A multi-layer metadata schema for digital folklore collections

Irene Lourdi

Libraries Computer Centre, National and Kapodistrian University of Athens & Laboratory on Digital Libraries and Electronic Publishing, Dept. of Archive & Libraries Sciences, Ionian University, Greece

Christos Papatheodorou

Laboratory on Digital Libraries and Electronic Publishing, Dept. of Archive & Libraries Sciences, Ionian University, Greece

Mara Nikolaidou

Harokopio University of Athens & Libraries Computer Centre, National and Kapodistrian University of Athens, Greece

Correspondence to: Irene Lourdi Libraries Computer Centre, National and Kapodistrian University of Athens, University Campus, Dept. of Informatics & Telecommunications, Ilisia, 15784, Athens, Hellas. Tel+302107275618. Email: elourdi@lib.uoa.gr

Abstract

Digital folklore collections are valuable sources for studying cultural and oral tradition of a country. The main difficulty in managing such collections is material heterogeneity (handwritten texts, photographs, 3D objects, sound recordings etc.) that imposes different digitization, description and maintenance practices. A multi-layer metadata model for the description of a digital folklore collection is presented. The proposed metadata policy considers a collection as a hierarchy of entities and combines different metadata schemas for the management of each entity. The metadata model integrates elements from different metadata schemas ensuring efficient information discovery from all structural levels. Furthermore, interoperability between the used metadata schemas is discussed and Topic Maps model is presented as an approach for developing mappings.

Keywords: Folklore collections, metadata policies, application profiles, topic maps

JIS Editorial Process
Received: 26th February 2006
Revised: 25th June 2006
JIS-0332 v2.0

1. Introduction

Digitizing cultural material has become a great priority for heritage institutions. A digital collection facilitates user access to the wealth of cultural heritage objects through various retrieval policies [1]. A large volume of papers focuses on digitization projects and strategies in heritage institutions [2][3][4], but in most cases, little attention is given to the analysis of documentation workflow.

This study is motivated by the digitization project of the folklore collection belonging to the Department of Greek Literature, University of Athens in Greece. The collection is dedicated to the tradition and customs of almost all the Greek regions. It is quite complex and lacks systematic documentation. Additionally, the material diversity imposes difficulties in digitization project, since the adopted metadata model must facilitate the effective description of both the structure and semantics of cultural objects.

Our purpose is to propose an efficient method to manage and expose the wealth of complex cultural heritage collections. We present a functional metadata policy covering both collection-level and item-level descriptions, as well as facilitating effective access to digital content. Furthermore, we discuss the implementation of the digital folklore collection in the integrated digital repository platform of University of Athens and related experience. Also we deal with metadata interoperability issues since a variety of metadata standards are used to describe heterogeneous material collections, and therefore search mechanisms for integrating the participating metadata schemas are needed.

The rest of the paper is organized as follows: in section 2 we define the problem and discuss about the benefits of our effort referring also to related work. In section 3 the main digital collection requirements are demonstrated and in section 4 the collection structure and the metadata model are presented. Further, a digital repository management architecture implementing the proposed model is presented in section 5 and a mapping solution for interoperability reasons is proposed in section 6. Finally, conclusions reside in the last section discussing also the contributions of our work.

2. Problem definition and related work

The folklore collection of the Greek Literature Department of the University of Athens mainly consists of hand-written travelling notebooks containing information about the way of living in various regions of Greece. The notebooks were written by the Department's students, since about 1967, and accumulated in its library. Each notebook is related to a specific region. The notebooks structure is based on a questionnaire prepared by folklore experts and organized into predefined chapters and sections, which are indicated in the table of contents. Further, many notebooks contain attached maps created by the author, photographs, lyrics and handcrafts collected also by the author and related with the specific region. Notebooks are a significant source for a folklore researcher since they contain primordial information collected from interviewed habitants of each visited place. They consist of a quite large collection, containing more than 4000 notebooks and 350000 pages.

For administrative reasons the folklore collection is divided into sub-collections according to the type of objects. The sub-collections are:

1. Notebooks sub-collection. Each notebook is a manuscript written by a student after local research and refers to a specific area or village. Most of the notebooks are accompanied by *photographs* of habitants and places and *small objects* stuck on the pages like artefacts (e.g. laces or doles). These attached items can also be a part of the further separate sub-collections:
2. Photographs sub-collection, accompanying the notebooks. The exact number is yet unknown.
3. Objects sub-collection, consists of objects that are either attached to the notebooks (unknown number) or exposed in the library, which are encountered to be about 1000.

Despite the cultural value of material, the collection has not been currently catalogued or registered to an electronic system. Therefore (a) users are obliged to read and look through all the notebooks (thousands of pages) to find any information (b) the physical material and consequently collection’ s intellectual content is exposed constantly to bad time effects and to unpredictable natural destructions. Thus, the expected benefits from the digitization project will be to provide users with access mechanisms to folklore resources by effectively organizing the material and contribute to the long-term preservation of the valuable cultural information. Moreover, digitization is expected to increase both the *distant* visitors, who view the collection over the Internet and retrieve information using specific access points such as subject, date, place etc., as well as *in-house* visitors, who mainly are folklore researchers that want to study closely the material and contribute to its documentation. Our approach focuses on two dimensions:

(a) the management, promotion and exploitation of the cultural information provided by each collection item and

(b) the multi-level description of the collection structure emerging the semantics and the relations of the items [5].

To the best of our knowledge there are similar projects dealing with composite collections. The “Discover Project” [6], manages the New Zealand’s National Library digital collection, which consists of various types of material (photographs, books, music etc). The metadata schema applied for the items in Discover contains elements from Dublin Core element set (DC) [7], qualified Dublin Core [8], Encoded Archival Description (EAD) [9] and also local defined elements. In the same vein a metadata application profile for collection-level description is proposed in [10] focusing on complex folklore collections. Another similar case is the “European Library Project” (TEL) [11], which is funded by European Commission. The main objective of the project is to set the right framework leading to a system for accessing various collections of the European National libraries. The TEL metadata model is proposed to be an application profile based on DC-Library Application profile covering the needs of collections and materials owned by the European National libraries. The project presents an interesting approach for metadata development and heterogeneous collections manipulation, since (a) it distinguishes the collection-level and item-level descriptions and (b) the proposed metadata models combine elements from various metadata standards.

An attempt for describing and organizing electronic documents and collections on the Web is the MODDEC metadata model [12]. In this case an extension of the work developed by Barreto [13] is proposed taking into account six important aspects: structure, intellectual content, relationships, internal and external organization and presentation formats. The main idea of this approach is that digital collections and objects are organized into a hierarchical structure and each hierarchy level is described by specific metadata. This consist a first attempt of organizing resources using a metadata framework that explores the associations between resources and their structural composition.

3. Requirements

3.1. Content organization

Folklore collections are usually characterized by material heterogeneity, since they contain objects such as handcraft objects, clothing, written texts, digital like music records, photographs, video recordings and other kinds of material coming from the daily activities of people. The volume and variety of the resources demands the organization of a cultural heritage collection into sub-collections following specific criteria like the type of objects, the corresponding chronological period, the geographic region, the common provenance or even the common usage of them [14].

