Managing Heterogeneous Digital Collections through a Unified Environment

M. Nikolaidou
Library Computer Center, Department of Informatics
University of Athens
Panepistimiopolis, 15771Athens, Greece
Email: mara@di.uoa.gr

Key Words

Digital Library Architecture, Heterogeneous Collection Management, Interoperability.

Abstract

Athens University digital library platform aims at integrating diverse collection requirements to support collection management in a unified manner. Collections vary in terms of the material included and the requirements imposed by potential users. We introduce the term collection management service to denote the automated collection definition and management within an integrated Digital Library framework. Enhanced capabilities, such as simplified collection definition, collection relationship exploration and collection integration, are provided. A digital library framework based on collection management service is also proposed. The implementation of Medical School collections is presented as an example to explore the advantages offered by the proposed framework, while collection definition process is discussed in detail.

Introduction

Athens University initiated a digital collection development project to provide enhanced educational capabilities and preserve the research material produced by its laboratories and researchers. Collections are accessed by students and researchers mainly for educational purposes. Requirements for the integrated Digital Library platform involve: supporting all collections in a unified manner, simplifying the creation of new collections and allowing the integration of existing ones. Promoting interoperability with other library systems (bibliographic catalogue, technical report/thesis collection, library portal) and external applications is also imposed.

Collections vary significantly regarding the material included and the requirements imposed by potential users. Collections may be archival or evolving in nature, include different types of material (e.g. digital material, music, photographs, videos, scanned documents) and are described by different metadata schemes (e.g. DC or EAD variations or even local schemes). Thus, while designing the Digital Library architecture, managing heterogeneous collections was exploited. We introduced the term *collection management service* to denote the automated collection definition and administration within an integrated Digital Library environment. Collections are developed by cataloguers and researchers working in specific university libraries, while the integrated DL environment is supported by the Libraries Computer Center.

The rest of the paper is organised as follows: The concept of an enhanced Collection Management Service is introduced in section 2, while the proposed digital library framework incorporating such a service is presented in section 3. The Collection Dictionary and its proposed structure and functionality is described in section 4. The Medical Digital Collection build based on the proposed framework and the experience obtained is discussed in section 5, while conclusions reside in section 6.

Collection Management Service

In an integrated digital library environment, heterogeneous collections in terms of structure and purpose are supported. Such issues have been exploited in *Greenstone* (Witten, 2001), where the *Collector* environment facilitates collection definition. The Collector environment incorporated in Greenstone facilitates collection structure definition by defining the structure of a collection material and the related metadata based upon a *collection directory*. Aggregating diverse collection-specific requirements and facilitating access to collections through a common access point enables the unified management of all digital material and promotes interoperability (Arms, 2002). The *collection dictionary* concept discussed in (Arms 2002) facilitates access to heterogeneous collections supporting OAI PMH. In this paper, we propose an enhanced *Collection Management Service*, incorporating capabilities similar to those provided by the Collector environment and a few extended services, such as the definition of collections based on existing ones and the handling of different kinds of relationships between collections to simplify collection definition process. The integration of external collections is also supported. A *collection dictionary* is used to maintain collection-related information. Collections are stored in c*ollection repositories,* providing storing and searching services. The proposed collection dictionary in this context offers the following enhanced capabilities:

- Defining collections in terms of structure and related metadata
- Defining relationships between collections (common metadata fields, sub-collection definition)
- Defining collection structure based on existing definitions by properly extending them
- Accessing collections by a common access point
- Integrating collections supported by different implementation environments, independently of digital object storing and searching mechanisms.

Proposed DL Framework

Collection Dictionary implementation is independent of individual collection repository implementations, provided that an access protocol supports search and storing of digital objects. The OAI PMH may be used, while the collection management service may fit within OAI framework as an OAI Service (Suleman, 2002). In the proposed framework (figure 1), collection repositories facilitate storing and searching; while the collection management provides elementary services to add/delete digital material, initiate collection search and forward search results. Results are presented using XML pages containing entity related information and links in the body part and the related metadata information in the header using RDF format. The collection dictionary enables unified access to all collections and the transparent implementation of additional services, such as collection search and cataloguing/processing workflow.

