications, games, music, movies, videos, images, and nearly three million public-domain books. The overall goal is universal access to all knowledge. https://archive.org/
- ***Open Library*** is an initiative of Internet Archive for facilitating access to digital book. The goal is to include a web database for every book that was ever published. It includes a 23 million catalog records of books. In addition, Open Library has over 6 million digital books in various formats. Several of the books is in the form of scanned raster images. https://openlibrary.org/
- ***HathiTrust*** is an online repository that provides access to a comprehensive body of published works for scholarship and education. The HathiTrust digital archive includes over 13.1 million volumes (4.7 billion pages). https://www.hathitrust.org/

More …

[ABO97] W. Y. Arms, C. Blanchi, E. A. Overly: An Architecture for Information in Digital Libraries. D-Lib Magazine 3(2) (1997)

[AGH+98] A. Adler, A. Gujar, B. L. Harrison, K. O'Hara, A. Sellen: A Diary Study of Work-Related Reading: Design Implications for Digital Reading Devices. CHI 1998: 241-248

[BML15] G. Buchanan, D. McKay, J. Levitt: Where My Books Go: Choice and Place in Digital Reading. JCDL 2015: 17-26

[CS14] L. Colombo, and M. P. Scipioni, M.P. Children reading ebooks on tablets: a study of the context of use. In *Procs. NordiCHI '14*. ACM, New York, NY, USA, 2014.

[Ga08] R. Gartner, Metadata for the digital libraries: state of the art and future directions (1.0). Peer reviewed report from the JISC Technology and Standards Watch. April, 2008, Bristol, UK

[Gr09] T. Gruber: Ontology. Encyclopedia of Database Systems 2009: 1963-1965

[IASA09] IASA Technical Committee, Guidelines on the Production and Preservation of Digital Audio Objects, ed. by Kevin Bradley. Second edition 2009. ISBN 978-91-976192-3-3 l

[Ma97] Catherine C. Marshall: Annotation: From Paper Books to Digital Library. ACM DL 1997: 131-140

[NPW+15] T. Nurmikko-Fuller, K. R. Page, P. Willcox, J. Jett, C. Maden, T. W. Cole, C. Fallaw, M. Senseney, J. S. Downie: Building Complex Research Collections in Digital Libraries: A Survey of Ontology Implications. JCDL 2015: 169-172

[PBT+12] J. Pearson, G. Buchanan, H. Thimbleby, and M. Jones, M. The Digital Reading Desk: A lightweight approach to digital notetaking. Interact. Comput. 24, 5 (September), 2012. 327-338.

[Pe00] H. Petroski., The Book on The Bookshelf, Vintage, 2000.

[TKE+09] C. Tenopir, C., D. King, D. W., Edwards, S. and Wu, L. Electronic journals and changes in scholarly article seeking and reading patterns. Aslib Proceedings, 61, 1, 5-32 2009.

[TLH+11] A. Thayer, C. P. Lee, L. H. Hwang, H. Sales, P. Sen, and N. Dalal, The imposition and superimposition of digital reading technology: the academic potential of ereaders. In CHI 2011 Vancouver, BC, Canada. ACM, 2011. 2917-2926.

[SFP98] B. N. Schilit, A. Golovchinsky, M. N. Price: Beyond Paper: Supporting Active Reading with Free Form Digital Ink Annotations. CHI 1998: 249-256

[Vl10[C. Van Le, "Opening the Doors to Digital Libraries: A Proposal to Exempt Digital Libraries From the Copyright Act," Case Western Reserve Journal of Law, Technology & The Internet, 1.2 (Spring 2010), 135.

Thank you!