Consequently, in order to manipulate and characterize a hybrid collection, it is required to divide the collection into sub-collections and represent the resulted internal structure. By defining a hierarchy in the collection, the characteristics inherited from the collection to sub-collections and to items, such as date, subject, geographic area etc., can be identified by setting particular access points. Furthermore, hierarchy is helpful for the digital collection administrators, since every structural level can be considered as a distinct entity and can be accredited with rich semantics. However, the internal collection structure encoding requires the application of structural metadata to describe the logical or physical relationships between the parts of the collection and of each compound object.

3.2. Metadata policy

The heterogeneity of folklore resources requires a metadata model that will combine elements from various metadata standards. The adopted metadata model should cover the following categories: (a) the content nature and characteristics of items (descriptive metadata) (b) the digitization technique and the technical requirements (technical metadata) (c) the meta-metadata information indicating the particular metadata standards used for the description of each material type (d) the access rights and the copyright of folklore material (rights metadata) and (e) the educational character or the purpose of every resource. Since the folklore objects have not been created in isolation but are parts of a collection, they should be described within the collection context [15]. Hence, collection-level and object-level documentation must be combined.

Further, it is important to deploy a metadata policy that will preserve the inheritance among entities hierarchy, avoid redundancy of information and link descriptions [16]. The desired policy will ensure both the efficient information retrieval as well as the protection of the original content by keeping all the valuable information for authenticity and preservation. Users must have the potential to discriminate the specific collection from the plethora of other similar in the web and decide whether it is of their interest. Finally, the metadata policy must be compatible with international protocols and standards to accomplish a high level of consistency, credibility and interoperability.

3.3. Access rights

The unique nature of cultural resources requires access policy to protect the authenticity of information without preventing resources retrieval. A digital folklore collection shall allow material usage in such a way that does not breach intellectual rights of cultural objects and simultaneously respect every user needs for learning and studying. There should be the possibility to refine specific user groups or even to provide different access rights for each level of the collection structure. There are no specific access restrictions for the folklore collection of the University of Athens, but the target to increase material usage and exposure should not happen against data safety and long-term preservation. Thus, digital folklore collection must conform to the policy of the holding institution and provide access to material up to the degree permitted.

3.4. Functionality - Usability

Digital collection functionality can be evaluated examining various parameters such as retrieval performance, usability, etc. The main target of the identification and digitization of folklore collections is to facilitate wide access to containing items and promote education, academic research and preservation of cultural heritage and folklore features. In other words, digital folklore objects are required to be transformed into valuable resources used as a reference from other related works. Users need to study old and fragile objects without concerning about possible destructions of physical material.

Hence, digital folklore collections should bring together various user groups of the scientific community and enable resource discovery and foster item-level access [17]. Information about cultural heritage and oral tradition of a country is interesting for a wide audience of varied educational level and preferences (students, historians, philologists, psychologists, ethnologists etc). Further by developing detailed content descriptions and usable digital services, efficient search across aggregations of varied and complex sub-collections and objects is accomplished in a robust, rich and user-friendly manner.

4. Information Architecture

4.1. Collection structure

A digital collection consists of digital objects that, in an abstract level, refer to real-world objects [18] usually organised in a hierarchical structure (collection, sub-collections and items). Figure 1 presents the structure of the folklore collection. The notebooks sub-collection is divided to the levels of chapter, subsection, page, text/objects/photographs. Each level contains a set of digital objects with their own attributes, description and behaviours. The photographs and handcrafts included or attached in any notebook can be documented separately as independent entities with the possibility to affiliate them either in the context of a notebook or in a different sub-collection.

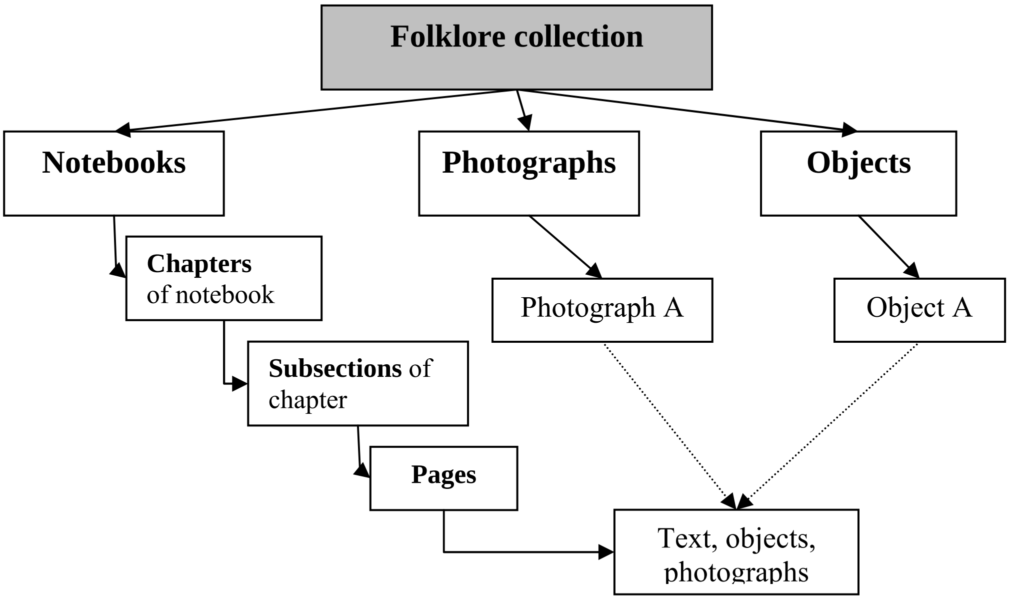


Fig.1 Description levels of folklore collection

It should be noticed that any other categorization of material would cause additional processing and classification effort. For example, if thematic categorization was preferred, then all wedding ceremony customs should consist a sub-collection. In such a case, librarians should re-organise the whole collection and therefore they spend extra time for that work. Moreover, setting particular access points supporting subject-based queries ensures thematic access to cultural content.

4.2. Metadata schema

The proposed metadata schema describes each collection level and presents the relationships between the corresponding digital objects. Also it is expected to enhance information discovery by offering users automatic facilities either by (a) browsing the notebooks one by one or (b) searching their content using keywords or combinations of several search criteria like time, place, usage etc. The proposed metadata model correlates all collection levels and defines the adequate elements for a rich documentation making simultaneously easier the cataloguer's job, who is obliged to fill the specific metadata fields for each level.

The metadata schema includes three kinds of data: *descriptive* (describe the resources intellectual content), *structural* (document the structure of the objects and the relationships between them) and *administrative* (provide information about the digitization process and the collection preservation) [19]. The following challenging issues have also been taken into account:

- Provide full documentation about the content and context of objects e.g. when they were created, by whom, their usage, their contents etc.
- Facilitate queries and information retrieval from all the structural levels of the collection.
- Combine different metadata standards in order to cover all the formats and types of physical and digital objects.
- Assure the greatest interoperability with other projects and applications.

