Journal of Computational Methods in Sciences and Engineering 6 (2006) S289–S298
IOS Press

Implementing digital folklore collections

Irene Lourdi[a,*], Mara Nikolaidou[b] and Christos Papatheodorou[c]
[a]*Department of Archive and Library Sciences, Ionian University, Corfu, Greece; Libraries Computer Centre, University of Athens, Athens, Greece*
E-mail: elourdi@lib.uoa.gr
[b]*Harokopio University of Athens, Greece; Libraries Computer Centre, University of Athens, Athens, Greece*
E-mail: mara@di.uoa.gr
[c]*Department of Archive and Library Sciences, Ionian University, Corfu, Greece*
E-mail: papatheodor@ionio.gr

Abstract. In this paper, a metadata model is analyzed to describe the digitized digital folklore collection of the Department of Greek Studies in the University of Athens. Folklore collection consists of different kinds of digitized material and requires a specialized data schema to describe its peculiarities. The collection big size and material variety results in representing the collection as a hierarchical structure, according to the type of objects, the corresponding chronological period and geographic region. The main goal is to preserve and popularize to all kinds of users the precious information regarding cultural collection data. For this purpose a metadata model is developed that enables efficient navigation to the notebooks sub-collection, as well as meaningful information retrieval to the collection objects.

Keywords: Metadata policies, cultural heritage material, application profiles

1. Introduction

Folklore collections are valuable sources for study and research the cultural heritage of a society or a group of people. They refer to various aspects of every-day life, such as: customs, music, architecture, clothing, handicraft, folk tales and oral tradition and reflect the common way of thinking and living. In order to preserve and popularize collections of cultural heritage through the web, many digitization projects take place that digitize and describe cultural objects with adequate metadata elements. University of Athens has initiated a similar project, aiming at the digitization and presentation of the Folklore collection belonging to the Department of Greek Studies.

Folklore collection consists of sub-collections of different kinds of material, such as the sub-collection of travelling notebooks, the sub-collection of sound recordings and the sub-collection of physical objects exposed in the library. The collection big size and the material variety demand the collection representation as a hierarchical structure, according to objects type, corresponding chronological period and geographic region. The main difficulty for managing folklore collections is the material heterogeneity

*Corresponding author: Irene Lourdi, Libraries Computer Centre, National and Kapodistrian University of Athens, University Campus, Department of Informatics & Telecommunications, Ilisia, 15784, Athens, Greece. Tel.: +30 210 727 5618; E-mail: elourdi@lib.uoa.gr.

(handwritten texts, photographs, 3D objects, sound recordings, maps) that requires the application of different digitization, description and maintenance practices. Further, a wide range of users of varied educational level and preferences (students, historians, philologists, psychologists, ethnologists) are interested in searching and retrieving information from the cultural heritage collections.

While designing and organizing the digital collection, some issues have to be taken into account, such as: a) to show the structure of the collection and its sub-collections by organizing the material into groups under specific criteria, b) to make a full diagram of the metadata model that will be used for the description of the material and c) to define the policy and the way the metadata model will affect the efficient retrieval of information by users.

This paper extends our previous work of defining a description model for collection level entity using as example the same folklore collection [1], and focuses on presenting a metadata model for describing the notebooks sub-collection. The metadata model facilitates efficient navigation to the notebooks sub-collection structures and meaningful information retrieval to the collection objects. The presented model is based on the concept of separating big and complex collections into simple and distinct entities that can be treated in a digital system as unique digital objects with their own attributes. This concept has been analyzed extensively in [2]. In the next section, we provide a short description of the notebook sub-collection structure and the requirements imposed for the cultural items. In Section 3, the metadata model for describing collection material is presented giving simultaneously emphasis on the representation of object relationships and related constraints. In Section 4 the benefits from the implementation of a similar proposed metadata policy description are discussed, while conclusions reside in Section 5.

2. Notebook sub-collection description

2.1. Collection structure

The travelling notebooks sub-collection is a very good sample of a collection with complexity and heterogeneity since it contains various kinds of material. The basic physical component of the collection is the notebook which has been written by the students of the University Folklore Department with the intention to write down in detail the cultural features of a place/village of Greece and keep that information for the future. The collection size is quite big, about 4000 written essays, and covers almost the whole country. Besides the text (handwritten or not), the notebook consists of photographs or pictures and small objects stuck on the pages by the students, in order to make the content of the narrative text more expressive and valid. Also it must be noticed that the structure of the notebooks follows a questionnaire prepared by folklore specialists. In the current situation, the notebooks have not been catalogued or registered to an electronic system, so users are obliged to read and look all the notebooks in order to find the information they want.

