A multi-layer metadata schema for digital folklore collections

Irene Lourdi, Christos Papatheodorou and Mara Nikolaidou

Journal of Information Science 2007; 33; 197
DOI: 10.1177/0165551506070711

A multi-layer metadata schema for digital folklore collections

Irene Lourdi

Library Computer Centre, National and Kapodistrian University of Athens, and Laboratory on Digital Libraries and Electronic Publishing, Department of Archive and Libraries Sciences, Ionian University, Greece

Christos Papatheodorou

Laboratory on Digital Libraries and Electronic Publishing, Department of Archive and Libraries Sciences, Ionian University, Greece

Mara Nikolaidou

Harokopio University of Athens, and Library Computer Centre, National and Kapodistrian University of Athens, Greece

Received 26 February 2006
Revised 25 June 2006

Abstract.

Digital folklore collections are valuable sources for studying the 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 recovery from all structural levels. Furthermore, interoperability between the used metadata schemas is discussed and a Topic Maps model is presented as an approach for developing mappings.

Keywords: folklore collections; metadata policies; application profiles; topic maps

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].

Correspondence to: Irene Lourdi, Libraries Computer Centre, National and Kapodistrian University of Athens, University Campus, Department of Informatics and Telecommunications, Ilisia, 15784, Athens, Greece. Email: elourdi@lib.uoa.gr

A large volume of papers focuses on digitization projects and strategies in heritage institutions [2–4], 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 at the 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 on the 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 the University of Athens, and related experience. We also 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 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 are given in the last section, with discussion on the contribution of our work.

2. Problem definition and related work

The folklore collection of the University of Athens Greek Literature Department consists of handwritten travelling notebooks containing information about the way of living in various regions of Greece. The notebooks were written by the Department's students, from about 1967, and are accumulated in the 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 also collected by the author and related to the specific region. Notebooks are a significant source for a folklore researcher since they contain primordial information collected from interviewed inhabitants of each visited place. They consist of a quite large collection, containing more than 4000 notebooks and 350,000 pages.

For administrative reasons the folklore collection is divided into sub-collections according to the type of object. 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 inhabitants and places and small objects stuck on the pages like artefacts (e.g. lace or dolls). 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 as yet unknown.

(3) Objects sub-collection, consisting of objects that are either attached to the notebooks (unknown number) or on display in the library, which are estimated to number about 1000.

Despite the cultural value of its material, the collection has not currently been catalogued or recorded in an electronic system. Therefore (a) users are obliged to read and look through all the notebooks (thousands of pages) to find any information, and (b) the physical material and consequently the collection's intellectual content is constantly exposed to the effects of time and to unpredictable natural destruction. Thus, the expected benefits from the digitization project are to provide users with access mechanisms to folklore resources by organizing the material effectively and contributing to the long-term preservation of its valuable cultural information. Moreover, digitization is expected to increase the number of both *long-distance* visitors, who view the collection over the internet and retrieve information using specific access points such as subject, date, place etc., and *in-house* visitors, who are mainly folklore researchers that want to study the material closely 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 from the semantics and the relations of the items [5].

To the best of our knowledge there are some similar projects dealing with composite collections. The 'Discover Project' [6] manages New Zealand National Library's 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 the Dublin Core element set (DC) [7], qualified Dublin Core [8], Encoded Archival Description (EAD) [9] and also locally 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 the 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 the 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 at 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 is a first attempt at 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, 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 object, the corresponding chronological period, the geographic region, common provenance or even common usage [14].

Consequently, in order to manipulate and characterize a hybrid collection, it is necessary to divide the collection into sub-collections and represent the resulting internal structure. By defining a hierarchy in the collection, the sub-collections and items inherit specific characteristics, such as date, subject, geographic area etc., which 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 are parts of the 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 inheritance among a hierarchy of entities, avoid redundancy of information and link descriptions [16]. The desired policy will ensure both efficient information retrieval and the protection of the original content by keeping all the valuable information for authenticity and preservation. Users must have the potential to discriminate between the specific collection and the plethora of other similar ones in the web and decide whether it is of 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 retrieval of resources. A digital folklore collection shall allow material usage in a way that does not breach the intellectual rights associated with cultural objects and simultaneously respect every user’s learning and studying needs. There should be the possibility to define 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 stand against data safety and long-term preservation. Thus, a 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 the items they contain and promote education, academic research and preservation of cultural heritage and folklore features. In other words, digital folklore objects need to be transformed into valuable resources that can be used as reference by other collections and for related studies. Users need to study old and fragile objects without being concerned about possible destruction 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 the cultural heritage and oral tradition of a country is interesting for a wide audience of varied educational levels 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.