A picture of the metadata categories that are ascribed to every hierarchical entity of the folklore collection is presented in table 1.

Table 1 Metadata categories

COLLECTION	Descriptive	
	Administrative for the physical	for the digital collection
	Structural	
NOTEBOOK	Descriptive	
	Administrative for the physical	for digital version of the notebook
	Structural	
CHAPTER	Descriptive	
	Administrative	
	Structural	
SUB-SECTION	Descriptive	
	Structural	
PAGE	Administrative	
	Structural	
PHOTOGRAPH	Descriptive	
	Administrative for physical	for digital versions of the photo
	Structural	
OBJECTS	Descriptive	
	Administrative	
	Structural	

The metadata model for the digital folklore collection is thoroughly presented in the following paragraphs. It is based on DC for both collection-level and item-level description while enriched with elements from other metadata standards or with local fields according to material requirements. The elements are grouped in categories, as shown in table 1, according to the described object nature (physical or digital). It is important to maintain information for the physical and digital version of the object, because physical characteristics affect also digital ones.

In the tables with the metadata fields for each component, some indications are given to accredit specific attributes to each element. These indications show: a) the metadata standard from which each element origins (e.g **DC**=Dublin Core, **L**=local according to the project requirements etc.) b) whether an element is mandatory (**M**=mandatory) and c) which elements are proposed to be filled automatically by the system taking values from lower or upper levels (**I**=inherit); this ensures the inheritance policy from one level to another.

4.3. Collection-level description

Collection-level description is based on an application profile, which has been proposed for describing a folklore collection as a distinct entity [13]. The application profile considers as core schema the Dublin Core Collection Description Application Profile (DC CD AP) [20] and extends it with elements from metadata standards that cover special characteristics of the collection entity. These standards are: General International Standard Archival Description (ISAD(G)) [21], the metadata model of Alexandria Digital Library (ADL) [22], Research Support Libraries Program (RSLP) [23] and IEEE-Learning Object Metadata (LOM) [24]. In table 2 abbreviations of them are encapsulated in the label of each element.

In particular the refined term of DC *table of contents* describes the contents of the collection. The element *relation*, with its additional refined terms, encodes the types of relations that are not covered by the elements of sub-collection, super-collection and associated collection given from DC CD AP. Also the element *dc_source* is applied to indicate the source of the collection and the element *dc_contributor* to express other persons related with the collection besides collector and owner.

Furthermore, elements from ISAD(G) were added such as: *legal status* of the collection and *note* to encode any other kind of information for the collection that does not fit in another field, e.g. information about the digitization project. The element *structure* from LOM defines the type of collection structure (e.g. hierarchical, linear or networked) and the element *location* from RSLP specifies the place of both the physical and digital collection.

However, the most important elements are *metadata schema & metadata mapping* taken from ADL schema, that encode the metadata standards used for the description of collection objects, and *scope/purpose* that specifies the purpose of the collection or how it will be used. This kind of information is valuable for the description of folklore collections because their heterogeneity requires mixing and combining separate metadata standards to describe each item according to its format.

The elements that are inherited automatically from the collection to the notebook schema are *physical location*, *language*, *rights & access rights*, while the values of elements *medium* and *size* are computed as summations of the corresponding elements in the notebooks schema.

Table 2 Collection Entity Metadata

COLLECTION SCHEMA		
DESCRIPTIVE METADATA		
(DC CD AP)_TITLE (M)	(DC CD AP)_COLLECTOR	(DC CD AP)_AUDIENCE
(DC CD AP)_ALTERNATIVE TITLE	(DC CD AP)_LANGUAGE	(RLSP)_ACCRUAL STATUS
(DC CD AP)_SUBJECT	(DC CD AP)_DESCRIPTION	(ISAD)_NOTE
(DC CD AP)_ACCUMULATION DATE RANGE	(DC CD AP)_COVERAGE SPATIAL	(ADL)_SCOPE/PURPOSE
(DC CD AP)_CUSTODIAL HISTORY	(DC CD AP)_COVERAGE TEMPORAL	(DC)_SOURCE
ADMINISTRATIVE METADATA for PHYSICAL COLLECTION		
(RSLP)_LOCATION_PHYSICAL	(DC CD AP)_OWNER	(DC CD AP)_ACCESS RIGHTS
(DC CD AP)_IDENTIFIER (M)	(DC CD AP)_TYPE	(ISAD)_ LEGAL STATUS
(DC CD AP)_SIZE (I)	(DC CD AP)_RIGHTS	(DC)_CONTRIBUTOR
ADMINISTRATIVE METADATA FOR DIGITAL COLLECTION		
(RSLP)_LOCATION_DIGITAL	(DC CD AP)_ACCESS RIGHTS	(ADL)_METADATA SCHEMA
(DC CD AP)_MEDIUM (I)	(DC CD AP)_RIGHTS	(ADL)_METADATA MAPPING
(DC CD AP)_SIZE		
STRUCTURAL METADATA		
(LOM)_STRUCTURE	(DC CD AP)_ASSOCIATED COLLECTIO	(DCTERMS)_RELATIONS
(DC CD AP)_SUB-COLLECTION	(DC CD AP)_SUPER-COLLECTION	(DCTERMS)_TABLE OF CONTENTS

4.4. Notebooks-level description

The description of the notebooks and their structural levels has been based mostly on DC. Moreover elements have been added from the bibliographic standard MARC [25] to describe characteristics of physical objects and MIX [26] to give technical information about the scanning process of notebooks, images and small objects placed inside the pages. The metadata fields used for notebook description are presented in table 3.

The element *dc_format_extent* holds the number of scanned pages calculating them automatically. The element *dc_description_tableofcontents* is automatically filled copying the values of the element *dc_title* of the related chapters. The elements *dc_coverage_spatial* and *dc_date_accumulated*, are inherited to the lower level of chapters. Also the element *dc_date_accumulated* is another label for the refined term *dc_date_created* and expresses the time period that was needed for the student to collect the information and write the notebook. The local element *other physical details* describes the physical characteristics of the notebook that maybe are not covered from the rest elements, such as whether a page of the notebook is missing. The local defined element *credibility* is added after specific request of the folklore experts because the notebooks contain primordial information that has not been studied before. The MARC element *binding information* is added to keep information about the way the notebooks have been compiled in a volume and according to which criteria.