For the best administration of the collection in the Digital Library system and for making the material easily retrievable, the notebooks sub-collection has been separated into smaller units. Defining structural levels for a digital collection offers the possibility to handle them as separate digital objects with their own characteristics and description data. These logical entities are described from the proposed metadata model and follow the hierarchical structure given in Fig. 1.

2.2. Requirements imposed

Notebooks representation on the web depends quite on the structure and organization of the collection. The developed metadata model for each notebook has to ascribe to each level the necessary attributes.

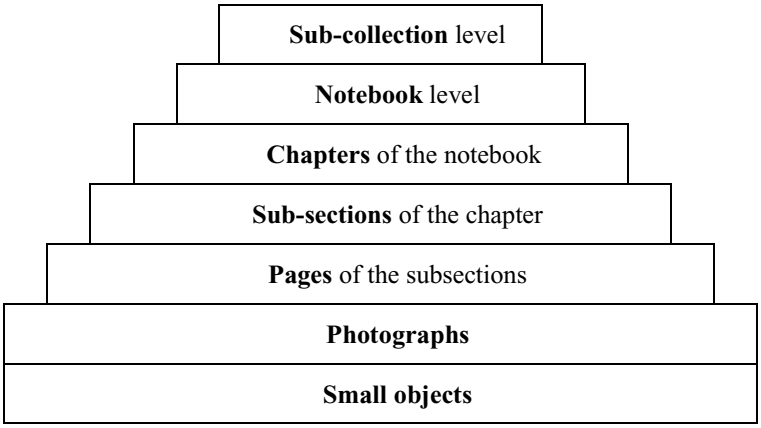

Fig. 1. Description levels of Notebook Sub-collection.

Thus, it is required: i) to express the subject coverage of the resources and the geographic region that every notebook covers ii) to express clearly the relationships that exist between the digital objects through all the levels iii) to contain elements concerning content description, administration and the notebooks stable structure iv) to be characterized from the policy of "inheritance" (transferring values) from the parent to the children and sometimes the opposite way (as it is explained later) and v) to contain elements concerning access rights in order to protect the copyrights of the oral tradition [3].

The description of the resources needs to be supported from a metadata model that depicts the existing collection structure collection and express further the distinctiveness and the relations between the various objects. Also rules must exist for the metadata model concerning how it will be implemented in the local digital library system. Since digital folklore collections are valuable sources for studying the cultural heritage of a country, the model, besides structure, must also express the semantic definition of folklore material. It is required to register all the cultural information concerning not only with the material creation and administration but also the customs and ideas this material represents and the constant changes that occur on folklore resources during time.

What we need is a metadata model that will make the digital collection functional for the users and will provide them the best way to access and retrieve what they want either by browsing the notebooks one by one or by searching in the contents of them using access points like subject, geographic region and time. In order to achieve such a functional digital collection, we have planned a metadata model that describes every level of the collection and defines the adequate elements to give a full and rich content description for them, while simultaneously generates digital operations making easier the cataloguer' s job.

3. Proposed metadata schema

It is important to emphasize that the resources nature and collection structure affect the digital collection description. The proposed metadata model has been designed taking into account the following issues related to the collection:

1. The notebooks have a stable structure, since students have written them according to the questionnaire prepared by folklore researchers.

COLLECTION	Descriptive	
	Administrative for the physical	For the digital collection
	Structural	
NOTEBOOK	Descriptive	
	Administrative for the physical	For digital version
	Structural	
CHAPTER	Descriptive	
	Administrative	
	Structural	
SUB-SECTION	Descriptive	
	Structural	
PAGE	Administrative	
	Structural	
PHOTOGRAPHS	Descriptive	
	Administrative for physical photograph	Administrative for digital versions of the photograph
	Structural	
OBJECTS	Descriptive	
	Administrative	
	Structural	

Fig. 2. Metadata categories.

2. The main target of digitization project is to adjust the digital collection to the users' needs and to provide "user centered" retrieval of information.
3. Users need to have complete access to all the defined structural levels of the collection by making queries and receiving responses from all the metadata templates.
4. The type and characteristics of the physical objects greatly affect the description of their corresponding digital objects. So in our case it is necessary to combine elements from different metadata standards in order to cover both the physical and digital versions of an object.

