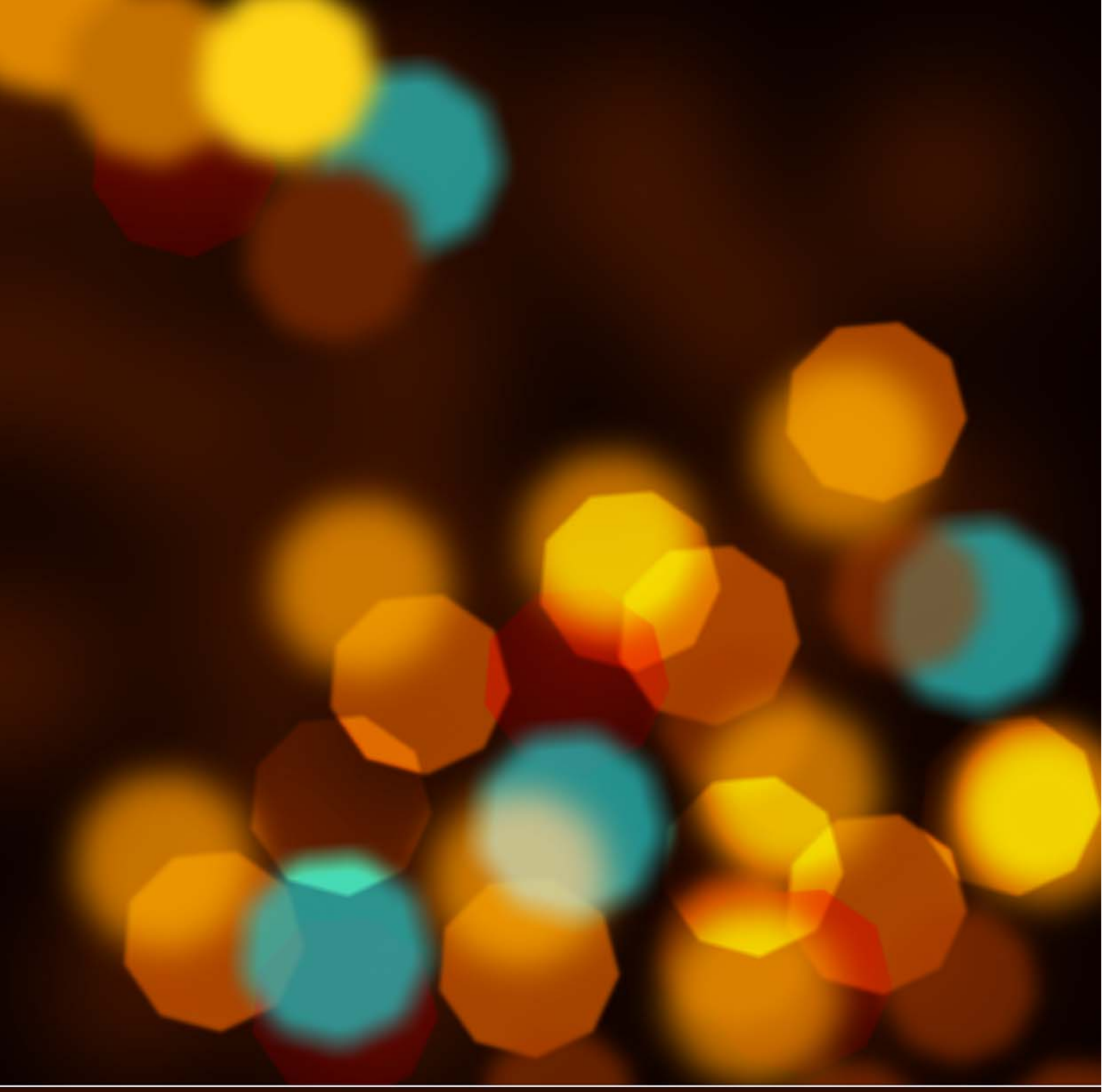

First Workshop on
Digital Information Management

Corfu, Greece, March 30-31, 2011

Corfu, 2011

Proceedings of the

First Workshop on Digital Information Management

March 30-31, 2011
Corfu, Greece

Workshop Information

Description: The workshop was organized by the Laboratory on Digital Libraries and Electronic Publishing, Department of Archives and Library Sciences, Ionian University, Greece and aimed to create a venue for unfolding research activity on the general field of Information Science. The workshop featured sessions for the dissemination of the research results of the Laboratory members, as well as tutorial sessions on interesting issues. It was addressed both to researchers and practitioners on the topics listed below, while it had been taken well care to hold a pedagogical aspect to attract the interested graduate students of the Department. The topics of the workshop included, but were not limited to:
- Information retrieval and digital libraries
- Semi-structured data management
- Information Integration
-- Metadata interoperability
-- Semantic interoperability
- Knowledge organization and management
-- Ontologies and conceptual modelling
- Social Networking in information contexts
- User studies
- Evaluation of Information Services and Systems

Workshop Chairs	**Program Committee**	**Website**
Christos Papatheodorou	Sarantos Kapidakis	Giannis Tsakonas
Manolis Gergatsoulis	Maria Monopoli	Lefteris Kalogeros
	Michalis Sfakakis	
	Spyros Veronikis	
	Giannis Tsakonas	

Additional Information

Webpage:
http://dlib.ionio.gr/workshop2011/

SlideShare webpage:
http://www.slideshare.net/event/first-workshop-on-digital-information-management

Table of Contents

Failed Queries: a Morpho-Syntactic Analysis Based on Transaction Log Files

Anna Mastora[1], Maria Monopoli[2] and Sarantos Kapidakis[1]
[1] Laboratory on Digital Libraries and Electronic Publishing, Department of Archives and Library Sciences, Ionian University, 72, Ioannou Theotoki str., GR-49100, Corfu, Greece.
[2] Library Section, Economic Research Department, Bank of Greece, 21, El. Venizelos Ave., Athens, GR-10250
{mastora, sarantos}@ionio.gr, mmonopoli@bankofgreece.gr

Abstract. The aim of the study is to elaborate on the procedure needed in order to analyze morpho-syntactically the typing-error queries submitted in Greek during the search process. In the context of our analysis a *failed query* is a query which returned no hits. The analysis showed that failed queries represent 36% of the submitted queries. More specifically, 19.6% of failed queries occurred due to typing errors. We discovered that for analyzing morpho-syntactically a Greek text corpus the PoS tools need to be rich in tags in order to work adequately. Open Xerox tokenizer performed well but with significant pre-processing of the queries and the analyzer seems to require additional tools to improve its performance. MS Word which was used for spelling corrections seems to perform satisfactorily. All tools were challenged in terms of named entities recognition.

Keywords: Failed queries, Morpho-syntactic analysis, PoS tagging, Typing errors

1 Introduction

Information retrieval techniques do not work effectively at all times. *Not working effectively* includes both not retrieving relevant documents, i.e. low recall, and retrieving non relevant documents, i.e. low precision. Part of studying what is not retrieved during an information search process is the analysis of *failed queries* or *failure analysis*. This is also the motivation of our study with respect to Natural Language Processing (NLP) techniques.