Table 3 Notebook Entity Metadata

NOTEBOOK SCHEMA		
DESCRIPTIVE METADATA		
DC_ TITLE (M)	DC_DATE_ACCUMULATED	COVERAGE_SPATIAL_SPECIFICATION (L)
DC_SUBTITLE	DC_COVERAGE SPATIAL (M)	COVERAGE_SPATIAL_ADDITIONAL INFO (L)
DC_CREATOR (M)	CREDIBILITY (L)	DC_DESCRIPTION_TABLEOFCONTENTS
DC_CONTRIBUTOR (ROLE)	DC_SUBJECT	SUBJECT_CLASSIFICATION (L)
ADMINISTRATIVE METADATA for Physical entity		
BINDING INFORMATION (MARC)	DC FORMAT_EXTENT (I)	FORMAT_DIMENSIONS (MARC)
DC_IDENTIFIER (M)	DC_SOURCE	DC_DATE_ACCUMULATED (M)
ADMINISTRATIVE METADATA for Digital entity		
DC_DATE_CREATED (M)	OTHER PHYSICAL DETAILS(L)	DC_FORMAT_ EXTENT(I)
DC_DATE AVAILABLE	LOCATION_DIGITAL (L)	DC_FORMAT_MEDIUM
STRUCTURAL METADATA		
DC_RELATION (IS PART OF)	DC_DESCRIPTION_TABLEOFCONTENTS (I)	

The chapter description is shown in table 4. Accordingly, the element *dc_description_tableofcontents* is automatically filled by aggregating the values of the element dc_title of the corresponding subsections. Also the refined term *dc_format_extent* is automatically filled calculating the number of the scanned pages that belong to the specific chapter in the case of the original notebook.

Table 4 Chapter Entity metadata

CHAPTER SCHEMA		
DESCRIPTIVE METADATA	ADMINISTRATIVE METADATA	STRUCTURAL METADATA
DC_TITLE (M)	DC_FORMAT_EXTENT (I) (FOR PHYSICAL)	DC_DESCRIPTION_TABLEOFCONTENTS (I)
DC_COVERAGE_SPATIAL	DC_IDENTIFIER (M)	DC_RELATION_(IS A CHAPTER OF…)
DC_DATE_ACCUMULATED		

The chapters of a notebook are divided further in “subsections” predefined according to the questionnaire prepared by folklore experts. As shown in table 5, the element *dc_description_tableofcontents* is filled mechanically from the titles/ names of the objects and photographs, if they have separate metadata records.

The element *dc_contributor* describes the person that gave the information of the chapter to the student. The element *dc_subject* is mandatory to be filled in this level, since the values of the element are copied automatically to the upper corresponding element *dc_subject* of notebook schema.

Table 5 Subsection Entity metadata

SUBSECTION SCHEMA	
DESCRIPTIVE METADATA	
DC_IDENTIFIER (M)	DC_SUBJECT (M)
DC_TITLE (M)	DC_CONTRIBUTOR
DC_DESCRIPTION_ABSTRACT	
STRUCTURAL METADATA	
DC_RELATION (IS SUBSECTION OF CHAPTER)	DC_DESCRIPTION_TABLE_OF_CONTENTS (I)
DC_RELATION_HAS “PHOTOGRAPH”/ “OBJECT”	

The elements *pixel size, scanning resolution* etc. describe technical details about the scanning process of each page and the element *other physical details* maintains characteristics of the specific page that need to be preserved. This technical information is encoded by elements from MIX schema [26], as it is shown in table 6.

Table 6 Page Entity metadata

PAGE SCHEMA		
DC_IDENTIFIER (M)	OTHER PHYSICAL DETAILS (MIX)	PHYSSCANRESOLUTION (MIX)
PIXEL SIZE (MIX)	DC_DATE CREATED (M)	DC_RELATION (IS PAGE OF THE SUBSECTION…)
FILE SIZE (MIX)		

Photographs and small objects related to notebooks are documented separately depending on the choice to affiliate them either in the context of the digital notebook or as independent entities in a different sub-collection (table 7).

Table 7 Photograph Entity metadata

PHOTOGRAPH SCHEMA		
DESCRIPTIVE METADATA		
DC_TITLE (M)	*DC_SUBJECT*	COLOR (MARC)
DC_DESCRIPTION_ABSTRACT	*DC_COVERAGE_SPATIAL*	NOTES (MARC)
DC_DATE CREATED	DC_COVERAGE_TEMPORAL	
ADMINISTRATIVE METADATA for physical		
DC_FORMAT_DIMENSIONS	TECHNIQUE (MIX)	DC_TYPE (M)
OTHER PHYSICAL DETAILS (MARC)		
ADMINISTRATIVE METADATA for digital		
SCANNER MODEL NAME AND NUMBER (MIX)	BIT DEPT (MIX)	DC_RIGHTS
SCANNINGSOFTWARE (MIX)	COMPRESSION LEVEL (MIX)	DC_FORMAT_MEDIUM
PHYSSCANRESOLUTION (MIX)	FILE SIZE (MIX)	DC_IDENTIFIER (M)
PIXEL SIZE (MIX)	DC_DATE CREATED (M)	
STRUCTURAL METADATA		
DC_RELATION (is referenced to subsection…)		

The element *dc_coverage_spatial* is copied automatically from the notebook to photograph schema and the element *dc_subject* is copied from the subsection schema that the photograph belongs to. If the photograph subjects do not agree with the subsection subjects, then are filled manually. Technical details regarding photograph scanning may differ in every case, due to the special features of each photograph. Therefore, it is wise to keep information about the digitization of the photographs separately for each one for preservation reasons. Like in scanned pages, MIX schema is also applied. The MARC element notes is used in case we want to give general additional information that does not fit in another element, for example how the student found the photo.

The elements used for describing physical objects are given in table 8. Dc_coverage_spatial and dc_subject are inherited from the upper levels and like in photographs, the same condition exist also for the subjects.

Table 8 Object Entity metadata

OBJECT SCHEMA		
DESCRIPTIVE METADATA		
DC_TITLE (M)	DC_SOURCE	*DC_SUBJECT*
DC_DESCRIPTION	*DC_COVERAGE_SPATIAL*	DC_COVERAGE_TEMPORAL
DC_DATE CREATED		
ADMINISTRATIVE METADATA for physical		
DC_IDENTIFIER (M)	BRIGHTNESS (MIX)	ADDITIONAL PHYSICAL CHARACTERISTICS (MIX)
DC_RIGHTS	DC_DATE AVAILABLE	DC_DATE CREATED (M)
DC_TYPE	FLASH (MIX)	DIGITAL CAMERA MODEL (MIX)
DC_FORMAT_MEDIUM (MATERIAL)	BLACK LIGHT (MIX)	EXPOSURE TIME (DURING SCANNING) (MIX)
DC_FORMAT_EXTENT	FOCAL LENGTH (MIX)	
STRUCTURAL METADATA		
DC_RELATION (is referenced to subsection…)		

The proposed metadata model satisfies the requirements discussed in section 3, such as material organization, collection and item description, rights preservation and enhanced resource discovery. The main benefit of the proposed policy is that every entity of the collection is treated as a separate digital object with the adequate metadata, providing plethora of access points for information discovery. The system filters thousands of pages transparently to users following particular paths to find the desired information.

As an example consider the following query: a user wants to find information about the wedding ceremony songs in the South part of Greece. In order to satisfy this query, searches to various levels can be combined, such as:

- which are the notebooks about this geographic region (notebook-level: spatial_coverage)
- do they have a chapter about marriage (notebook-level: table of contents)
- which is the right chapter (chapter-level, title) and which subsection (subsection-level: title and subject).