3.1. Categories of metadata standards

According to the NISO paper [4], metadata elements are separated in the following categories: descriptive metadata, structural metadata and administrative metadata. Specifically, "descriptive metadata" are responsible for the description of the resources content in order to help users find the item they are looking for. On the other hand, "structural metadata" describe the structure of the resources and how they are internally organized. The specific metadata category is quite important in cases of composite objects with complex structure and multiple structural levels (like the folklore notebooks). Finally, the "administrative metadata" provide information about the administration of resources and technical information about when and how the described item has been created or digitized.

According to these categories, we have created a hierarchical picture of the metadata elements preserved for every logical entity of our notebook collection. The metadata categories covered from the proposed metadata model is given in Fig. 2.

3.2. The metadata model

The proposed model for the collection notebooks description combines elements from a variety of metadata standards to describe and cover the characteristics of many thematically interlinked subcollections with composite objects. The model is mostly based on the Dublin Core Metadata Initiative [5] for both collection-level description and item-level description. In order to cover the requirements we set

Table 1
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)	SUBJECT_CLASSIFICATION (L)
DC_CONTRIBUTOR (ROLE)	DC_SUBJECT	
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_TABLEOFCONTENTS (I)	

in chapter 2.2 above, we have extended the Dublin Core element set with some locally defined elements or with elements taken from other metadata standards suitable for the specific types of resources.

Specifically, we have used Marc [6] for describing the physical objects characteristics and NISO "technical data for still images" [7] to provide technical information about the scanning process of the notebooks, images and small objects, which are placed inside the pages. The schema for the collection-level description is an application profile [8] since it is based on the Dublin Core Collection Description Application Profile [9] and is enriched with elements from other metadata standards for collection description like: ISAD(G) [10], the metadata model of Alexandria Digital Library (ADL) [11], Research Support Libraries Program (RSLP) [12] and IEEE-Learning Object Metadata (LOM) [13], as it is described in [1].

In the following tables we present only a part of the metadata model that deals with the entities of notebook, chapters, subsections and pages, in order to give a general but explanative picture of how the model functions. The elements are separated in the categories described in Fig. 2, and according to the described resource nature (physical or digital). It is necessary to keep information both for the physical and the digital version of each entity, since the characteristics of the physical item affect also the digital ones.

There are also defined some indications to express attributes for each metadata field. These indications show: i) from which metadata standard each element comes from (**DC** = Dublin Core, **Marc**, **L** = local (made for our project) ii) whether the element is mandatory to be filled (**M** = mandatory) and iii) which elements are set to be automatically filled from the system taking values from lower or upper levels (**I** = inherit).

The values of some metadata elements are inherited to the corresponding elements of next levels, in order not to be filled again. For example the elements that pass from the notebook to the chapters are: "Coverage_spatial" and "date_created". Also from the chapter to the subsections is inherited the element "coverage_spatial" with its values.

Table 2
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		

Table 3
Subsection entity metadata

SUBSECTION SCHEMA	
DESCRIPTIVE METADATA	
DC_IDENTIFIER (M)	DC_SUBJECT (M)
DC_TITLE (M)	SUBJECT_CLASSIFICATION (L)
DC_DESCRIPTION_ ABSTRACT	DC_COVERAGE_TEMPORAL
DC_CONTRIBUTOR	
STRUCTURAL METADATA	
DC_RELATION (IS SUBSECTION OF CHAPTER	DC_DESCRIPTION_CONTENTS (I)
DC_RELATION_HAS PHOTOGRAPH / OBJECT	

Table 4
Page entity metadata

PAGE SCHEMA	
DC_IDENTIFIER (M)	FILE SIZE
SCAN PIXEL SIZE (NISO)	OTHER PHYSICAL DETAILS (NISO)
SCANNING RESOLUTION (NISO)	RELATION (IS PAGE OF THE SUBSECTION)
SCAN BIT DEPTH (NISO)	DC_DATE CREATED (M)

Further, the elements “subject” and “format extent” for notebooks are automatically filled by taking values from the templates of the chapters and the element “coverage_temporal” also by taking the values of the same element from the templates of the chapters. The same exists for the element “description_contents” that is filled automatically with the values from the element “title” from the chapters, without being the cataloguer responsible to write them down manually. Similarly, the contents of chapters are automatically filled from the titles of the subsections. The element “format_extent” in chapter is filled by adding the number of the pages that belong to the specific chapter.