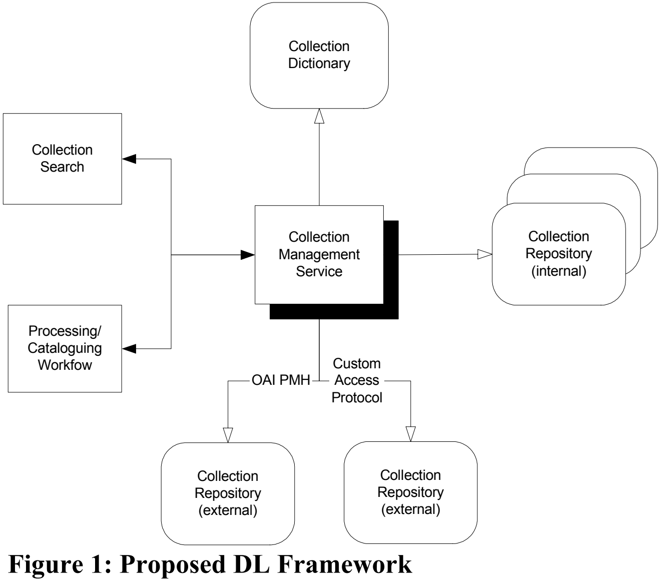
Collection Dictionary
Collection Search
Collection Management Service
Collection Repository (internal)
Processing/ Cataloguing Workfow
OAI PMH
Custom Access Protocol
Collection Repository (external)
Collection Repository (external)

Figure 1: Proposed DL Framework

ollection Management service provides elementary services to create a new collection in a specific repository, add/delete digital material, initiate collection search and forward search results. It is a multithreaded environment implemented using component programming. The service consists of two main modules. *Repository Access* module is responsible for interacting with the repository platform using predefined APIs.

This module is activated upon request whenever there is a need to store or retrieve data or metadata information to/from the repository and its implementation is strongly related with the repository architecture.

Different repository access modules may be included within Collection Management Service to facilitate access to different repository environments. As indicated in figure 1, *internal* and *external* repositories may be defined. Internal repositories are used to create collections and add/retrieve digital material, while external repositories are only used to retrieve metadata and data information regarding external collections. The OAI/PMH is currently supported for external collection search and the relevant OAI Repository Access module is currently under development. Custom access protocols may also be developed to access existing collections, which are not easily migrated within internal repository platforms. As internal repository platforms IBM Content Manager Platform (IBM, 2000) and Greenstone (Witten, 2001) are currently supported. Both of them are chosen since they are already used to develop specific collections, as explained in section 5.

The main services provided by the Collection Manager module are collection management, object processing and cataloguing, user management and collection search. These high-level services are implemented using a set of interacting services organised in a multilayer hierarchy. Each of them interacts with lower-level services further decomposed to lower ones. Only the low-level services, e.g. the ones responsible for storing/retrieving objects or collection definitions initialise the Repository Access module. For example, the *Object Cataloguing* service, a high-level service, interacts with *Object Composer* and *Metadata Manager* services to complete its tasks. The *Object Composer* service, as a low-level service, interacts with Repository Access Module to access the repository platform. System extendibility and transferability is one of the main advantages of the hierarchical architecture, as only the Repository Access Module must be adjusted to repositories built in different platforms. To facilitate unified collection administration the Collection Manager module supports uniformly storing and accessing objects having different structure and being characterised by different metadata. Thus, the same services can be supported for different collections.