In this study we explore the failed queries caused due to typing errors. The grouped queries are analyzed morpho-syntactically in order to develop a clear image of the required process before stepping to the next phases of the data analysis in the future.

2 Aims and Objectives

The aim of the study is to elaborate on the procedure needed in order to analyze morpho-syntactically the typing-error queries submitted in Greek during the search process.

The objectives of the study are twofold. First, we explore the extent and types of failed queries due to typing errors. Second, we explore the procedure and feasibility of their morpho-syntactic analysis.

3 Related Research

The discussion concerning what constitutes a *failed query* is extensive [1, 2, 3] Different perspectives of search failures are presented. Some researchers consider *failure* in terms of precision and recall applying retrieval effectiveness measures. Others examine *failure* in terms of user satisfaction applying users' criteria to measure whether a query failed or not. Others use transaction log files and treat input terms either as "bag of words" or apply relevance feedback and assign more interpretations to the result set. Finally, there are techniques which study the human behavior by observation.

Significant interest has been expressed on failed queries as the outcome of subject searching [4, 5]. This strategy has been identified as the most common for delivering failed queries due to various reasons but mostly because of the inherent difficulty of matching the index terms to the users' queries. This identified difficulty and the documented analysis [6] which supports that for information needs related to environmental issues users tend to perform subject searching explain the focus of our study on subject searching.

A considerable aspect of the research on failed queries is the techniques used for Natural Language Processing. These techniques are essential especially in highly inflectional languages [7] such as the Greek language. While the main goal at all times is to assign the proper semantic information to each query, this cannot be accomplished without prior identification of the morho-syntactic information of the terms used. The techniques applied for this purpose are the Part of Speech (PoS) tagging which is accompanied by more detailed morpho-syntactic information (see Fig.2 for an example).

4 Definitions and Methodology

In this section we provide the definitions of the terminology used in our study as well as the analysis on the methodology used.

4.1 Definitions

Through the study of related research, as presented in the previous section, what becomes obvious is that failed queries constitute a disputable area concerning the very definition of what actually should be considered as a *failed query*.

In the context of our analysis a *failed query* is a query which returned no hits. We took into consideration the objections on the issue yet we support this decision by the fact that the analysis of the data was based on terms extracted from transaction log files without any relevance feedback from the users' perspective. This is also why we proceeded with a morpho-syntactic analysis leaving for later phases the processes related to word-sense disambiguation. An additional factor which strengthens our decision is that both the content of the database and the information needs belonged to the same domain and it was expected that most queries would return hits.

The *morpho-syntactic analysis* of the data is a cognitive process that constitutes an intermediate layer between morphological and syntactic analysis and aims to assign unambiguous morpho-syntactic information to words of texts [8].

The *morpho-syntactic information* consists of the morphological origin and the morpho-syntactic properties of a word. For example, the word *ανθρώπου* is the genitive singular form of the masculine noun *άνθρωπος* [8].

Inflectional languages are the languages with a high morpheme-per-word ratio whereas the *morpheme* is the smallest meaningful linguistic unit. The Greek language is considered a highly inflectional language.

More definitions on terminology used across this paper can be found in the corresponding sections.

4.2 Methodology

The data analyzed in this paper was gathered from an in vitro experiment with the participation of 27 undergraduate students at the Department of Archives and Library Sciences at the Ionian University in Corfu. They were given 13 information needs related to environmental issues and asked to submit appropriate queries in order to retrieve relevant documents. The database they were searching in contained material mainly from the environmental domain.

For the purpose of this experiment we selected and customized approximately 14,400 bibliographic records of the Evonymos Ecological Library[1]. The queries were submitted in Greek as well as the records contained information only in Greek. This is a significant factor when analyzing data in the context of Natural Language Processing because it eliminates the possibilities of arbitrarily assigning characteristics to words due to the intervening stage of their translation.

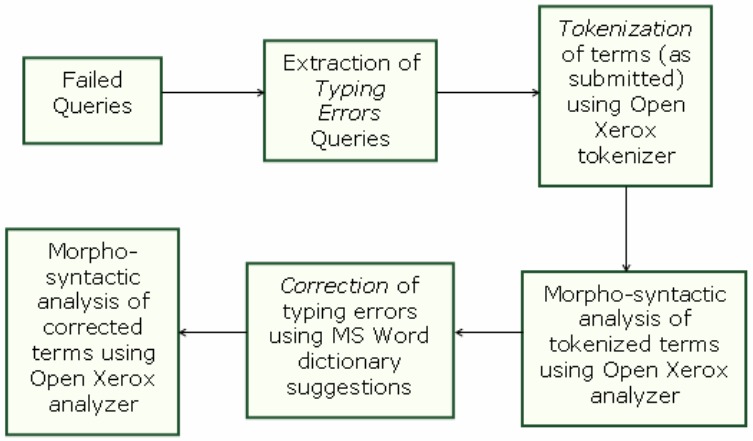

Fig. 1. Synopsis of the procedures' workflow during the processing of the data.

The participants could search only in the *Subject* field. According to Jones et al. [2] users rarely change default settings. This observation suggests that the customization of the interface did not record either an unrealistic or biased users' behavior. The transaction log files kept in a Zclient consist of one xml document per user per session. All participants logged in the system using their matriculation numbers thus making it easier to potentially relocate them for providing feedback at a later stage of the research.

Concerning the processing of the data, the first step involved the selection of *failed queries* and, more specifically, the selection of *typing error queries*. The next step involved the tokenization of the selected corpus of queries and then their morpho-syntactic analysis. Following was the processing of correcting the spelling errors of the tokens and running from scratch the analyzer. Figure 1 above visualizes the workflow of the data processing while Figure 2 below gives an example of the processed data.

[1] Full database available at http://www.evonymos.org/index.html (last accessed 17 April 2011).

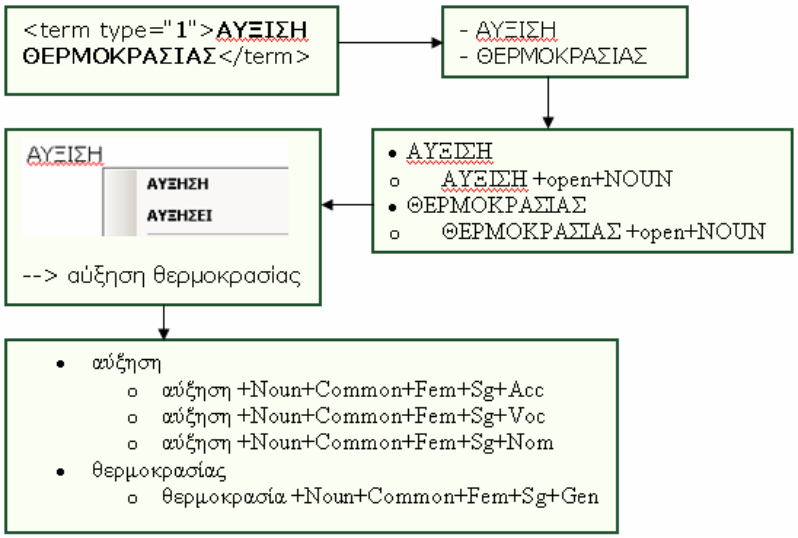

Fig. 2. Example of the processing of the data.