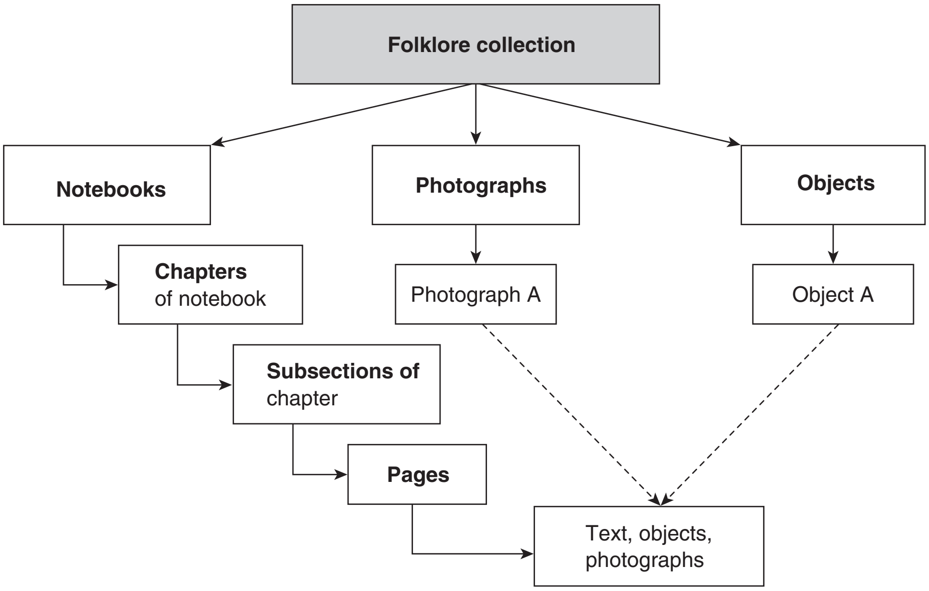


Fig. 1. Description levels of folklore collection.

4. Information architecture

4.1. Collection structure

A digital collection consists of digital objects that, on an abstract level, refer to real-world objects [18] usually organized in a hierarchical structure (collection, sub-collections and items). Figure 1 presents the structure of the folklore collection. The notebooks sub-collection is divided into the levels of chapter, subsection, page, text/objects/photographs. Each level contains a set of digital objects with their own attributes, description and behaviour. The photographs and handcrafts included in or attached to 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.

It should be noted 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 form a sub-collection. In such a case, librarians should re-organize the whole collection and therefore spend extra time on 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 rich documentation and simultaneously making easier the task of the cataloguer, who is obliged to fill in the specific metadata fields for each level.

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

Table 1
Metadata categories

Collection	Descriptive	
	Administrative for the physical	For the digital collection
	Structural	
Notebook	Descriptive	
	Administrative for the physical	For the digital version of the notebook
	Structural	
Chapter	Descriptive	
	Administrative	
	Structural	
Sub-section	Descriptive	
	Structural	
Page	Administrative	
	Structural	
Photograph	Descriptive	
	Administrative for the physical	For the digital versions of the photo
	Structural	
Objects	Descriptive	
	Administrative	
	Structural	

- providing full documentation about the content and context of objects e.g. when they were created, by whom, their usage, their contents etc.;
- facilitating queries and information retrieval from all the structural levels of the collection;
- combining different metadata standards in order to cover all the formats and types of physical and digital object;
- assuring 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.

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 being 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's nature (physical or digital). It is important to maintain information for the physical and digital version of the object, because physical characteristics also affect digital ones.

In the table sections showing the metadata fields for each component, some indications are given to allot specific attributes to each element. These indications show:

(a) the metadata standard from which each element originates (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 designed to be filled in 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 the special characteristics of the collection entity. These standards are: General International Standard Archival Description (ISAD(G)) [21], the metadata model

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	RSLP_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_FORMAT_MEDIUM (I)	DC CD AP_RIGHTS	ADL_METADATA MAPPING
DC CD AP_SIZE		
	Structural metadata	
LOM_STRUCTURE	DC CD AP_ASSOCIATED COLLECTION	DCTERMS_RELATION
DC CD AP_SUB-COLLECTION	DC CD AP_SUPER-COLLECTION	DC_DESCRIPTION_TABLE OF CONTENTS

of Alexandria Digital Library (ADL) [22], Research Support Libraries Program (RSLP) [23] and IEEE's Learning Object Metadata (LOM) [24]. In Table 2 their abbreviations are encapsulated in the label of each element.