To support unified collection management, dynamic interface creation is supported based on information stored within Collection Dictionary. The same access interface is used for all collections, while screens presented to the user are dynamically formed based on the collection description. It is also multilingual, currently supporting Greek and English. Bilingual support increased interface complexity. Different search options are supported for different types of users (e.g. the undergraduate student and the researcher) using user profiles, which are maintained in the Collection Dictionary.

Collection Dictionary

For each collection added in dictionary, the corresponding *collection description* must be defined. We consider that a collection description should comprise three different parts: *collection properties, object structure* and *object metadata*.

The term *digital object* is used to denote material stored within a digital library. Digital objects are usually compound objects consisting of *parts* of different medium type (e.g. text, image, sound, video), which are indexed by different tools. Another object may also be considered as an object part. As an example, we indicate travelling notes in the *Folklore Collection,* which includes material regarding customs and traditions of specific regions of Greece. Travelling notebooks are composed by notes and maps created by the author and also lyrics or handcrafts related with a specific region. The lyrics and handcrafts included in a notebook must be treated both as parts of it and as independent objects belonging in a different collection. The *object structure* is the skeleton used to construct all digital objects belonging in a specific collection (regardless of whether object parts are mandatory), thus it is assumed that all the objects belonging in a specific collection have the same structure.

The digital objects belonging in a specific collection are characterized by a common metadata set (both at object and part level). Metadata may be *general*, common in all collections, and *collection specific*, useful only for a specific collection. For either general or domain specific schemes four categories of metadata are maintain (Besser, 2002): *descriptive*, used to describe the material, *technical* related to object/part type/format and storing properties, *administrative*, used for access control and *educational*, related to educational categorization (e.g. corresponding course or lecture). The metadata scheme used may be a standard one (e.g. DC), a variation of it or even a local one. Implementation properties, e.g. whether a field is bilingual, multi-valued or mandatory, are also included. The value type of each field is also recorded. Restricted value lists are also supported for specific fields. Technical metadata are strongly related to each part medium type (e.g. text, image) and include information regarding object format and storing properties, digitization process (if any) and the original object (if digitized) (Niu, 2002). The technical metadata of an object describe its structure. Administrative and most of technical metadata are not accessed by users. The types of metadata maintained are shown in Table I.

Table 1: Metadata types

Type	Purpose	Level
descriptive	Bibliographic description	object
educational	Educational categorization	object
technical	Technical description	object/part
administrative	Used for access control	object/part

Collection properties consist of collection-based metadata (e.g. EAD header), structural information, relationships between different collections and access information. *Access information* includes the repository where the collection is stored and the access protocol (e.g. OAI PMH). Collection could be *internal*, implemented within the integrated DL platform, or *external*. This feature allowed the integration of existing digital collections, although we had to implement custom access protocols to support specific commercial environments. For example, the University Historical Archive was already digitized using a specific commercial platform with no interoperability features. In order to integrate *Historical Archive Collection* we implemented a custom access protocol. *Relationships* between collections are also defined (e.g. sub-collections). Each sub-collection may be described by different metadata schemes. The relationships between them must be explored. *Collection interoperability rules* may also be defined. This corresponds to the definition of a mapping scheme between the metadata characterizing each collection. A mapping scheme may be partial (as usually concerns a metadata subset) and enables common collection search.

A *collection description* can be derived from an existing description by extending the object structure and metadata model, e.g. a collection description can be defined as the descendant of an existing collection description, while additional object parts and metadata fields can also be defined. This feature allows flexibility during collection definition and facilitates collection description in a simplified manner. The mappings introduced for a specific collection are also valid for its descendants.

Case Study

Using the proposed DL framework, Medical Collections of Athens Medical School were developed. Athens Medical School (AMS) is one of the largest medical research institutions in Greece, where educational activities are combined with every-day practice in the University Hospital. The laboratories operating in the University Hospital produce a large amount of research material, mainly consisting of medical images and videos in digital format. This material should be exploited for educational purposes.