The main drawback of the proposed metadata schema is its complexity, since it requires much time for cataloguing and filling all elements. This can be encountered against the usability of the schema but the unquestionable value of preserving and providing cultural information for a society makes this problem seem less important. However, in order to facilitate cataloguers in their daily work values of specific elements are copied to higher or lower level corresponding elements.

5. Implementation and evaluation

5.1. Implementation

Pergamos digital repository system [27] has been developed by the Libraries Computer Centre of the University of Athens in Greece to host all digital collections of the University. Pergamos is based on Fedora, a Java based open-source flexible and extensible digital repository management system. It can implement the proposed metadata policy, since it supports many metadata models, which can be either local or extensions of DC set.

Within Pergamos, each digital object is consisted of four parts: a) the metadata sets that describe the content of the object, b) the structure, that indicates its associations and any linking information i.e. the existing links among the objects of a collection, c) the files of the digital object i.e. the digital instances of the physical object, if it has any (for example a photograph may have various instances like a TIFF image, JPEG or thumbnail) and d) the behaviors, a set of methods that define how the object can be manipulated in the system.

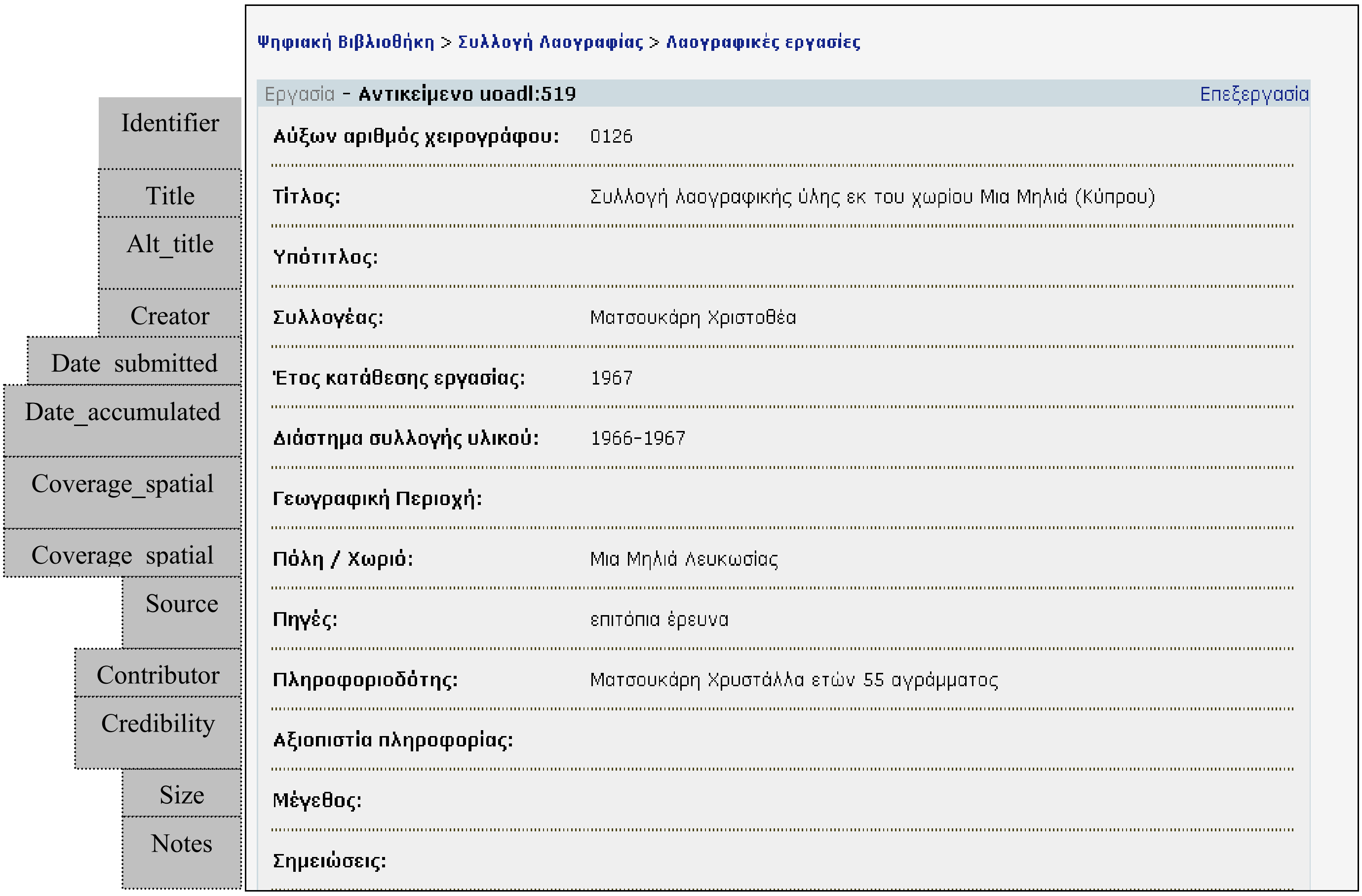
```
<collection
 id="folklore.notebooks">
        ....
        <structure>
          <child type="notebook"></child>
        </structure>
        <DOTypes><DOType id="notebook"> <label lang="en">Notebook</label>
            <MDSets><MDSet id="dc"> <label lang="en">Dublin Core Metadata</label>
        <fields>
            <field id="dc:identifier_physical" indexed="true" mandatory="true" repeatable="false" viewable="true">
            <label lang="en">Call number</label> <description lang="en">An unambiguous reference to the resource
within a given context</description></field>
            <field id="dc:title" textarea="true"></field>
             <field id="dc:title_alternative" repeatable="false" textarea="true"><label lang="en">Subtitle</label><label
lang="en">Υπότιτλος</label></field>
             <field id="dc:date" indexed="true" repeatable="false" viewable="true"><label lang="en">Year
submitted</label>
                <description lang="en">Year of submission of the notebook</description></field>
             <field id="dc:date_accumulated"><label lang="en">Accumulated range</label><description
lang="en">Range of material accumulation</description></field>
             <field id="dc:creator"><label lang="en">Creator</label> <description lang="en">An entity primarily
responsible for making the content of the resource</description></field>
             <field id="dc:contributor" indexed="true" mandatory="false" repeatable="true" textarea="false"
viewable="true">
                   <label lang="en">Interviewee</label> <description lang="en">A person that gave information to the
creator of the notebook</description> </field>
                         .........
```

Fig. 2 XML file representing the metadata of notebook prototype

Pergamos proposes a mechanism for digital objects generation and manipulation based on the *prototype instantiation process*. According to [28], a *digital object prototype* specifies the digital objects' constitutional parts (metadata sets, files, internal structure and behaviors definitions). Each digital object is an instance of a digital object prototype and the process of generating a digital object is called prototype instantiation. This process ensures that the resulted objects will be aligned to the specifications of the prototype. Hence, all information in Pergamos is stored in terms of digital objects prototypes and their instances.