In general, the model has been designed to take advantage the operations of the digital system that hosts the collection in order to gain time and make the process of material documentation easy and effective. Hence, there is an internal interchange of values between the fields of the notebook structural

levels based on conditions we have set for our project' s needs. The conditions and rules, that address the proposed metadata model, are analyzed in the following paragraphs.

3.3. Metadata model rules

Metadata model rules define the function and presentation of the digital entities. According to the general policy we propose the following rules:

1. The Dublin Core elements follow the encoding schemes that are defined by the DCMI, for example the dates must have the format of ISO 8601 "standard for dates and times" (W3CDTF) [14].
2. The element "DC_Description_contents" will be filled automatically by taking values from the element "DC_title" from the lower levels. This is a way to earn time and effort for the cataloguer in order no to fill every time the contents of each level by hand. So the contents of each chapter come from the title of each subsection. In case that somebody wants to fill the contents by hand it is proposed to fill the element "description_abstract" that is a free text.
3. The element "dc_format_extent" is also proposed to be filled automatically by the system taking values from the same element but from the lower levels (if they have been filled).
4. The element "dc_publisher" is proposed to express the entity responsible for making available the content of the collection to the web. So in our case represents our Department "Libraries Computer Centre" or the Library of Folklore Department.
5. In general the elements dc_subject and DC_coverage will be filled by values coming from locally defined vocabularies or lists with authority subjects. It is required to keep a specific level of homogeneity and to describe fully the content of the resources.
6. About the element "DC_rights" it is proposed to be inherited to all the structural levels automatically, with the assumption that all the rights are common for all the resources of the collection. In case that the rights of a digital entity are different from the whole collection's then it is proposed to be filled manually.

3.4. Local extensions -refinements

Due to the nature of folklore collection and the complexity that characterizes the resources and the content, we have extended some elements of Dublin Core by setting refinements, in order to give more precisely the content and the context of the collection. These refinements are:

1. The element "dc_Subject" is proposed to be characterized more precisely, so we have set a local refinement: Subject_Classification, that corresponds to the tag 080 of Marc21.
2. The element "dc_coverage_Spatial" is important for easily retrieving information from the thousands notebooks for a specific region or village. Especially in the collection of notebooks, places are characterized from a unique hierarchy that corresponds to the greek local government (village, town, province, nomarchy, area). In order to keep this hierarchy and to provide this information to the users we set two more refinements: "coverage_spatial_specification", that is to define the hierarchy that characterizes the place (village-town. . .) that the notebook is about and "coverage_spatial_info", that is to give information about the place e.g. the current name of the place.
3. Further, is proposed to refine each person referred in elements like "dc_creator", "dc_owner" "dc_collector" and "dc_contributor" with the local refinement "role" by taking values from the Marc list [15]. Also it is proposed for administrative reasons to give information for the entities "dc_owner" and "dc_collector" using the attributes of "vcard namespace", as it is given in RSLP metadata schema.

4. In order to express exactly the relationships that exist between every logical entity of the collection (collection, notebook, chapter. . .) we have extended the given refinements (qualifiers) of the element of "DC_relation". For example: we have extended the refinement "DC_relation_HasPart" by saying for the chapter: "is a chapter of the notebook. . . " or for the sub-section "has photograph. . . or has object. . .), declaring exactly the kind of relation between the two digital objects.

3.5. Functional inheritance of attributes

The collection big size (it contains almost 4000 notebooks) and the notebook complex structure (text, photographs and objects) makes even more difficult the work of the cataloguer to describe the containing items with all the required details. It requires too much effort and time to describe all the chapters, subsections and the additional material with detail, even though the cultural data that lay inside are quite valuable and difficult for someone easily to find.

For that reason, it is decided that the Digital Library system must support the policy of inheritance of the attributes from one level to another. It has been defined in the system that the values of many metadata elements from the model, are inherited automatically from one logical entity to another in order not to fill them every time. By this way it is possible to earn time and effort from the daily cataloguer' s job. Except from of the policy of transferring the values of elements between the various levels of the notebook, we have also set to happen the opposite. Some elements are filled automatically by taking values from lower levels (as it has already been said in chapter 3). In order to get in function this kind of policy, some "expressive tools" are required, that will implement the inheritance of the attributes from one level to another (up or down).

4. Benefits from the proposed description policy

4.1. Information retrieval

The presented model focuses both on the notebooks collection and its components, so that the user can access the collection and every notebook separately. Users have the possibility to find information either by browsing the list of all the notebooks or by looking the table of contents of each notebook or by searching in the contents of each digital entity using keywords and values from specific lists of geographical places, subjects, persons and chronological periods. Further, the digital system allows the combination of selection criteria and the combination of searching in various levels or the collection.