In particular the refined term *dc_description_tableofcontents* describes the contents of the collection. The element *dc_relation*, with its additional refined terms, encodes the types of relation 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 to the collection besides collector and owner.

Furthermore, elements from ISAD(G) were added such as: *isad_legal status* of the collection and *isad_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 *lom_structure* defines the type of collection structure (e.g. hierarchical, linear or networked) and the *rslp_location* elements specify the place of both the physical and digital collections.

However, the most important elements are *adl_metadata schema* and *adl_metadata mapping*, which encode the metadata standards used for the description of collection objects, and *adl_scope/purpose,* which 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 by the notebook schema are *rslp_location_physical*, *dccdap_language*, *dccdap_rights* and *dccdap_access rights*, while the values of elements *dc_format_medium* and *dccdap_size* are computed as summations of the corresponding elements in the notebooks schema.

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

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_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	
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_TABLE OF CONTENTS (I)	

element *dc_title* for the related chapters. The elements *dc_coveragespatial* and *dc_date_accumulated*, are inherited by 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 *otherphysicaldetails* describes the physical characteristics of the notebook that may not be covered by the other elements, such as whether a page of the notebook is missing. The locally defined element *credibility* is added at the specific request of the folklore experts because the notebooks contain primordial information that has not been studied before. The MARC element *bindinginformation* is added to keep information about the way the notebooks have been compiled in a volume and according to which criteria.

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 scanned pages that belong to the specific chapter in the case of the original notebook.

The chapters of a notebook are divided further into 'subsections' predefined according to a 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 in the chapter to the student. The element *dc_subject* is mandatory at this level, since the values of the element are copied automatically to the upper corresponding element *dc_subject* in the notebook schema.

The elements *pixelSize, physScanResolution* etc. in Table 6 describe technical details about the scanning process of each page and the element *otherphysicaldetails* maintains characteristics of the specific pages that need to be preserved. This technical information is encoded by elements from the MIX schema [26].

Table 4
Chapter entity metadata

Chapter schema		
Descriptive metadata	Administrative metadata	Structural metadata
DC_TITLE (M)	DC_FORMAT_EXTENT (I) (FOR PHYSICAL)	DC_TABLE OF CONTENTS (I)
DC_COVERAGE_SPATIAL	DC_IDENTIFIER (M)	DC_RELATION_(IS A CHAPTER OF…)
DC_DATE_ACCUMULATED		

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_table of contents (I)
dc_relation_has "photograph"/ "object"	

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).

The element *dc_coverage_spatial* is copied automatically from the notebook to the 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, they 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. As 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.

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…)		

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 elements used for describing physical objects are given in Table 8. The elements *dc_coverage_spatial* and *dc_subject* are inherited from the upper levels and the same condition as for photographs exists also for the subjects.

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 adequate metadata, providing a 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 southern 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: *tableofcontents*); or
- 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 the elements. This can be weighed 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 the values of specific elements are copied to corresponding elements at higher or lower levels.

5. Implementation and evaluation

5.1. Implementation

Pergamos digital repository system [27] has been developed by the Library 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 the DC set.

Within Pergamos, each digital object consists of four parts:

(a) the metadata sets that describe the content of the object;

(b) the structure, which 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) behaviour, a set of methods that define how the object can be manipulated in the system.

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’ constituent parts (metadata sets, files, internal structure and behaviour 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 resultant objects will be aligned with the specifications of the prototype. Hence, all information in Pergamos is stored in terms of digital object prototypes and their instances.