Medical Collections facilitate access to medical material produced by Medical School laboratories for educational purposes. Each Laboratory develops its own collection. Collections vary in terms of the material included and the requirements imposed by potential users. Beside the general metadata describing all medical material, domain-specific metadata related to each laboratory must be maintained. To satisfy the requirements of Medical Collections a complex *processing/cataloguing workflow service* (figure 1) must be supported, since the research material is added directly by the researcher, while he/she also participates in metadata creation.

Researchers usually use the material stored within the Library to create presentation and on-line tutorials for medical students. The presentations are treated as composite medical objects stored within the DL consisting of sequences of images, notes and videos. An additional service was developed (*Create Presentation/Lesson* application) to support collaborative editing of simple presentations consisting of discrete steps of either presenting a medical object stored within the Library or writing simple comments. Although this service was developed for the specific application, it can be used for all other collections as well without any modifications, since it is built on information extracted from Collection Dictionary.

Collection Repository consists of IBM Content Manager components. The IBM Content Manager is a middleware platform providing tools for storing, searching and managing digital content. Sets of APIs are supported to enable access through standard programming interfaces (IBM, 2000). The main servers forming the middleware platform are: *Library Server,* maintaining the data dictionary, and

Object Server, facilitating storing and retrieving digital data as objects. Library Server is responsible for data integrity.

The *IBM Repository Access* module is responsible for interacting with IBM Content Manager Platform using predefined APIs. This module is activated upon request whenever there is a need to store or retrieve data or metadata information to/from the Content Manager, but it cannot facilitate structuring and administrating a digital collection. This functionality is included in the *Collection Manager* module providing services to external clients. Thus, clients do not interact with the Content Manager platform and consequently have no knowledge of its existence. This ensures system modularity and extendibility and enables supporting different data and metadata models at data storing and data management levels. Repository access module is also responsible for DOI assignment and management.

Since all collections include medical material, we have decided to define a generic *Medical Collection* and use it as a prototype to create all laboratory specific collections. Medical collection consists from three sub-collections, *medical image collection*, which includes compound objects consisting of different analysis images (*medical image objects*), *medical video collection*, which include compound objects consisting of different analysis videos (*medical video objects*) and *presentation collection*, which consists of presentations edited by researcher using the material included in the two aforementioned collections (*presentation objects*) (figure 2). For each collection the object structure and metadata model must be defined. The metadata kept for both collections are similar, except the technical ones, which differ. The following parts are included in the *medical image/video objects*:

1. *Original image/video*. It is the original image/video produced in the Laboratory. It is of high quality; it cannot be efficiently transferred over the Web and should be strongly protected regarding copyright issues. Thus, access to it is restricted.
2. *Derivative image/video*: It is produced from the original usually in JPEG/MPEG format to be accessed through the Web. Access to it is restricted.
3. *Thumbnail Image*, to be shown in the Collection Search application.
4. *Description* in Greek and English (as the application should be bilingual).

The original image and the description are produced by the researcher, while all other formats are produced by the cataloguer during image processing.

The metadata scheme introduced to describe the *medical image/video objects* is based on Dublin Core (Dublin Core Initiative), although it also supports customisations for medical material and specific educational metadata, such as *course* or *lecture*. Dublin Core is a widely adopted scheme, used in medical images archives and health care applications (Sakai, 2001; Davenport, 2001). The DC.Subject field was extended to support NML medical subject headings and local thesaurus schemes. The DC.Date field is used to maintain information related with the creation and management of images. The DC.Format field is properly extended to maintain information related with the image file characteristics. The Dublin Core Identifier field is used to store the Medical Image DOI produced automatically. Implementation properties, e.g. whether a field is bilingual, multi-valued or mandatory, are also included in Appendix I. The value type of each field is also recorded. Restricted value lists are supported for specific fields. As indicated in Appendix I, a lot of DC fields and subfields, such as DC.Type, obtained default values.