According to the digital folklore collection, the proposed architecture is implemented as a set of prototypes, each one describing a different hierarchy level. So each collection and sub-collection is an instance of the collection prototype, while the notebook prototype is used for the implementation of notebook objects (notebook level) etc. Further, all digital objects are stored in Pergamos as XML documents. An indicative part of the XML schema for the notebook prototype is presented in figure 2. Specific attributes are defined for each element such as whether it is mandatory, repeatable or indexed.

The XML files of digital object prototypes are presented to the users by web-based and quite friendly templates. The fields and attributes described in notebook prototype of figure 2 are included in the notebook cataloguing form presented in figure 3.



Ψηφιακή Βιβλιοθήκη > Συλλογή Λαογραφίας > Λαογραφικές εργασίες

Εργασια - **Αντικείμενο uoadl:519** Επεξεργασία

Αύξων αριθμός χειρογράφου:	0126
Τίτλος:	Συλλογή λαογραφικής ύλης εκ του χωρίου Μια Μηλιά (Κύπρου)
Υπότιτλος:	
Συλλογέας:	Ματσουκάρη Χριστοθέα
Έτος κατάθεσης εργασίας:	1967
Διάστημα συλλογής υλικού:	1966-1967
Γεωγραφική Περιοχή:	
Πόλη / Χωριό:	Μια Μηλιά Λευκωσίας
Πηγές:	επιτόπια έρευνα
Πληροφοριοδότης:	Ματσουκάρη Χρυστάλλα ετών 55 αγράμματος
Αξιοπιστία πληροφορίας:	
Μέγεθος:	
Σημειώσεις:	

Fig. 3 Notebook cataloguing template

Cataloguing template currently is in Greek language and so for being understandable the English label is given for each element. On the upper left side the current hierarchy level is shown (*Digital library*> *Folklore collection*> *Folklore notebooks*). The system offers specific features helping the cataloguer to save time, such as: the opportunity to fill an element with values taken from a list or to define elements with default values. Now in Pergamos are registered almost 1400 digitized notebooks with their metadata records. A team of three persons is responsible for digitizing and cataloguing the material and their experience and comments for the system environment and function are presented in the next paragraph of evaluation.

5.2. Evaluation

Pergamos evaluation will be held in two phases. The first one corresponds to a formative evaluation process and involves the surveillance of the metadata schema and templates implementation as well the easiness and usability of the cataloguing process. The second phase, which is a summative evaluation process, will start when the implementation of the search and retrieval functionalities is completed and involves several experiments to study the system performance, usefulness and usability.

The first evaluation phase has already started and a focus group, consisted of three folklore experts, two librarians and three information systems developers, has been surveyed. The survey results were encouraging because in general the focus group is satisfied by the system implementation. More specifically:

- Folklore experts: they are satisfied by the rich semantically and well-organized information structure. However, since the search mechanisms are under development they are not yet able to estimate the system usefulness.
- Librarians: the interfaces are friendly but they are cautious about the required cataloguing time. For the present the cataloguers fill as many elements as it is possible with the intention to register automatically first all the notebooks and then to continue with the cataloguing of the full details of each one.
- Developers: the system is well structured and provides much functionality for acquiring and managing all the characteristics of the collection and for preserving all the data about scanning and storing processes.

The second phase will be held when the implementation of the search and retrieval functionalities are completed and will deploy both empirical and automated techniques such as cognitive walkthroughs and log analysis. Several experiments will be done based on different user samples and their results will be compared in order to reveal and analyze the types of information behavior and needs [29]. In this way significant recommendations will be emerged concerning the system functionalities and user interaction improvement.

6. Interoperability issues

Current information landscape has changed since users need to have access to available cultural information without concerning about institutional and national boundaries. The intention of having the folklore collection wide exposure is one of the main concerns and therefore interoperability with other digital cultural collections and resources is in highest priorities. Since each cultural organization implements different description and preservation models, it is essential to facilitate mappings between them. There have been many studies and projects about mappings and crosswalks. MetaNet [30], is a metadata thesaurus providing the required semantic knowledge which facilitates the semantic mapping between metadata terms from different domains or standards. Another approach is based on concept-based mapping and defines relationships between the metadata semantics [31].

Among the information society, a good practice for interoperability and metadata exchange between different digital collection systems is considered to be the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [32]. MOSC project [33] examines the advantages of applying OAI-PMH to concurrently maximize the metadata exposure from digital systems of varying cultural organizations like museums, libraries and archives. The fact that Pergamos system fully supports OAI-PMH and OAI-PMH harvests metadata expressed in DC, address to map our proposed metadata model for collection and item level description to DC element set. Even though it is clear that there will be a loss of information by making a mapping from a rich metadata model to a standard (DC contains 16 core elements), the compatibility with international protocols contributes to the maintenance of a high level of consistency in retrieving information.

As a tool for generating the mapping, the Topic Maps (TMs) model and especially XTM 1.0 syntax [34] (DTD) was chosen. TMs provide are quite powerful in managing and creating links between different metadata vocabularies and are inherently flexible for defining various kinds of relationships. In a TM concepts can be organized in various ways and therefore there is not a unique way to create a TM.

We consider the mapping as a table correlating the semantics of two different metadata schemas (vocabularies). For each metadata element of the source schema we locate a semantically related element of the target schema. In particular we consider each metadata element as a *topic* and we define *types of associations* between the metadata elements. An association correlates two metadata elements that belong to different schemas. In an association each of the elements has a specific *role*.

As a demonstrator of our effort we selected as source schema the proposed collection level application profile and as target schema the DC CD AP. As mentioned earlier we used DC CD AP as core schema for the collection – level description and we enriched it with additional elements of other standards. We selected to present the mapping of our collection-level schema to the DC CD AP for two reasons: (a) the collection-level description is the starting access point to the plethora of the offered digital content in web and (b) DC CD AP is based on DC on which OAI-PMH is based. Since the collection description enables users to select the material they are really interested in, the collection-level metadata should be semantically rich, interoperable enough and suitable for effective navigation and retrieval. Concluding developing mappings with TMs facilitates mostly the semantic interoperability of the metadata schemas, while OAI-PMH focuses on syntactic issues.

The mapping procedure is as follows:

a) Every metadata element is considered as a “topic” with its own attributes, according to the metadata standard that comes from.

b) We define 3 topic types categorizing the elements of the two schemas: descriptive, administrative and structural metadata, according to [19]. Each metadata element is an instance of one of the above types.

c) We define specific “association” types correlating a couple of elements from the two different schemas: a) *equivalence:* for mapping elements that have the same meaning, b) *refinement:* to express a relationship between an element and its qualifier following exactly the DC and c) *hierarchical*: to connect elements that can be considered as broader and narrower concepts.