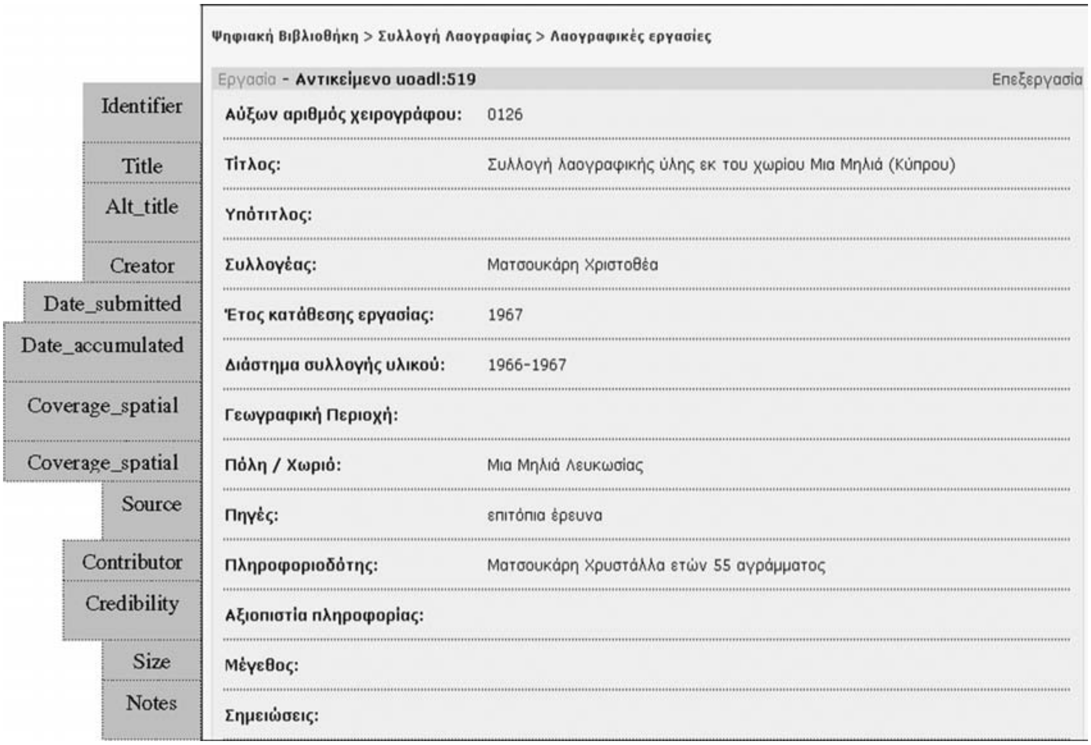
The proposed architecture consists of a set of prototypes, each one describing a different hierarchy level. So each collection and sub-collection is an instance of the collection prototype, while the notebook prototype is used for the implementation of notebook objects (notebook level) etc. Further, 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 user-friendly templates. The fields and attributes described in the notebook prototype shown in Figure 2 are included in the notebook cataloguing form presented in Figure 3.