5 Results

This section presents the findings of our study. There were 1,284 queries submitted overall, while 459 of them were failed queries, i.e. 36%, meaning that they returned no hits. Consistent to our previous work [9], we further categorized those failed queries to four subcategories, namely *Valid terms with no hits*, *Typing errors*, *Inseparable terms* and *Undefined terms.*

The failed queries subcategory named *Valid terms with no hits* is the most populated one with a ratio of 75.8% and includes terms which were valid both morphologically and syntactically yet they did not deliver any hits. The second subcategory, i.e. *Typing errors*, comes next in delivering failed queries with a ratio of 19.6%. Third appears the subcategory containing the words which were not separated during typing. They represent a ratio of 2.4%. And, finally, the last subcategory includes some undefined terms, meaning words that do not appear in official dictionaries, in 2.2% of the failed queries overall.

Since the focus of this study is on *Typing error* queries, we analyzed them further by dividing them, based on previous work [9], to five new subcategories, namely *Substitutions*, *Transpositions*, *Omissions*, *Insertions* and *Divisions*. *Substitutions* include the changing of a letter with another letter, like in the case of typing ***φ***εωθερμική instead of the correct ***γ***εωθερμική. *Transpositions* include the cases where one or more characters within a word do not appear in the right order, for example *αν****ιτ***ρρήσεις instead of the correct *αντ****ι***ρρήσεις. *Omissions* include the cases where one or more characters within a word are missing, for example *οπω****φ***όρα δέντρα instead of the correct *οπω****ρο***φόρα δέντρα. *Insertions* include the cases where one or more characters are added within a word, as in the case of *μεσσ*όγειος instead of the correct *με****σ***όγειος. Finally, the last subcategory of typing error queries is *Divisions*, including splitting terms which should appear as one. Table 1 right below shows the distribution percentage of each subcategory.

Table 1. Categorization of failed queries due to typing errors (percentage, %).

Substitutions	Transpositions	Omissions	Insertions	Divisions
36.7	4.4	28.9	28.9	1.1

At this point we remind that the total number of failed queries was 459 out of 1,284 submitted queries. Ninety (90) of the failed queries were due to typing errors. In order to proceed to the morpho-syntactic analysis we had to identify the tokens to analyze. For this purpose we uploaded the terms to the Open Xerox Tokenizer[2]. The outcome of this process was 156 tokens.

The following step was to explore whether the Open Xerox analyzer[3] would directly identify the misspelled tokens during the morpho-syntactic analysis. As shown in Table 2, the tool did not manage to recognize the misspelled tokens, thus, performing poorly since it only managed to identify 20 out of 156 tokens.

Table 2. Categorization of identified tokens when analyzed as submitted (exact numbers).

Regular words	Punctuation	Pronouns	Prepositions	Others
10	5	3	1	1

In order to overcome the barrier of this poor performance we proceeded with the correction of the identified tokens using the spelling suggestions of the MS Word's default dictionary. During this stage, since the data was processed manually, we interfered with the results by assigning the semantically correct suggestion to each token. Table 3 below shows the performance of the MS Word dictionary.

Table 3. Categorization of MS Word correction suggestions.

Action	**Percentage (%)**	**Actual number**
No suggestion required	30.1	47
No suggestion provided	12.8	20
Irrelevant suggestion	3.2	5
MS Word's 1st suggestion=correct	45.5	71
MS Word's 2nd suggestion=correct	7.1	11
MS Word's 3rd suggestion=correct	1.3	2
Total	100	156

As shown in Table 3, for approximately 30% of the cases no suggestion was required. This includes the tokens which did not contain any typing error. Their assignment to typing error queries was due to the fact that they belonged to multi-word terms in which at least one typing error was identified. After the tokenization stage, these tokens were isolated from the original term and when processed during the next stage, that is the stage of typing errors' correction, no intervention was required. Punctuation was also included in this category.

After having corrected the originally identified tokens, we proceeded with the morpho-syntactic analyzer anew. This time it performed significantly better identifying 139 out of 156 tokens. Table 4 below shows a categorization of the missed identifications. We need to mention at this point that in the documentation for the Part of Speech tag set for Greek it is mentioned that the analyzer identifies words in other languages and tags them as *+FM*, i.e. Foreign Words[4]. We observed an inconsistency concerning this feature since words in English included in our corpus were not identified as expected. Instead they were rather arbitrarily assigned a general tag, like *noun*.

[2] Available at http://open.xerox.com/Services/fst-nlp-tools/Consume/175 (last accessed 17 April 2011).

[3] Available at http://open.xerox.com/Services/fst-nlp-tools/Consume/176 (last accessed 17 April 2011).

[4] The full Part of Speech (PoS) tag set for Greek is available here http://open.xerox.com/Services/fst-nlp-tools/Pages/Greek%20Part-of-Speech%20Tagset (last accessed 17 April 2011).

Table 4. Categorization of corrected tokens not recognized during the morpho-syntactic analysis.

Category of the token	Percentage (%)	Actual number
Named entities	17.6	3
Regular words	35.3	6
Truncated words	29.4	5
English words	11.8	2
Punctuation	5.9	1
Total	100	17

6 Conclusions

The analysis of *failed queries* shows that they represent 36% of the submitted queries overall in our experiment. More specifically, 19.6% of failed queries are due to typing errors. During Natural Language Processing the queries which contain typing errors require more steps and extra mechanisms involved in order to achieve a trustworthy and effective morpho-syntactic analysis. This is both a practical and a substantial problem to solve considering their proportion within the overall submitted queries.

In the process of data analysis we discovered that the tools for morpho-syntactic analysis for the Greek language need to be rich in tags in order to work adequately. Since the Greek language is a highly inflectional language it requires the combination of more mechanisms, such as dictionaries, discovering synsets etc., for proper analysis. This practice affects the complexity of the tools used but it seems inevitable. Such tools should aim at making the less possible suggestions for each segment and that the suggestion is as close as it gets to the *true* sense of the segment, where by *true* is meant the sense which the user intended.

Transaction log files serve as good starting points for processing the data quantitatively but more measures need to be applied in order to extract adequate qualitative information for the terms used in submitted queries.

Concerning the tools we used for the analysis of our data we observed important deficiencies which complicated the process. First, we observed that in order for the Open Xerox tokenizer to work properly all input words should be lower case and stress marked. This caused extra load of work because we had to convert the words submitted in capitalized form and stress them. Additionally, we had to implement this step to all the words that were originally in lower case but had no stress mark as well.

Another challenge of the tools used was that they did not recognize *named entities*. This covers a whole separate field of research but within our dataset the use of named entities was not extensive and did not severely affect the outcome. In other cases, however, this could play a significant role.