The proposed metadata model facilitates users with additional searching capabilities. For example, in case that a user searches for information about the marriage customs of *"Helateia"* village, he/she has direct access to the chapters and subsections of the notebooks that contain information about this matter as also to any photos or objects that satisfy the searching criteria. The distinction of a notebook into digital components with their own attributes and metadata description, facilitates information retrieval since users have access to all the collection contents without spending time in browsing each notebook and page.

4.2. Flexible composite objects administration

The folklore notebook is represented as a set of objects that belong to specific types: notebook, chapter, subsection etc. Every type of object corresponds to a metadata schema which contains the appropriate

informative data about their content, administration and structure. The types have been defined according to the physical composition of the notebooks and they are pointing to specific data objects. Notebooks and chapters contain descriptive, administrative and structural metadata, while page objects contain information about the scanning process without descriptive metadata.

Separating a composite object, like the folklore notebooks, into smaller data objects and defining for each digital object the appropriate metadata elements allows administrators and cataloguers to handle them with flexibility. This means that there is the potentiality in a digital system:

1. To express relations between independent data objects belonging to the same or a different collection.
2. To represent the collection structure
3. To define specific access rights or restrictions to each object depending the collection level
4. To extend navigation functionality in the composite object like in the folklore notebook

4.3. Support for big heterogeneous collections

The point described above depicts the requirement for the efficient handling of multiple, heterogeneous collections by a digital library system.The heterogeneity and the big amount of the resources warrant the need for separating the material into collections and sub-collections to represent complex structures and to accredit rich semantics to any level. Defining sub-collections as separate entities and giving a high-level metadata collection description simplifies the collection management and helps the navigation, discovery and selection of cultural content [16]. Also the collection-level description facilitates the retrieval of cultural data since users can decide whether the collection is of their interest without getting into details about the containing objects.

5. Conclusions

The paper intends to define and implement a general metadata model that facilitates the retrieval of information of digital folklore collections consisting of heterogeneous resources. Taking into account that, the notebooks collection, to which this study focuses on, currently is not functional or easily accessible by users, a description policy for affectively describing and administering large digital folklore collections is presented. The description policy proposes the division of composite collections into hierarchical levels further processed as separate digital entities/objects with their own semantics and documentation. The main intention is to facilitate easy information retrieval from big collections and to comprehend the content and context of the containing resources.

References

[1] I. Lourdi and 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.
[2] G. Pyrounakis, K. Saidis, M. Nikolaidou and 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.
[3] T. Cole, Creating a Framework of guidance for building good digital collections, *First Monday* **7**(5) (2001).
[4] National Information Standards Organization (NISO), Understanding metadata, ed., NISO Press, 2004.
[5] Dublin Core Metadata Initiative (DCMI), Dublin Core Metadata Element Set Version 1.1: Reference Description, http://dublincore.org/.

[6] The Library of Congress, Marc Standards, http://www.loc.gov/marc/.
[7] National Information Standards Organization and AIIM International, Data Dictionary – Technical Metadata for Digital Still Images, Released as a Draft Standard for Trial Use June 1, 2002–December 31, 2003.
[8] R. Heery and M. Patel, Application profiles: mixing and matching metadata schemas, *Ariadne issue* **25** (2000).
[9] Dublin Core Metadata Initiative (DCMI), Dublin Core Collection Description Application Profile, 2006, http://dublincore.org/groups/collections/collection-application-profile/2006-02-24.
[10] International Council on Archives (ICA), International Standard Archival Description ISAD(G), 2nd ed., 2000.
[11] Alexandria Digital Library, The ADL Collection Metadata DTD, http://www.alexandria.ucsb.edu/middleware/dtds/ADL-collection-metadata.dtd.
[12] Research Support Libraries Program (RSLP), Collection Description Schema, http://www.ukoln.ac.uk/metadata/rslp/.
[13] Computer Society/Learning Technology Standards Committee, IEEE Standard for Learning Object Metadata, ed., IEEE Standards Department, 2002.
[14] W3C, Date and Time Formats, http://www.w3.org/TR/NOTE-datetime.
[15] The Library of Congress, List MARC Code Lists for Relators, Sources, Description Conventions, http://www.loc.gov/marc/relators/.
[16] L. Dempsey, Scientific, Industrial, and Cultural Heritage: a shared approach, *Ariadne issue* **22** (1999).