The cataloguing template is currently in the Greek language and so to aid understanding an English label is given for each element in Table 3. On the upper left side the current hierarchy level is shown (*Digital library* > *Folklore collection* > *Folklore notebooks*). The system offers specific features helping

```
<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">τττττττττ</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.



Ψηφιακή Βιβλιοθήκη > Συλλογή Λαογραφίας > Λαογραφικές εργασίες

Εργασία - 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.

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. Almost 1400 digitized notebooks with their metadata records are now registered in Pergamos. 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 Section 5.2.

5.2. Evaluation

The Pergamos evaluation will be carried out in two phases. The first one corresponds to a formative evaluation process and involves the surveillance of the metadata schema and template 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's performance, usefulness and usability.

The first evaluation phase has already started and a focus group, consisting 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 with the semantically rich and well-organized information structure. However, since the search mechanisms are under development they are not yet able to estimate the system's usefulness.
- Librarians – they find the interfaces friendly but they are cautious about the required cataloguing time. For the present, the cataloguers fill as many elements as is possible with the intention of first registering automatically all the notebooks and then continuing with the cataloguing of the full details from each one.

- Developers – they feel the system is well structured and provides a lot of 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 is 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 analyse the types of information behaviour and needs [29]. In this way significant recommendations will emerge concerning the system functionalities and user interaction improvement.

6. Interoperability issues

The current information landscape has changed since users need to have access to available cultural information without concerning themselves with institutional and national boundaries. The intention of giving the folklore collection wide exposure is one of the main concerns and therefore interoperability with other digital cultural collections and resources is one of the 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 to facilitate semantic mapping between metadata terms from different domains or standards. Another approach is based on concept-based mapping and defines relationships between metadata semantics [31].

In the era of the information society, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [32] is considered to be good practice for interoperability and metadata exchange between different digital collection systems. The MOSC project [33] examines the advantages of applying OAI-PMH to maximize the metadata exposure from digital systems of varying cultural organizations like museums, libraries and archives concurrently. Since Pergamos fully supports OAI-PMH and therefore harvests metadata expressed in DC, our metadata model is mapped to the 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 no unique way to create a TM.

We consider the TM 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 demonstration 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 use the DC CD AP for the mapping of our collection-level schema for two reasons:

(a) the collection-level description is the starting access point to the plethora of digital content offered on the 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. In conclusion, developing mappings with TMs mostly facilitates 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 a ‘topic’ with its own attributes, according to the metadata standard that it comes from.

(b) We define three 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*, to map elements that have the same meaning, (b) *refinement*, to express a relationship between an element and its qualifier following the DC exactly, 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.

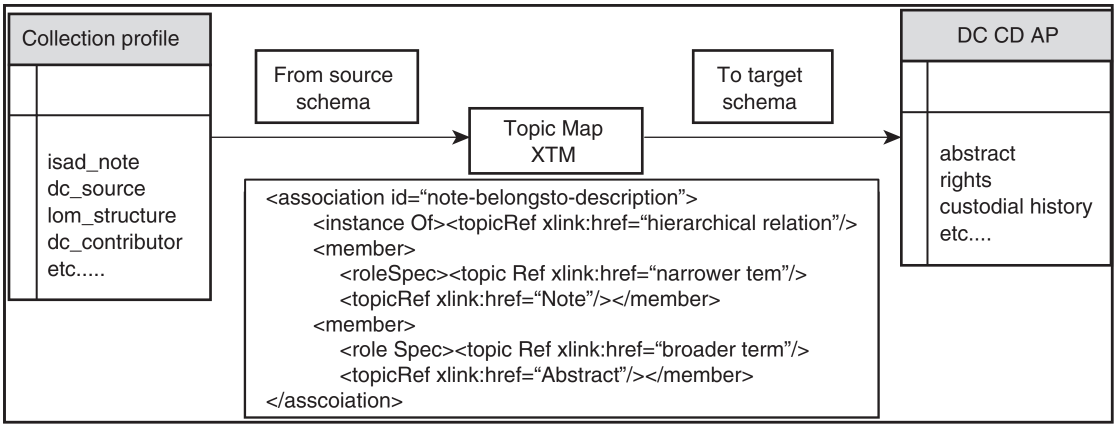
Table 9 presents the defined association and role types for the mapping. In the table the common elements of the source (the application profile for collection level) and the target (DC CD AP) schemas do not appear.

Figure 4 provides an illustration of how the topic map implements the mapping between the two schemas. Assuming 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.

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 was mapped to the concepts of an ontology. The main advantage of this approach is that it 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

Table 9
Mapping elements

Collection application profile		dc cd ap		
Element	Role	Element	Role	Association type
(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