7 Future Work

Future planning concerning this work includes research on *named entities recognition, language identification* and *word-sense disambiguation* in order to achieve higher rates of morpho-syntactic analysis. All three aforementioned areas are important in terms of analyzing the input of the user and delivering better results.

Another aspect of future research on this area is the exploration of how and to what extent could we incorporate Knowledge Organization Systems (KOS) to query expansion techniques in terms of improving the retrieved result set in cases of prior failed queries.

Acknowledgement: This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

References

1. Tonta, Y., 1992. Analysis of Search Failures in Document Retrieval Systems: A Review. *Public-Access Computer Systems Review*, 3(1), pp. 4-53.
2. Jones, S. et al., 2000. A Transaction Log Analysis of a Digital Library. *International Journal on Digital Libraries*, 3, pp. 152-169.
3. Pu, H.-T., 2008. An analysis of failed queries for web image retrieval. *Journal of Information Science*, 34(3), p.275–289.
4. Lau, E.P. & Goh, D.H.-L., 2006. In search of query patterns: a case study of a university OPAC. *Information Processing and Management: an International Journal*, 42(5), pp. 1316–1329.
5. Villén-Rueda, L. et al. 2007. The Use of OPAC in a Large Academic Library: A Transactional Log Analysis Study of Subject Searching. *The Journal of Academic Librarianship*, 33(3), pp. 327-337.
6. Nicholas, D. et al., 2008. User diversity: as demonstrated by deep log analysis. *The Electronic Library*, 26(1), pp. 21-38.
7. Acedański, S., 2010. A morphosyntactic Brill Tagger for inflectional languages. In *Proceedings of the 7th international conference on Advances in natural language processing*. IceTAL'10. Berlin, Heidelberg: Springer-Verlag, pp. 3–14.
8. Orphanos, G., 2000. *Computational morphosyntactic analysis of modern Greek*. Unpublished PhD thesis. Patras: University of Patras. School of engineering. Department of computer engineering and Informatics.
9. Mastora, A. et al., 2007. Exploring users' online search behaviour: a preliminary study in a library collection, 2nd DELOS Conference on Digital Libraries, Pisa, Italy, December 5-7.

Mapping Encoded Archival Description to CIDOC CRM

Lina Bountouri and Manolis Gergatsoulis

Database & Information Systems Group (DBIS),
Laboratory on Digital Libraries and Electronic Publishing,
Department of Archives and Library Science, Ionian University,
Ioannou Theotoki 72, 49100 Corfu, Greece.
{boudouri, manolis}@ionio.gr

Abstract. In this paper we analyze the semantics of the archival description, expressed through the Encoded Archival Description. Through this analysis it is concluded that an EAD document is a hierarchy of documentation elements and attributes and that through this documentation the archive is semantically expressed through three different hierarchies (hierarchy of physical objects, hierarchy of information objects and hierarchy of linguistic objects). The semantic views of the archive as well as their interrelationships are mapped to the CIDOC CRM.
Keywords: Metadata interoperability, Encoded Archival Description, Ontologies, CIDOC CRM, Mappings.

1 Introduction

Cultural heritage institutions, archives, libraries, and museums host and develop various collections with heterogeneous types of material, often described by different metadata schemas. Managing metadata as an integrated set of objects is vital for information retrieval and (meta)data exchange. To achieve this, *interoperability* techniques have been applied. One of the widely approved and implemented techniques is the *Ontology-Based Integration*. Ontologies play a vital role in semantic interoperability and integration scenarios, and they are preferred in regard to other schemas, due to their ability to conceptualize particular domains of interest and express their rich semantics in a formal manner. One of their main roles in an interoperability scenario is to act as the *mediated schema* between heterogeneous information systems [14, 4].

This paper builds upon an ontology-based metadata integration architecture, which considers the *CIDOC Conceptual Reference Model* (*CIDOC CRM*) ontology [3] as the *mediator*. The proposed architecture considers a set of data sources each of them providing information encoded by a different metadata schema (e.g. EAD, VRA, DC, MODS, etc). Each schema is semantically mapped to the CIDOC CRM based mediator, which may also retain its own database of metadata encoded in CIDOC CRM format. Various integration scenarios can be built on this architecture.

The main research result of this paper is the mapping of the *Encoded Archival Description* (*EAD*) [9] to CIDOC CRM. In order to create this mapping, we firstly analyzed the main concepts of the archive and of its components parts, as well as the main concepts of the archival description, which are the hierarchical structure and the inheritance of information between the hierarchical levels of description. These concepts (being expressed through EAD) have to be mapped to the ontology so as to promote the semantic integration. Part of the mapping procedure was to properly define these highly complex semantic structures in order to be expressed by the CIDOC CRM. Furthermore, the EAD descriptive fields must be also mapped to the ontology. This research work is the first complete effort to define the semantic mappings of the EAD to the CIDOC CRM.

2 Preliminaries

2.1 CIDOC Conceptual Reference Model

The *CIDOC CRM* is a core ontology, which consists of a hierarchy of 86 *entities* (or *classes*) and 137 *properties*. A *class* (also called *entity*) groups items (called *class instances*) that share one or more common characteristics. A class may be the *domain* or the *range* of *properties*, which are binary relations between classes. An *instance of a property* is a relation between an instance of its domain and an instance of its range. A property can be interpreted in both directions (active and passive voice), with two distinct but related interpretations. A *subclass* is a class that specializes another class (its *superclass*). A class may have one or more immediate superclasses. When a class A is a subclass of a class B then all the instances of A are also instances of B. A subclass inherits the properties declared on its superclasses without exception (*strict inheritance*) in addition to having none, one or more properties of its own.

A *subproperty* is a property that specializes another property (its *superproperty*). If a property P is a subproperty of a property Q then a) all instances of P are also instances of Q, b) the domain of P is the same or a subclass of the domain of Q, and c) the range of P is the same or a subclass of the range of Q. Some properties are associated with an additional property (called *property of property*) whose domain contains the property instances and whose range is the class E55 Type. Properties of properties are used to specialize the meaning of their parent properties. A sample of CIDOC CRM properties is shown in Table 1.

CIDOC CRM expresses semantics as a sequence of path(s) of the form entity-property-entity. It is an event-based model and its main notions are the temporal entities. As a consequence, the presence of CIDOC CRM entities, such as actors, dates, places and objects, implies their participation to an event or activity [11].