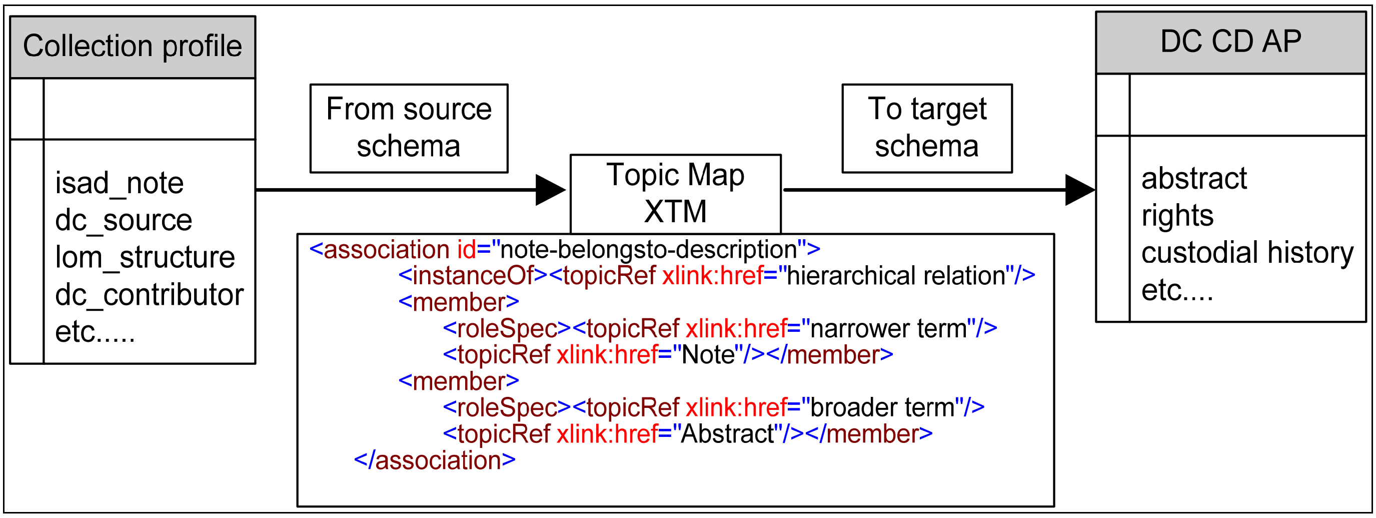
d) Each element in an association has a specific role. We have set the following couples of role types: *equivalent terms* for the “equivalence” association, *broader - narrower term* for the “hierarchical” and *element type – qualifier* for the “refinement” association.

Table 9 presents the defined association and role types for the mapping. In the table the common elements of the source (the application profile for collection level) and the target (DC CD AP) schemas are not appeared.

Table 9 Mapping elements

COLLECTION APPLICATION PROFILE		DC CD AP		ASSOCIATION TYPE
ELEMENT	ROLE	ELEMENT	ROLE	
(ISAD)_NOTE	NARROWER TERM	ABSTRACT	BROADER TERM	HIERARCHICAL
(ISAD)_LEGAL STATUS	NARROWER TERM	ABSTRACT	BROADER TERM	HIERARCHICAL
(ADL)_SCOPE/PURPOSE	NARROWER TERM	ABSTRACT	BROADER TERM	HIERARCHICAL
(DC)_SOURCE	NARROWER TERM	CUSTODIAL HISTORY	BROADER TERM	HIERARCHICAL
(RSLP)_LOCATION_PHYSICAL	EQUIVALENT TERM	IS LOCATED AT	EQUIVALENT TERM	EQUIVALENCE
(RSLP)_ACCRUAL STATUS	BROADER TERM	ACCRUAL_PERIODICITY ACCRUAL_POLICY ACCRUAL_METHOD	NARROWER TERMS	HIERARCHICAL
(LOM)_STRUCTURE	NARROWER TERM	CATALOGUE OR DESCRIPTION	BROADER TERM	HIERARCHICAL
(DCTERMS)_TABLE OF CONTENTS	QUALIFIER	CATALOGUE OR DESCRIPTION	ELEMENT TYPE	REFINEMENT
(DCTERMS)_RELATION	ELEMENT TYPE	ASSOCIATED PUBLICATION	QUALIFIER	REFINEMENT
(DC)_CONTRIBUTOR	NARROWER TERM	ABSTRACT	BROADER TERM	HIERARCHICAL
(RSLP)_LOCATION_DIGITAL	EQUIVALENT TERM	IS AVAILABLE VIA	EQUIVALENT TERM	EQUIVALENCE
(ADL)_METADATA SCHEMA	NARROWER TERM	CATALOGUE OR DESCRIPTION	BROADER TERM	HIERARCHICAL
(ADL)_METADATA MAPPING	NARROWER TERM	CATALOGUE OR DESCRIPTION	BROADER TERM	HIERARCHICAL

Figure 5 provides an illustrating example of how the topic map implements the mapping between the two schemas. Defining for example that the element *isad_note* is related with *dccdap_abstract* through a hierarchical association, the topic map operates as a mediator helping the translation from the source schema elements to the target ones.



```
<association id="note-belongsto-description">
        <instanceOf><topicRef xlink:href="hierarchical relation"/>
        <member>
            <roleSpec><topicRef xlink:href="narrower term"/>
            <topicRef xlink:href="Note"/></member>
        <member>
            <roleSpec><topicRef xlink:href="broader term"/>
            <topicRef xlink:href="Abstract"/></member>
    </association>
```

Fig. 5 The role of TM between the metadata schemas

Current research on semantic interoperability also focuses on the development of metadata mappings to domain-specific ontologies. In this way several metadata schemas could interoperate if each of them is mapped to the concepts of an ontology. The main advantage of this approach is that minimizes the effort for the development of mappings and crosswalks between different metadata standards. Moreover these mappings constitute core ontologies that extend the initial ontology and describe, represent and co-relate all the vocabularies of a domain [35]. Therefore we plan in the future to create a mapping of our multi-layered model to CIDOC/CRM [36] ontology, which consist a complete, event-oriented reference model for the cultural heritage.

7. Conclusions

The current research presents a hierarchical metadata model for manipulating complex folklore collections, designed according to material requirements. Specifically, the model follows the internal collection structure and focuses both on the collection and item characteristics. It contains elements from various metadata standards trying to cover the peculiarities of physical and digital objects, while it respects any restrictions regarding material usage. The expectation for semantic rich and meaningful information retrieval is satisfied, since all collection components are treated as separate digital objects with their own related information, structure and behavior in the system. Users have the possibility not only to browse the collection contents but also to find exactly, where their requested information resides inside the millions of text pages.

Furthermore, a suitable infrastructure for developing new and improved information services is presented with Pergamos system. Pergamos supports the described metadata policy, since it is object oriented system and manipulates digital objects according to the prototype instantiation process. A first attempt for evaluating the system, based on a focus group, is discussed while an evaluation plan was presented and will be executed when the system is completed. Finally, interoperability concerns are explored in order to achieve wide exposure of material and communicate with other collections. The development of topic maps provides a different approach for creating mappings between metadata schemas. Moreover the mapping of our model to CIDOC/CRM is in our future plans due to the advantages of the ontology driven semantic interoperability.

In general, the exposition of the folklore collection in an attractive way for users can be a powerful mean to enhance the cultural heritage study and research. The digital cultural objects can be transformed to valuable sources of information that will be referenced and used by other applications and systems.

