Implementing Digital Folklore Collections

Irene Lourdi

Department of Archive and Library Sciences, Ionian University, Corfu, Greece

Libraries Computer Centre, University of Athens, Athens, Greece
elourdi@lib.uoa.gr

Mara Nikolaidou

Libraries Computer Centre, University of Athens, Athens, Greece
mara@di.uoa.gr

Christos Papatheodorou

Department of Archive and Library Sciences, Ionian University, Corfu, Greece
papatheodor@ionio.gr

Abstract

In this paper, we present a metadata model 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. The volume and the variety of material results to collection representation as a hierarchical structure, according to the type of objects, the corresponding chronological period and geographic region. Our goal was to preserve and popularize to every user all the precious information regarding collection material. For this purpose we develop a metadata model that enables efficient navigation to the notebooks sub-collection structures, as well as meaningful information retrieval to the collection objects.

1. Introduction

Folklore collections are valuable sources for study and research about 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, one could digitize them and made them accessible through the web. University of Athens has initiated a 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 volume and the variety of material results to collection representation as a hierarchical structure, according to the type of objects, the corresponding chronological period and geographic region. The main difficulty for managing such collections is the heterogeneity of the material (handwritten texts, photographs, 3D objects, sound recordings, maps) that requires the application of different digitization, description and maintenance practices. Furthermore, 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 collections.

In such cases, it is required: 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 on the definition of a description model for the collection level (Lourdi, 2004) and focuses on the description of notebooks sub-collection. Our goal was to preserve and popularize to every user all the precious information regarding collection material. For this purpose we develop a metadata model that enables efficient navigation to the notebooks sub-collection structures, as well as meaningful information retrieval to the collection objects. In the next section, we provide a short description of the notebook sub-collection. In section 3, the metadata model introduced to describe collection material is presented. We emphasize on the representation of object relationships and related constraints. Conclusions relay in section 4.

2. Notebook Sub-collection Description

Collection Structure

The travelling notebooks sub-collection is a very good sample of a collection with complexity and heterogeneity because it contains a variety of material. The basic physical component of the collection is the notebook that it has been written by every student of the Folklore Department of the University with the intention to write down in detail the cultural features of a place/ village of Greece and to keep that information for the future. The size of the collection is quite big, about 4000 written essays and 2 million pages that cover almost the whole country. Besides the text (handwritten or not) the notebook consists of photographs or pictures and some small objects that have been stuck on the pages by the students on purpose, in order to make the content of the page or of the chapter more expressive and more valid. Also it must be noticed that the structure of the notebooks follows a standard questionnaire prepared by the Department professors. In the current situation, the notebooks have not been catalogued and have not been registered to any system, which means that the 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 easy retrievable, the sub-collection of notebooks has been separated in several levels that will be treated as separate digital objects with their own characteristics and their own description. These logical entities are the levels of description, covered by the metadata model, and they follow an hierarchical structure as it is given in figure 1.
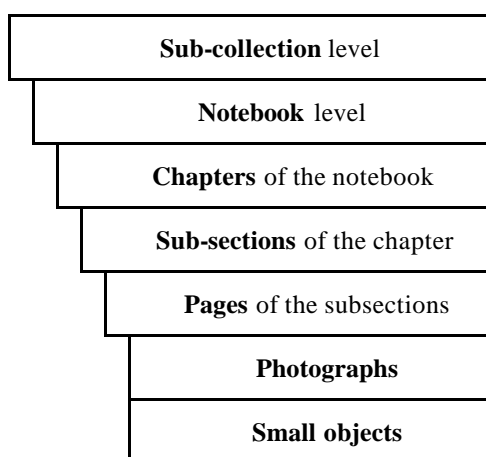
Sub-collection level
Notebook level
Chapters of the notebook
Sub-sections of the chapter
Pages of the subsections
Photographs
Small objects

Figure 1.Description levels of Notebook Sub-collection

The representation of the notebooks on the web depends quite on the structure of the collection, hence and the metadata model has to ascribe to each level all the material attributes. Therefore it has to combine elements from various metadata standards. Thus, it is required: i) to express the subject coverage of the resources and the geographic region that covers ii) to express clearly the relationships that exist between the digital objects through all the levels iii) to contain elements concerning description, administration and elements that will show the stable structure of the notebooks iv) to be characterized by the policy of "inheritance" (transferring attributes) from the parent to the children and sometimes the opposite way (as it will be explained later) and v) to contain elements concerning rights in order to protect the copyrights of the oral tradition. (Cole,2001)

The description of the resources needs to be done by a metadata model that will depict the existing structure of the collection in order to express further the distinctiveness as well as the relations between the objects and will give a strict policy of how it will be implemented in the local system of digital library. The digital folklore collections are valuable source for studying the cultural heritage of a country, so the data model besides structure must also express the semantic definition of folklore material.

What we need is a metadata model that will make the digital collection functional for the users and that 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 keywords. In order to achieve our target we planned a metadata model that correlates to every level of the collection and defines which are the adequate elements to give a full and rich description of the content and simultaneously by making easier for the cataloguer to fill these elements in every level.