2.2 Encoded Archival Description

The *archival description* documents the *archive*, which is a complex set of materials sharing common provenance, regardless of form or medium. The description involves a hierarchical and progressive documentation, beginning from the

Property id & Name	Entity - Domain	Entity - Range
P1 is identified by (identifies)	E1 CRM Entity	E41 Appellation
P2 has type (is type of)	E1 CRM Entity	E55 Type
P14 carried out by (performed)	E7 Activity	E39 Actor
P67 refers to (is referred to by)	E89 Propositional Object	E1 CRM Entity
P70 documents (is documented in)	E31 Document	E1 CRM Entity
P71 lists (is listed in)	E32 Authority Document	E55 Type
P102 has title (is title of)	E71 Man-Made Thing	E35 Title
P106 is composed of (forms part of)	E90 Symbolic Object	E90 Symbolic Object
P108 has produced (was produced by)	E12 Production	E24 Physical Man-Made Thing
P128 carries (is carried by)	E24 Physical Man-Made Thing	E73 Information Object

Table 1. A sample of CIDOC CRM properties.

archive, and proceeding with its sub-components, the sub-components of sub-components, and so on, often reaching the item level (e.g. a map). In parallel, it supports the inheritance of information between the hierarchical levels. *Finding aids* materialize archival descriptions and the *EAD* [9, 8] is the most widely used schema that supports the creation of electronic finding aids. An *EAD document*, starting from the ead root element, consists of three elements: the *EAD Header* (eadheader), which is the mandatory element including the metadata for the EAD document, the *Front Matter* (frontmatter), which carries optional information for the printed finding aid (if any), and the mandatory *Archival Description* (archdesc), which provides information on the archive's content and context of creation, such as:

- core identification information (incorporated in the did element), e.g. the archive's creator (origination) and title (unittitle),
- administrative and supplemental information that facilitate the use of the archival materials, such as the biography or history (bioghist), and
- description of the components, bundled in a wrapper element called dsc that encodes the hierarchical groupings of the archival components being described. An archival component is an easily recognizable archival entity, characterized by an attribute level as series, subseries, file, item etc, and it may be in any level within the hierarchical structure of the description. Components are deployed as nested elements, called either c or c01 to c12.

Example 1 presents an archival description on the level of fonds. Basic descriptive identification information for the archive, such as the title (unittitle), the creation date (unitdate), the identifier of the archive (unitid) and its creator (origination), is given inside the did element. Administrative and supplemental information is provided through the bioghist and controlaccess elements. Description of subordinate components is presented inside the dsc element, where two components are provided through c01 elements (both on the level of series) and include basic identification information, such as unittitle, unitdate, etc.

Example 1. In this example a fragment of an EAD document is presented:

```
<ead>
 <eadheader>...</eadheader>
 <archdesc level="fonds">
  <did>
    <unitid countrycode="GR" repositorycode="IU">ARC.14</unitid>
    <unittitle>Ionian University Archive</unittitle>
    <unitdate>1984 - 2007</unitdate>
    <origination>
      <corpname>Ionian University</corpname>
    </origination>
  </did>
  <bioghist>
    <p>The Ionian University was founded in 1984...</p>
  </bioghist>
  <controlaccess>
    <corpname>Ionian University</corpname>
  </controlaccess>
  <dsc>
   <c01 level="series">
    <did>
      <unitid countrycode="GR" repositorycode="IU">ARC.14/1</unitid>
      <unittitle>R. C. Archives</unittitle>
      <unitdate>1998 - 2007</unitdate>
      <origination>
        <corpname>I. U. Research Committee</corpname>
      </origination>
    </did>
   </c01>
   <c01 level="series">
    <did>
      <unitid countrycode="GR" repositorycode="IU">ARC.14/2</unitid>
      <unittitle>I. U. Library Archives</unittitle>
      <unitdate>1998 - 2000</unitdate>
      <origination>
        <corpname>I. U. Library</corpname>
      </origination>
    </did>
   </c01>
  </dsc>
 </archdesc>
</ead>
```

3 The archive and the archival description: the main concepts

According to [6] "*an ontology is a specification of a conceptualization*". More specifically, the CIDOC CRM ontology is the specification of the Cultural Heritage conceptualization. Based on that fact, a necessary step that must be taken

before the mapping of a metadata schema to a domain ontology is to capture its concepts, aiming to map them to the ontology. In general terms, the concepts of a metadata schema are related to:

- the semantics of the description (in this case, the semantics of the archival description),
- the semantics of the information resource they describe (in this case, the semantics of the archive), and
- the semantics of its descriptive fields (in this case, the descriptive fields - elements and attributes - of the EAD).

The main semantic concepts of an archive, expressed through its description, are [13]:

- the archive is a *physical object* that acts as evidence for the functions/activities of the human or of the corporate body that created it, and
- the archive is an *information object* that includes information in different *formats* and *languages*.

The basic characteristic of the archive and of the archival description is the hierarchical and multilevel tree-based structure including also the principal of inheritance of information. An archive usually consists of a large number of components, which form the hierarchical relationship from the upper level of description (e.g. the archive) to the lower levels of description (e.g. the subfonds, the series, the files etc).

As far as the hierarchical structure is concerned, since an archive follows it, its semantic concepts are also expressed through this structure. As a result, an archive as a set of physical objects may contain one or more subfonds, which are also a set of physical objects and they may also contain one or more series, which are also a set of physical objects. In parallel, an archive as a set of information objects consists of one or more information objects, for instance the subfonds, which in turn consists of one or more information objects, such as the series etc.

The archival description is expressed in a machine readable way through the EAD. The EAD includes - apart from the archival description - the metadata of the EAD document and of the archival description. To express this documentation, an EAD document is structured as a tree having as root the element ead, which includes three subelements: the eadheader, the frontmatter and the archdesc.

Analytically, the root element ead includes the whole EAD document. The element eadheader includes the metadata of the machine readable archival description and the element frontmatter includes information for the creation, publication and use of the finding aid. Finally, the archdesc element includes the description of the archive and of its components (c01-c12 and c) defining also the hierarchical and multilevel tree-based structure, according to Figure 1.

In this figure, an illustrative structure of an archive is expressed through the EAD and in particular through the archdesc and its subelements c01-c05 for the components. Note that the description of the archive is expressed though

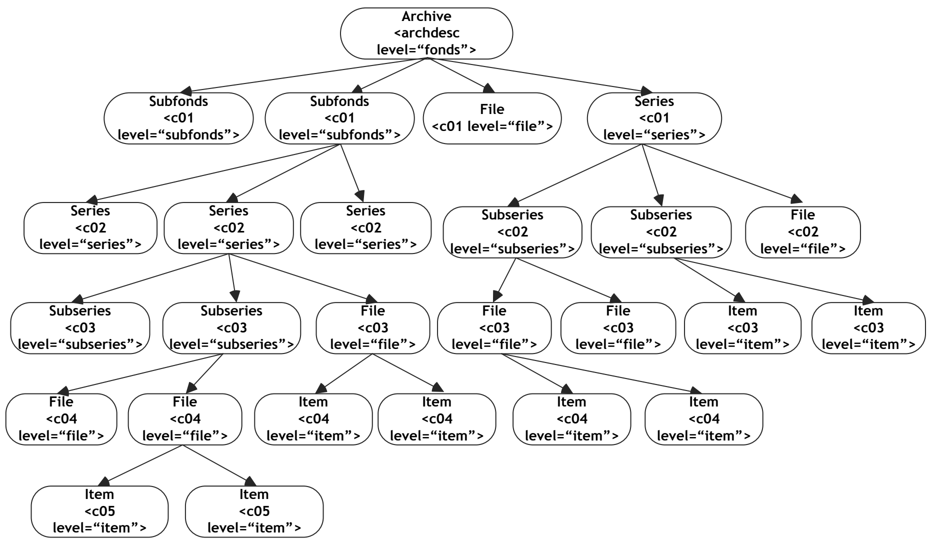

Fig. 1. The illustrative structure of an archive expressed through EAD.

the element archdesc declaring also the level of description which is the fonds (level="fonds"). The first level components (being one level lower than the archive) are expressed through the element c01, defining also the level of description for every archival entity that each c01 represents, for instance the subfonds (level="subfonds"), the series (level="series") and the file (level="file"). Lower levels may follow.