```
<association id="note-belongsto-description">
      <instance Of><topicRef xlink:href="hierarchical relation"/>
      <member>
         <roleSpec><topic Ref xlink:href="narrower tem"/>
         <topicRef xlink:href="Note"/></member>
      <member>
         <role Spec><topic Ref xlink:href="broader term"/>
         <topicRef xlink:href="Abstract"/></member>
</asscoiation>
```

Fig. 4. The role of TM between the metadata schemas.

describe, represent and correlate 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 comprises 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 on 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 semantically rich and meaningful information retrieval is satisfied, since all collection components are treated as separate digital objects with their own related information, structure and behaviour 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 the Pergamos system. Pergamos supports the described metadata policy, since it is an object-oriented system and manipulates digital objects according to the prototype instantiation process. A first attempt to evaluate the system, based on a focus group, is discussed while an evaluation plan is 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 means to enhance cultural heritage study and research. Digital cultural objects can be transformed into valuable sources of information that will be referenced and used by other applications and systems.

Acknowledgments

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

References

[1] NISO Framework Advisory Group, *A Framework of Guidance for Building Good Digital Collections* (NISO, Bethesda, MD, 2004).

[2] H. Ding, *Challenges in Building Semantic Interoperable Digital Library System* (2005). Available at: www.idi.ntnu.no/grupper/su/courses/dif8901 (accessed 11 September 2006).

[3] G Crane and K Wulfman, Towards a cultural heritage digital library. In: H.C. Chen et al. (eds), *Proceedings of the Joint Conference on Digital Libraries JCDL ´03* (IEEE Computer Society, Houston, TX, 2003) 75–86.

[4] T. Veen and R. Claypha, Metadata in the context of the European Library Project. In: *Proceedings of the International Conference on Dublin Core and Metadata for e-Communities*, *Florence, 2002* (Firenze University Press, Florence, 2002) 19–26.

[5] L. Dempsey, *Scientific, industrial, and cultural heritage: a shared approach* (2000). Available at: www.ariadne.ac.uk/issue22/dempsey/ (accessed 11 September 2006).

[6] K. Rollitt, A. Kebbell and D. Campbell, Using Dublin Core for DISCOVER: a New Zealand visual art and music resource for schools. In: *Proceedings of the International Conference on Dublin Core and Metadata for e-Communities*, *Florence, 2002* (Firenze University Press, Florence, 2002) 251–6.

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

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

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

[10] I. Lourdi and C. Papatheodorou, A metadata application profile for collection-level description of digital folklore resources. In: *Proceedings of the 3rd International Workshop on Presenting and Exploring Heritage on the Web, PEH 2004* (IEEE Computer Society, Zaragoza, 2004) 90–94.

[11] *The European Library* (2005–6). Available at: www.theeuropeanlibrary.org (accessed 18 January 2006).

[12] Ana Maria de C. Moura; G. da Costa Pereira and 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) 16–31.

[13] C. M. Barreto, *A metadata model for describing electronic documents on the web* (Unpublished Master's thesis, IME-RJ, 1999). [In Portuguese]

[14] P. Johnston and B. Robinson, *Collections and Collection Description*. *Collection Description Focus Briefing Paper*, *No 1* (2002). Available at: www.ukoln.ac.uk/cdfocus/briefings/bp1/bpl.pdf (accessed 11 September 2006).

[15] University of Bath, UKOLN, *Minerva Project: Technical guidelines for digital cultural content creation programmes, version 1.0, revised 8 April 2004* (2004). Available at: www.minervaeurope.org/structure/workinggroups/serveprov/documents.htm (accessed 11 September 2006).

[16] M. Sweet and D. Thomas, Archives described at collection level, *D-Lib Magazine* 6(9) (2000). Available at: www.dlib.org/dlib/september00/sweet/09sweet.html (accessed 11 September 2006).

[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). Available at: www.dlib.org/dlib/march05/bartolo/03bartolo.html (accessed 11 September 2006).

[18] Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS). Blue Book Issue1* (CCSDS, Washington, DC, 2002).

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

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

[21] International Council of Archives, *International Standard Archival Description ISAD(G)* (2000). Available at: www.ica.org/biblio/cds/isad_g_2e.pdf (accessed 11 September 2006).

[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*, 50(13) (1999), 1169–81.

[23] Research Support Libraries Program, *Collection Description Schema* (2001). Available at: www.ukoln.ac.uk/metadata/rslp/ (accessed 18 January 2006).

[24] IEEE 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, *MARC Standards* (2002). 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* (2004). Available at: www.loc.gov/standards/mix/ (accessed 18 January 2006).

[27] G. Pyrounakis, K. Saidis, M. Nikolaidou and I. Lourdi, Designing an integrated digital library framework to support multiple heterogeneous collections. In: R. Heery and L. Lyon (eds), *Proceedings of the 8th European Conference, ECDL 2004, UK, 2004* (Springer, Berlin/Heidelberg, 2004) 26–37.

[28] G. Pyrounakis, K. Saidis and M. Nikolaidou, On the effective manipulation of digital objects: a prototype-based instantiation approach. In: A. Rauber et al. (eds), *Proceedings of the 9th European Conference, ECDL 2005, UK, 2004* (Springer, Berlin/Heidelberg, 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, *International Journal of Digital Libraries* 2006 (in press).

[30] J Hunter, MetaNet – a metadata term thesaurus to enable semantic interoperability between metadata domains, *Journal of Digital Information* 1(8) (2001). Available at: http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Hunter/ (accessed 11 September 2006).

[31] M. Doerr, Semantic problems of thesaurus mapping, *Journal of Digital Information* 1(8) (2001). Available at: http://jodi.tamu.ed/Articles/v01/i08/Doerr/ (accessed 11 September 2006).

[32] Open Archives Initiative, *The Open Archives Initiative Protocol for Metadata Harvesting* (2004) Available at: 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* (2001). Available at: www.topicmaps.org/xtm/1.0/ (accessed 18 January 2006).

[35] D. Tudhope and C. Binding, A case study of a faceted approach to knowledge organisation and retrieval in the cultural heritage sector, *Digicult,* (2004) 28–33. [Thematic Issue 6: Resource Discovery Technologies for the Heritage Sector]

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