The *Presentation Objects* consists of two parts: a *description* (in Greek and English) and a multi-valued part, named *presentation pages,* which contains links to objects of the Video and Image Collections. The technical metadata associated with this part contain information regarding its structure. The Dublin core *relation* field (*has_part* type) was used for its description.

The *Medical Collection* is practically empty, while all other collections are easily defined as its descendants by adding *collection-specific metadata fields* and extending the properties of medical image/video objects. *Collection-specific* metadata scheme can be defined using Dublin Core basic fields or extensions or even local fields.

As an example, we discuss the definition of the *Histological Collection*. As indicated in figure 2, the *Description* of *Histological Collection*, corresponding to the Laboratory of Histology, is derived from *Medical Collection Description.* There are no alterations in the *Presentation Collection* Description. Since the laboratory produce digital images, only the *Medical Video sub-collection* remains idle. The *Medical Image Collection* description was extended. The *Medical Image Object* *structure* was extended by adding two new parts: *Watermarked Image*, produced from the derivative image watermarked using the symbols of the University and the corresponding Laboratory, and *Screen Size Image*, a medium-quality image produced from the derivative image to be easy shown through the Web. C*ollection-specific metadata fields* are added in the descriptive metadata. These fields are considered as *local* (they are not Dublin Core fields), since they are useful only when searching the specific collection. Metadata description is accompanied by implementation properties, such as indicators of whether the field is bilingual or not, multi-valued or not, mandatory or not, and field type. While defining the Histological Collection, only additional features have to be described contributing to the simplification of collection definition process.