It is important to notice that for every archival entity various XML elements and attributes are implemented as the descriptive fields, in which archivists can provide all the necessary information for the archive and its components.

Consequently, in order to define the semantic mapping of the EAD to the CIDOC CRM, the following concepts must be mapped:

- the tree-based hierarchical structure of the archive and of the archival description, which is expressed through the archdesc, c01-c12 and c elements, and the inheritance property of the archival description,
- the semantic views of the archive, and
- the descriptive fields, which are expressed through the XML subelements and attributes of the archdesc, c01-c12 and c elements.

In this paper, emphasis is given to the mapping of the subelements' and attributes' semantics for the archdesc, c01-c12 and c elements, given that they encode the documentation of the archive.

4 The archive and the archival description: the mapping of the main concepts

4.1 The EAD document as a hierarchy of documentation elements and attributes

As already mentioned, the ead root element includes the whole documentation of the EAD document. The documentation concept is expressed in CIDOC CRM through the class E31 Document, which includes instances that are immaterial objects defining and documenting the reality, such as the sentences of a text, the databases etc. As a result, the ead element is mapped to this class, creating and mapping the whole EAD document to an instance of this class.

Respectively, the eadheader, frontmatter and archdesc elements are also mapped to instances of the class E31 Document, since: a) the eadheader semantically includes the documentation of the machine readable archival description, b) the frontmatter includes the documentation of the printed finding aid, and c) the archdesc includes the documentation of the archive. Provided that the c01-c12 and c elements "carry" the documentation of the archival components, they are also mapped to instances of the E31 Document class.

The aforementioned instances of the E31 Document class express the semantics of the main EAD elements that form the basic structure of an EAD document. What is more, the archdesc, c01-c12 and c elements express at the same time the structure of the archival description, which is one of the main archival characteristics that must be mapped to the ontology. The hierarchical structure between the instances of the E31 Document class representing the ead and the eadheader, frontmatter, archdesc, c01-c12 and c elements is expressed in the CIDOC CRM ontology starting by the instance of the E31 Document representing the ead element. From this point, three new paths begin leading to three instances of the E31 Document class representing respectively the mapping of eadheader, frontmatter and archdesc. The instance of the E31 Document class representing the root element ead is linked through the P106 is composed of property to the instances of these three classes.

Correspondingly, the instances of the E31 Document representing the archival components (c01-c12 and c) are linked between them as part of the tree-based hierarchical structure via the P106 is composed of property. The tree structure obtained by mapping the EAD structure to the ontology is named as the "*Hierarchy of Documentation Elements and Attributes*" ("*HDEA*") and it is pictured in Figure 2.

4.2 The archive as a hierarchy of physical objects

An archive is a physical object, since it is a physical product of a person, a family or of a corporate body [13]. In addition, an archive as a physical object has an internal, well defined structure. In other words, an archive physically includes its components parts, which in turn include other components parts and so forth. Therefore, these archival physical objects also follow the hierarchical and

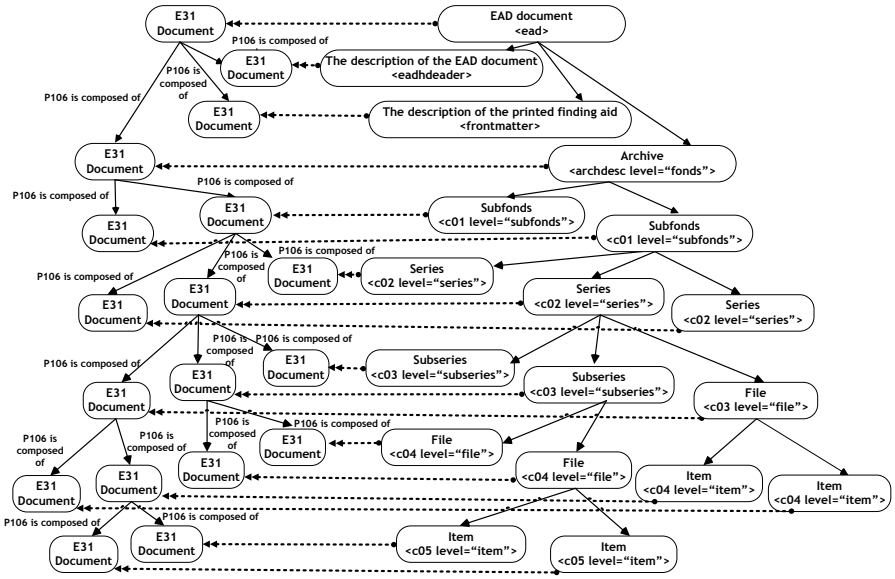

Fig. 2. Hierarchy of Documentation Elements and Attributes (HDEA).

multilevel structure. The structure of the archive as a physical object is hence expressed though the archdesc, c01-c12 and c.

In CIDOC CRM, the E22 Man-Made Object class defines the instances of the physical objects that have been created by human activity. According to this definition, every physical object expressed in EAD through the archdesc, c01-c12 and c elements is mapped to an instance of this class.

Moreover, in order to map their in between hierarchical relationship, these instances are linked via the P46 is composed of property. As it is presented in Figure 3, the tree structure obtained by mapping the archive and its components as a set of physical objects to the ontology is named as the "*Hierarchy of Physical Objects*" ("*HPO*").

4.3 The archive as a hierarchy of information objects

An archive is also an information object, since it carries information in one or more languages. An archive serves different purposes (for instance information purposes) and it is not only an evidence of the activity that produced it [13]. Both the archive and its component parts carry information. In detail, an archive contains information on its components as a set; an archival component (e.g. a subfonds) contains information on its components as a set and so on. For that reason, the informational aspect of the archive and of its components follow the hierarchical and multilevel tree structure.