3. Proposed Metadata Scheme

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

i. The stable structure, that characterizes the notebooks.

ii. Our target to adjust the digital collection to the users' needs. The basic intention is to provide "user centered" retrieval of information.

iii. Retrieval of information by making queries and receiving responses from all the defined structural levels of the collection.

iv. The type and the characteristics of the natural objects affect very much the description of the corresponding digital objects. So in our case it is necessary to combine different metadata standards in order to cover both cases (physical and digital versions).

Categories of metadata standards

According to the paper of NISO "Understanding metadata" (NISO guide) metadata elements are separated in the following categories: descriptive metadata, structural metadata and administrative metadata. More specifically, "descriptive metadata" are responsible for the description of the resources 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 organized. They are very important when it is about a compound object with complex structure and multiple levels. Finally, the "administrative metadata" give information about the administration of the resources and technical information about when and how the described resource has been created.

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	
PHOTO	Descriptive	
	Administrative for physical photograph	Administrative for digital versions of the hotograph
	Structural	
OBJECTS	Descriptive	
	Administrative	
	Structural	

Figure 2. Metadata Categories

According to these categories, we have made a hierarchical picture of the metadata elements that will be preserved for every logical entity of our collection (fig. 2)

The metadata model

The proposed model for the description of collection of notebooks combines elements from a variety of metadata schemes to describe many thematically interlinked sub-collections with compound objects. The model is mostly based on the Dublin Core Metadata Initiative for both collection-level description and for item-level description. In order to cover the requirements we set above, we have extended Dublin Core with some further local elements or we have enriched it by using elements taken from other metadata standards, that are suitable for types of resources.

More specifically, we have used Marc (4) for describing characteristics of the physical objects and NISO "technical data for still images" (5) to give technical information about the scanning of the notebooks, the images and the small objects that they are inside them. The model for the collection description as an entity is based on the Dublin Core Collection Description Application Profile (6) and is enriched with elements from other metadata standards for collection description like: ISAD (7), the metadata model of Alexandria Digital Library (ADL) (8), RSLP (9) and IEEE-Learning Object Metadata (LOM)(10).

In 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 is functioning. The elements are separated in the categories, we described in figure2, and according to the nature of the described resource (physical or digital). We believe that it is necessary to keep information both for the physical and for the digital version as well, because the characteristics of the physical item affect also the digital ones. In the model all the elements are optional except from someone that are indicated as mandatory.

Also there are some indications that express attributes of each element. The indications show: i) from which metadata standard the element comes from (**DC**=Dublin Core, **Marc**, **L**=local (made for our project) ii) if the element is mandatory to be filled in order to continue with the description (**M**=mandatory) and iii) the elements that are proposed to be filled automatically by the system from values that exist in lower levels (**I**=inherit).

NOTEBOOK	
DESCRIPTIVE METADATA	
DC_ TITLE (M)	DC_COVERAGE SPATIAL
DC_SUBTITLE	COVERAGE_SPATIAL_ SPECIFICATION (L)
DC_CREATOR (M)	DC_COVERAGE_ TEMPORAL (I)
DC_CONTRIBUTOR (ROLE)	COVERAGE_SPATIAL_ ADDITIONAL INFO (L)
DC_DATE_CREATED	DC_SUBJECT (I)
DC_DESCRIPTION_ ABSTRACT	SUBJECT_ CLASSIFICATION (L)
ADMINISTRATIVE METADATA for Physical entity	
BINDING INFORMATION (MARC)	FORMAT_DIMENSIONS (MARC)
DC_IDENTIFIER (M)	DC FORMAT_EXTENT (I)
DC_SOURCE	
ADMINISTRATIVE METADATA for Digital entity	
LOCATION_DIGITAL (L)	DC_FORMAT_ EXTENT (I)
DC_DATE_CREATED (M)	DC_FORMAT_MEDIUM
OTHER PHYSICAL DETAILS(L)	DC_PUBLISHER
DC_DATE AVAILABLE	
STRUCTURAL METADATA	
ORGANIZATION AND ARRANGEMENT OF MATERIAL (MARC)	DC_DESCRIPTION_ CONTENTS (I)
DC_RELATION (IS PART OF)	

Figure 3: Notebook Entity Metadata

It is important that the template of every logical entity some metadata elements are proposed to be inherited to the next level, in order not to be filled again. For example the elements that are proposed to pass from the notebook to the chapters are: "Coverage_spatial" and "date_created". Also from the chapter to the subsections is proposed to be inherited the element of "coverage_spatial" with its value.

Further, the elements of "subject" and "format extent" in notebooks are proposed to be filled automatically by taking values from the templates of the chapters and the element and "coverage_temporal" also by taking the values of the same element from the templates of the chapters. The same goes for the element "description_contents" that is filled automatically with the values from the element

of "title" from the chapters. By this way we have a full list of the contents without the cataloguer being responsible to write them down. The same is for the contents of the chapters that are coming from the titles of the subsections. The element "format_extent" in chapter is proposed to be filled by adding the number of the pages that belong to the specific chapter.