8. Acknowledgments

The authors wish to thank George Pyrounakis and Kostas Saidis, Libraries Computer Centre of National & Kapodistrian University of Athens, Greece for their valuable help.

9. References

[1] NISO Framework Advisory Group., A framework of guidance for building good digital collections, 2nd Edition (2004).

[2] Ding, Challenges in building semantic interoperable digital library System. Object-oriented Systems Course Orientation, Spring 2005.

[3] G Crane, K Wulfman, Towards a cultural heritage digital library, In IEEE Computer Society (ed.), JCDL ΄03: Proceedings of Joint Conference on Digital Libraries (2003)

[4] T. Veen, R. Claypha, Metadata in the context of the European Library Project, In Conf. on Dublin Core and Metadata for e-Communities, 20002: Proceedings of International Conf. on Dublin Core and Metadata for e-Communities, (Firenze, 2002).

[5] L. Dempsey, Scientific, Industrial, and Cultural Heritage: a shared approach. Ariadne Issue 22 (1999)

[6] K. Rollitt, A. Kebbell, D. Campbell, Using Dublin Core for DISCOVER: a New Zealand visual art and music resource for schools, Int. Conf. on Dublin Core and Metadata for e-Communities, 2002: Proceedings of International Conf. on Dublin Core and Metadata for e-Communities, (Firenze, 2002).

[7] Dublin Core Metadata Initiative, Dublin Core element set (DC), Available at: http://dublincore.org/ (accessed 18 January 2006).

[8] Dublin Core Metadata Initiative (DCMI), DCMI metadata terms, Available at: http://dublincore.org/documents/dcmi-terms/ (accessed 18 January 2006).

[9] Network Development and MARC Standards Office, Encoded Archival Description (EAD), Available at: http://www.loc.gov/ead/ (accessed 18 January 2006).

[10] I. Lourdi, C. Papatheodorou, A metadata application profile for collection-level description of digital folklore resources, IEEE Computer Society, PEH 2004: Proceedings of 3rd International Workshop on Presenting and Exploring Heritage on the Web, (Spain, 2004).

[11] The European Library, Available at: http://www.theeuropeanlibrary.org (accessed 18 January 2006).

[12] Ana Maria de C. Moura; G. da Costa Pereira, M. L Machado Campos, A metadata approach to manage and organize electronic documents and collections on the web, Journal of the Brazilian Computer Society, 8(1), 2002

[13] C. M. Barreto. A Metadata Model for Describing Electronic Documents on the Web (in Portuguese), Master Thesis, IME-RJ, Aug. 1999.

[14] P. Johnston, B. Robinson, Collections and Collection Description. *Collection Description Focus Briefing Paper*, No 1, 2002.

[15] Minerva Project by UKOLN, University of Bath, Technical guidelines for digital cultural content creation programmes, version 1.0, revised 08 April 2004

[16] M. Sweet and D. Thomas, Archives Described at collection level, D-Lib Magazine, 6(9) (2000).

[17] L.M. Bartolo, C.S. Lowe, D.R. Sadoway, A.C. Powell and S.C. Glotzer, NSDL MatDL: Exploring Digital Library Roles, D-Lib Magazine, 11(3) (2005).

[18] Consultive committee for Space Data Systems (CCSDS), “Reference Model for an Open Archival Information System (OAIS).” Blue Book, Issue 1 (2002).

[19] NISO, *Understanding Metadata* (NISO press, 2004).

[20] Dublin Core Metadata Initiative, Dublin Core Collection Description Application Profile, Available at: http://www.ukoln.ac.uk/metadata/dcmi/collection-application-profile/2003-08-25/ (accessed 18 January 2006).

[21] International Council of Archives (ICA), International Standard Archival Description ISAD(G), Available at: http://www.ica.org/biblio/cds/isad_g_2e.pdf

[22] L. Hill, G. Janee, R. Dolin, J. Frew and M. Larsgaard, Collection Metadata Solutions for Digital Library Applications, Journal of the American Society for Information Science (JASIS), 50 (13) (1999), 1169-1181.

[23] Research Support Libraries Program (RSLP), Collection Description Schema, Available at: http://www.ukoln.ac.uk/metadata/rslp/ (accessed 18 January 2006).

[24] Computer Society/Learning Technology Standards Committee, IEEE Standard for Learning Object Metadata, Available at: http://ltsc.ieee.org/wg12/par1484-12-1.html (accessed 18 January 2006).

[25] Network Development and MARC Standards Office of the Library of Congress, 2002, Marc Standards, Available at: http://www.loc.gov/marc/ (accessed 18 January 2006).

[26] Network Development and MARC Standards Office of the Library of Congress, MIX: NISO Metadata for images in XML Schema, Available at: http://www.loc.gov/standards/mix/ (accessed 18 January 2006).

[27] G. Pyrounakis, K. Saidis, M. Nikolaidou, I. Lourdi, Designing an Integrated Digital Library Framework to Support Multiple Heterogeneous Collections, In: Springer-Verlag GmbH (ed.), ECDL 2004, Proceedings of 8th European Conference, (UK, 2004).

[28] G. Pyrounakis, K. Saidis, M. Nikolaidou, On the Effective Manipulation of Digital Objects: A Prototype-Based Instantiation Approach. In: Springer-Verlag GmbH (ed.), ECDL 2005: Proceedings of 9th European Conference, (Austria, 2005).

[29] N. Fuhr, G. Tsakonas, T. Aalberg, M. Agosti, P. Hansen, S. Kapidakis, C.-P. Klas, L. Kovacs, M. Landoni, A. Micsik, C. Papatheodorou, C. Peters, and I. Solvberg. Evaluation of digital libraries. (submitted), 2006.

[30] J Hunter, MetaNet - A Metadata Term Thesaurus to Enable Semantic Interoperability Between Metadata Domains, Journal of Digital Information (JODI), 1(8) (2001).

[31] M. Doerr, Semantic Problems of Thesaurus Mapping, Journal of Digital Information (JODI), 1(8) (2001).

[32] Open Archives Initiative, The Open Archives Initiative Protocol for Metadata Harvesting, Available at: http://www.openarchives.org/OAI/openarchivesprotocol.html (accessed 18 January 2006).

[33] E. Roel, The MOSC Project: Using the OAl-PMH to Bridge Metadata Cultural Difference across Museums, Archives, and Libraries, Journal of Information Technology and Libraries, 24(1) (2005) 22-24.

[34] TopicMaps.Org, XML Topic Maps (XTM) 1.0, Available at: http://www.topicmaps.org/xtm/1.0/ (accessed 18 January 2006).

[35] D. Tudhope, C. Binding, A Case Study of a Faceted Approach to Knowledge Organisation and Retrieval in the Cultural Heritage Sector. Digicult, Thematic Issue 6 - Resource Discovery Technologies for the Heritage Sector (2004) pp. 28-33.

[36] ICOM. The CIDOC Conceptual Reference Model, Available at: http://cidoc.ics.forth.gr/ (accessed 18 January 2006].