To map to the CIDOC CRM ontology the concept of the archive as an object carrying information, the E73 Information Object class is used. This class includes

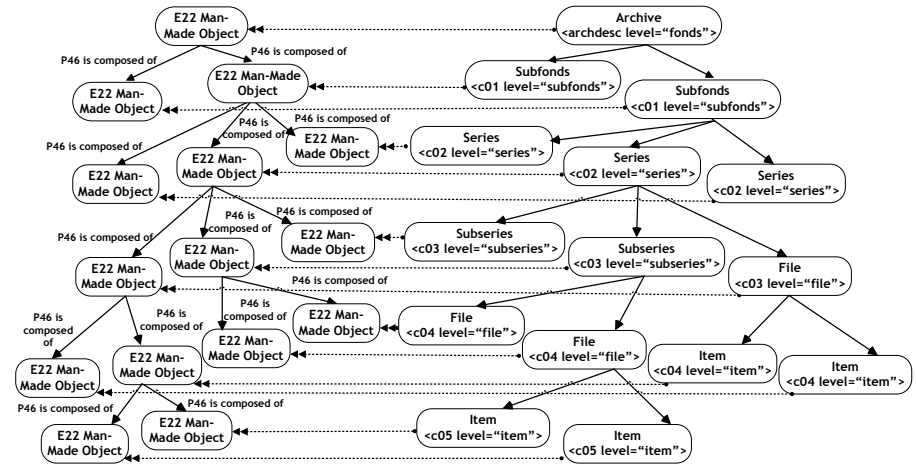

Fig. 3. Hierarchy of Physical Objects (HPO).

instances for the immaterial objects, which can be carried through any carrier. This semantic analysis comes to fully express the informational aspect of the archive, which is indeed immaterial and independent of any medium carrier [7].

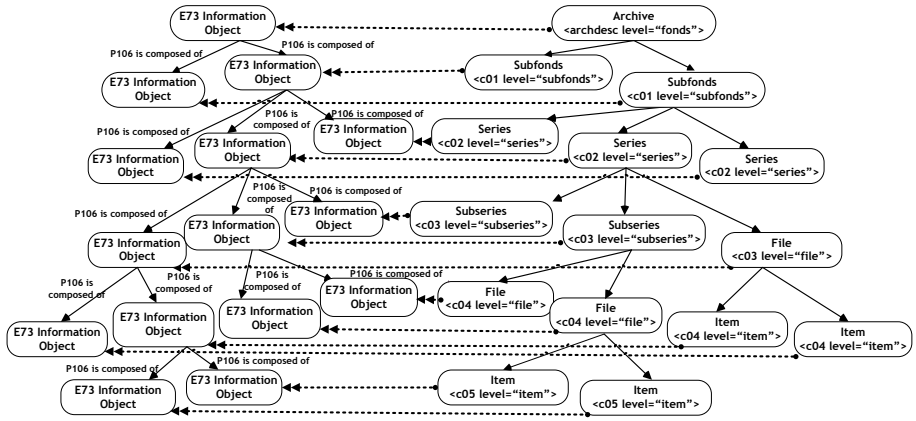

Fig. 4. Hierarchy of Information Objects (HIO).

In this context, the archdesc, c01-c12 and c elements are mapped to instances of the E73 Information Object. The expression of the hierarchical structure between these instances is defined through the P106 is composed of property. The tree structure obtained by mapping the archive and its components as a set of information objects to the ontology (Figure 4) is named as the "*Hierarchy of*

Information Objects" ("*HIO*") and it maps the semantics and the structure of the archive as an information object.

4.4 The archive as a hierarchy of linguistic objects

As mentioned in Section 4.3, an archive carries information in one or more languages, hence it is also a linguistic object. In CIDOC CRM, the E33 Linguistic Object class contains instances of information that can be expressed in one or more languages. Consequently, the semantic combination of the E73 Information Object and of the E33 Linguistic Object classes covers the semantic view of the archive as an information and linguistic object. Aiming to express these semantics, the archdesc, c01-c12 and c elements are mapped to instances of the E33 Linguistic Object class.

The expression of the hierarchical structure between these instances is defined through the P106 is composed of property, creating a hierarchy that maps the semantics and the tree structure obtained by mapping the archive and its components as a set of linguistic objects to the ontology. This tree is named as the "*Hierarchy of Linguistic Objects*" ("*HLO*") and it is pictured in Figure 5.

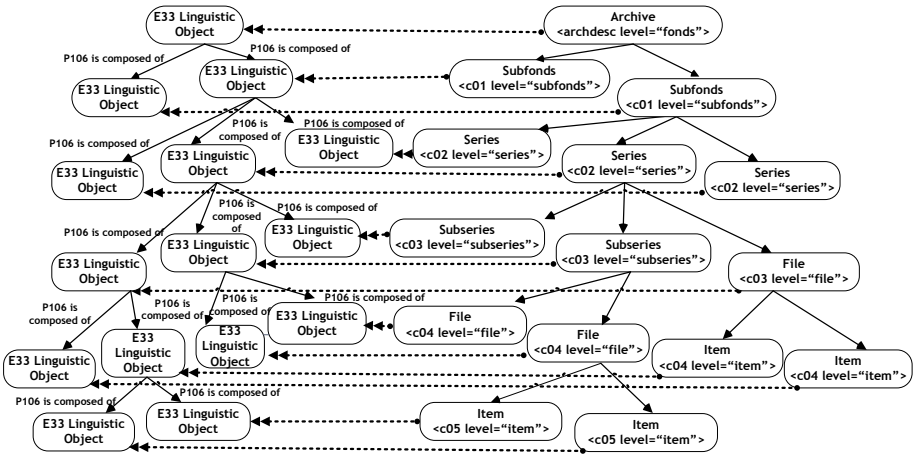

Fig. 5. Hierarchy of Linguistic Objects (HLO).

5 The relationships between the semantic views of an archive

Based on the mapping of the semantic views of the archive to the CIDOC CRM ontology, we conclude that the archive and its components are mapped to three

different CIDOC CRM hierarchies (*HPO*, *HIO* and *HLO*), each of them representing a different structured semantic view of the archive. Besides, the archival description is mapped to another hierarchy of the ontology (*HDEA*).

Note that these four hierarchies have the same structure and they only differ in terms of the names of the classes appearing in the tree nodes. It is clear that these hierarchies refer to the same object (the archive), which is documented by the archival description. Moreover, based on the analysis of Section 4, it is concluded that these hierarchies are semantically related to each other. Hence, it is necessary to: a) relate the four hierarchies with the tree of the EAD (and in particular with the archdesc, c01-c12 and c elements), since it is the metadata schema that expresses the archival description, and b) to associate these four hierarchies, since they all refer to the same object, the archive.

As the archdesc, c01-c12 and c elements are firstly referred to the archival description, which incorporates the semantic views of the archive, the *HDEA* is the starting point for the association of the different views. In detail, the *HDEA* refers to the archive as a physical object. Furthermore, the archive as a physical object carries information. Moreover, the analysis of this information can produce additional information for the archive. An illustrative example is the abstract of the archive's content as well as the controlled access points. Finally, an archive can also be a carrier of linguistic content, since the information it carries is usually expressed via written and/or oral speech, independently of the medium that carries this content.

In order to show an example of the hierarchies' association, the node <c01 level="subfonds"> of the EAD structure is chosen, and more specifically the node that contains three archival series in the Figure 1. This node is mapped to an instance of the E31 Document class expressing the documentation of this specific node that represents a subfonds. In detail, this instance documents its corresponding node in the *HPO* (see Figure 3), which is an instance of the E22 Man-Made Object class that represented the subfonds as a physical object. This relationship is expressed in the CIDOC CRM ontology through the P70 documents property that has as a domain the instances of the E31 Document class and as range the instances of the E1 CRM Entity class. For that reason, it can associate the instance of the E31 Document class to its corresponding instance of the E22 Man-Made Object class (since E22 Man-Made Object is a subclass of E1 CRM Entity).