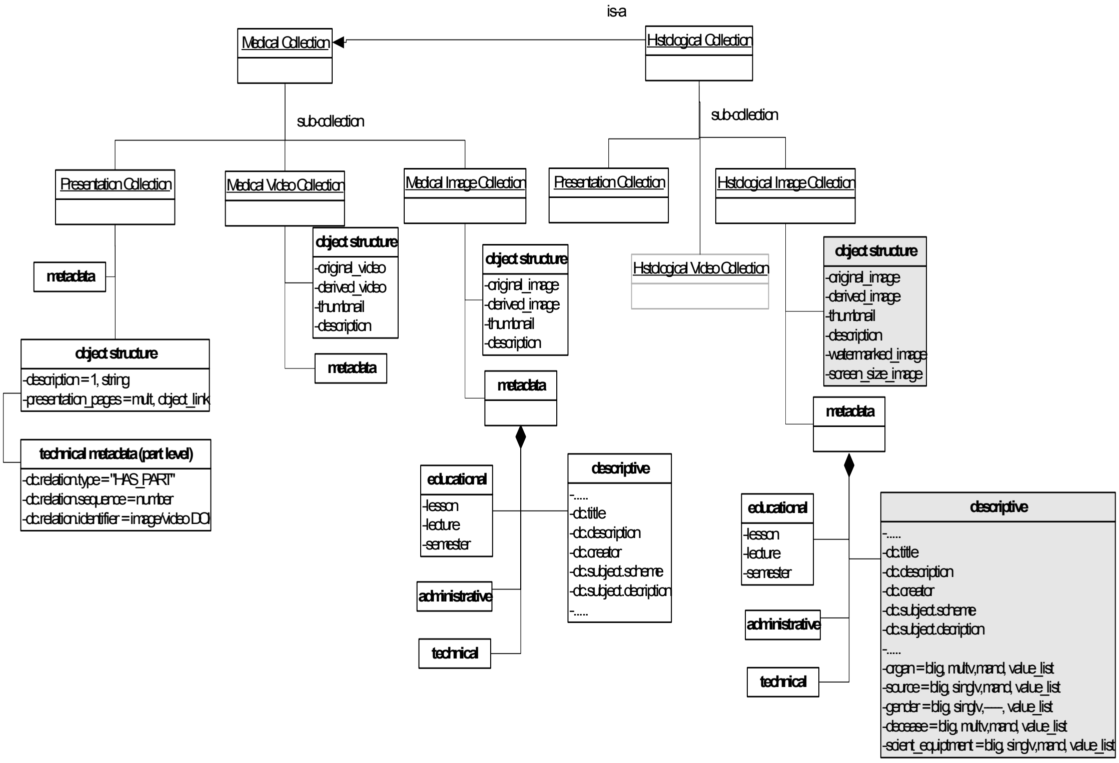

Figure 2: Medical and Histological Collection Definition

One of the problems we faced during system employment was to integrate Medical Object creation in researcher's daily work. Selection of material takes place while the researcher is examining properly processed tissue samples with specialised microscopes capable of producing digital material. The digital material can be stored in a hard disk but, in order to add it in the library, the researcher must leave his position in the microscope and enter the *Processing/Cataloguing* workflow application operating in the workstation next to it. This was not feasible due to researcher' workload. Thus, it was decided to store images during tissue examination, while review and characterisation of images are performed by researchers on a weekly basis.

Conclusions

We introduced the concept of enhanced collection management service to denote automated collection management. In order to support collection management service, a collection dictionary was implemented enabling collection definition in an advanced fashion. The Collection Management Service responsible for managing the Collection Dictionary, also facilitates access to collections through a common access point enables the unified management of all digital material and promotes interoperability, as the provided services, as processing and cataloguing workflow, are implemented independently of the Repository implementation platform.

Collections are described in detail regarding their structure, supported metadata and the relationships between them. Derived collection definition enabled the description of all collections by extending the description of a simple one (medical collection) that provided basic capabilities. It contributed significantly to the simplification of the overall process, as metadata field definition is rather time-consuming, especially when dealing with fields with bilingual and predefined values. The definition and administration of composite objects as those belonging it the Medical Presentation Collection proved to be straightforward and efficient.

Processing and cataloguing workflow application as well as collection search support dynamic interface creation, as screens presented to the user are dynamically formed based on the collection description. Dynamic interface creation promotes application flexibility, while it did not affected performance (seven different medical collections are currently supported). The user interface is simple and user friendly. Both researchers and cataloguers were able to use it after a few hours of training.

References

Arms, W.Y., et. al. (2002) **A Spectrum in Interoperability**, *D-Lib Magazine*, Vol. 8, No 1, available http://www.dlib.org.dlib/january02/01arms.html.

Besser, H. (2002) **The Next Stage: Moving from Isolated Collections to Interoperable Digital Libraries**, *First Monday On-line Journal*, Vol 7, No 6, available http://firstmoday.org/issues/issue7_6/besser/index.html.

Davenport Robertson, W., Leadem, E.M., Dube, J., Greenberg, J. (2001) **Design and Implementation of the National Institute of Environmental Health Sciences Dublin Core Metadata Schema**, in *Proceedings of Int. Conf. On Dublin Core and Metadata Applications 2001*, National Institute of Informatics, pp. 193-199.

Dublin Core Metadata Initiative, available (June 2002) http://www.dublincore.org/.

IBM Corporation (2000), **Content Manager Documentation – Planning and Installation Guide**.

Niu, J. (2002) **A Metadata Framework Developed at the Tsinghua University Library to Aid in the Preservation of Digital Resources**, *D-Lib Magazine*, Vol 8, No 11, available http://www.dlib.org.dlib/november02/ 01niu.html.

Sakai, Y.(2001) **Metadata for Evidence Based Medicine resources**, in *Proceedings of Int. Conf. On Dublin Core and Metadata Applications 2001*, National Institute of Informatics, pp. 81-85.

Suleman, H, Fox, E.A. (2002) **Designing Protocols in Support of DL Componentization**, in *Proceedings of ECDL'02*, Springer Verlag LNCS 2458, 568-582.

Witten, I. H., Bainbridge, D. and Boddie, S.F. (2001) **Greenstone: Open-source digital library software with end-user collection building**, *On-line Information Review*, Vol. 25, No 5. pp 288-298.