CHAPTER	
DESCRIPTIVE METADATA	
DC_TITLE (M)	DC_DESCRIPTION_ ABSTRACT
DC_COVERAGE_TEMPORAL(I)	
ADMINISTRATIVE METADATA	
DC_FORMAT_EXTENT(I)	DC_IDENTIFIER (M)
TECHNICAL METADATA	
DC_DESCRIPTION_ CONTENTS (I)	DC_RELATION_ (is a chapter of the notebook…)

Figure 4: Chapter Entity Metadata

SUBSECTION	
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_RELATION_HAS PHOTOGRAPH"/"OBJECT"
DC_DESCRIPTION_CONTENTS (I)	

Figure 5: Subsection Entity Metadata

PAGE	
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)

Figure 6: Page Entity Metadata

In general the model has been designed in a way that the templates of the logical entities are affected by each other and can be internally functional based on conditions that we have set, like the following ones. Our intention is by

making advantage the possibilities of Digital Library system to gain time and to make the session of describing the material easy and effective.

Metadata Model Rules

Metadata model rules define the function and the 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].

2) The element of "DC_Description_contents" will be filled automatically by taking values from the element of "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 of 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 of "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.

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:

i. The element "DC_Subject" is proposed to be characterized more precisely, so we have set a local refinement: Subject_Classification, that corresponds to Marc21 080.

ii. The element "DC_Coverage_Spatial" is very important for the searching of our folklore collection. Especially in the collection of notebooks the places have 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.

iii. 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 **(6)** Also it is proposed to give more information for the entities "DC_owner" and "DC_collector" using the attributes of "vcard namespace", as it is in RSLP metadata schema.

iv. In order to express exactly the relationships that exist between every logical entity of the collection (collection, notebook, chapter…) we have extended more 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…).

Functional Inheritance of Attributes

In order to define a minimum set of metadata elements that can describe the folklore collection of written notebooks as a whole entity and each notebook as a set of different structural levels. The big size of the collection (it contains almost 4000 notebooks) and the fact that the notebooks have such a complex structure (text, photographs and objects) makes even more difficult the work of the cataloguer to describe all the resources with all the details that the nature and the content of the collection requires.

For that reason, it has been decided that the system of the Digital Library must support the policy of inheritance of the attributes from one level to another. It has been defined in the

system that many metadata elements of the metadata model, we described above, 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 cataloguers to describe all the resources fully.

Except from of the policy of transferring elements and their values from upper to lower levels, we have also set to be happening the opposite. Some elements are filled automatically by adding and taking values from lower levels (as it has already been said in chapter…). In order to get in function this kind of policy, are required some "expressive tools" that will implement the inheritance of the attributes from one level to another (up or down).

Information Retrieval

The proposed model focuses both on the notebooks collection and on the components of it, so that the user can access both the collection and every notebook separately. The user has the possibility to find information either by browsing the list of all the notebooks and by looking the table of contents of each notebook or either by searching in the content of each digital entity by using keywords and values from specific lists of geographical places, subjects, persons and chronological periods. Further the system must allow 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: When a user searches for information about the customs of marriage in the *"Helateia"* village, he/she has direct access to specific chapters or the subsection of the notebooks dealing with this matter and any photos that fitting the same criteria. In its traditional form, the user might retrieve specific notebooks, but he/she has to search their content himself/herself.

4. Conclusions

Our scope was to define and implement a general metadata model that facilitates the retrieval of information of digital folklore collections consisting of heterogeneous resources. In its current state, the notebooks collection is not functional or easily accessible by users . Thus, we try to establish a model for affectively describing and administering large digital folklore collections. Our intention was to facilitate the users to easily retrieve all the information included within the notebooks and to comprehend the content and context of these resources.

5. References

1) Lourdi I. and Papatheodorou C., 2004 "A metadata application profile for collection-level description of digital folklore resources", *proceedings of PEH DEXA Workshop,* Spain, 2004.

2) COLE T., 2001 "Creating a Framework of guidance for building good digital collections", *journal First Monday* 7(5).

3) "Understanding metadata", NISO guide, http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

4) Marc Standards http://www.loc.gov/marc/ [viewed 01/03/2004]

5) NISO, *Draft Data Dictionary: Technical Metadata for Images*, Version 1.0, July 5, 2000

6) Dublin Core Collection Description Application Profile http://www.ukoln.ac.uk/metadata/dcmi/collection-application-profile/2003-08-25/[viewed 1/10/2004]

7) International Standard Archival Description ISAD(G) http://www.ica.org/biblio/cds/isad_g_2e.pdf

8) HILL, L. *et al.* Collection Metadata Solutions for Digital Library Applications, 1998 (ADL description)

9) RSLP Collection Description Schema http://www.ukoln.ac.uk/metadata/rslp/ [viewed 01/03/2004]

10) IEEE LOM 3.1 http://grouper.ieee.org/LTSC/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf

11) Marc Standards, "List of relator codes", http://www.loc.gov/marc/relators/relaterm.html