To continue with, the subfonds as a physical object carries information and thus it is also an information object, hence an instance of the E73 Information Object maps the <c01 level="subfonds"> node in the *HIO* of the Figure 4. The relationship between these two instances (i.e. the instance of E22 Man-Made Object and E73 Information Object representing the same element (<c01 level="subfonds">) can be expressed through the P128 carries property, which has as domain the E24 Physical Man-Made Thing class and as range the E73 Information Object class. For that reason, it relates the instance of the E22 Man-Made Object class (which is a subclass of the E24 Physical Man-Made Thing

class) to the component of the archive documented in the instance of the E73 Information Object class.

The component of the archive documented in the <c01 level="subfonds"> element may also be an information object that carries information in one or more languages and this semantic view can be expressed as an instance of the E33 Linguistic Object, being in the same position in the *HLO* as it is in the *HIO* (see Figure 5). The relationship between these two instances is expressed in the CIDOC CRM ontology through the P67 refers to property, which has as a domain the E89 Propositional Object and as a range the E1 CRM Entity, hence linking the instance of the E73 Information Object (which is a subclass of the E89 Propositional Object) to its corresponding instance of the E33 Linguistic Object (which is a subclass of the E1 CRM Entity).

As a consequence, these four hierarchies are linked in a way that allows the expression of their in between relationship inside the CIDOC CRM ontology. This "chain of relationships" is expressed through the following CIDOC CRM path:

> E31 Document → P106 is composed of → E31 Document → P70 documents → E22 Man-Made Object → P128 carries → E73 Information Object → P67 refers to → E33 Linguistic Object

This path declares that an EAD document includes the archival description (E31 Document → P106 is composed of →), which is the documentation (E31 Document → P70 documents) of a physical object that has been created by human activity (E22 Man-Made Object) and that carries (P128 carries) information which is immaterial and can be carried by any physical medium (E73 Information Object). To finish, the information carried by the archive can be expressed in one or more languages (P76 refers to → E33 Linguistic Object).

This "chain of relationships" expresses in the CIDOC CRM ontology the semantics for every archival unit (encoded in archdesc, c01-c12 and c) defining a horizontal relationship between them in every descriptive level. Therefore, the instances representing the archival units and being expressed in a vertical relationship inside the four hierarchies (*HDEA*, *HPO*, *HIO* and *HLO*) are also interconnected horizontally so as to express the relationship between the different semantic hierarchies of the archive and its description (see Figure 6).

6 Associating the EAD descriptive fields with the semantic hierarchies

Besides the mapping of the archdesc, c01-c12 and c elements studied in the previous sections, the mappings for the EAD descriptive fields that include the information for the content and context of the archive are also provided.

With the intention of defining the mappings of these elements/attributes to the CIDOC CRM, we are based on their semantics as they appear in the EAD Tag Library [8] and the published best practices and implementation guidelines for the EAD (for example the [10]). Derived from this investigation, we associate

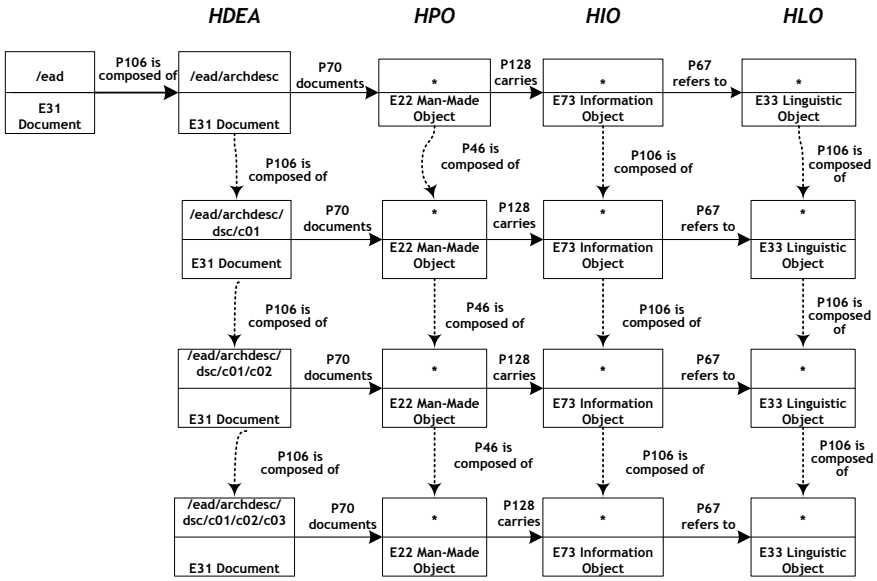

Fig. 6. Parallel Hierarchies.

the information content of the elements and attributes to one or more of the semantic hierarchies. More specifically, to map these elements/attributes to the CIDOC CRM the following steps are followed:

- Step 1: Associate the element/attribute to the semantic hierarchy(ies).
- Step 2: Select an appropriate CIDOC CRM class to map the element/attribute.
- Step 3: Associate the class selected in Step 2 (by constructing the appropriate paths) with the proper semantic hierarchy(ies) selected in Step 1.

6.1 Association of the element/attribute to the semantic hierarchy(ies)

A class' instance (on which an element/attribute is mapped) may be associated to: a) the *HDEA*, when it provides information for the archival documentation, b) the *HPO*, when it provides information for the archive as a physical object, c) the *HIO*, when it provides information for the archive as an information object, and d) the *HLO*, when it provides information for the archive as a linguistic object.

In Table 2 several EAD elements and attributes and the semantic hierarchy(ies) they are associated with are presented. In order to come up with this proposal the semantics of every node, and based on them, its association with one or more hierarchies are defined. In the following paragraphs, examples of nodes associated to the four hierarchies are presented.

More analytically, the information that refers to the EAD document and the archival description is semantically associated with the *HDEA*. For instance, the attributes of the archdesc element are referred to the *HDEA*, provided that they encode information for the archival description. Illustrative examples are the attributes audience and relatedencoding[1]:

- audience: This attribute provides information to help controlling whether the information contained in the element (to which the audience is attached) should be available to all viewers or only to the repository staff.
- relatedencoding: This attribute defines a descriptive encoding system, such as MARC 21, to which certain EAD elements can be mapped using the encodinganalog attribute.

Hence, given their meaning, both attributes are associated with the *HDEA*.

Subnode of the archdesc or c01-c12	HDEA	HPO	HIO	HLO
@audience	x			
@level	x	x		
@otherlevel	x	x		
@relatedencoding	x			
@type	x			
accessrestrict		x	x	
altformavail			x	
arrangement		x	x	
bioghist		x		
controlaccess			x	
fileplan		x	x	
phystech		x	x	
relatedmaterial			x	
scopecontent			x	
separatedmaterial		x		
userestrict			x	
did/unittitle			x	
did/note			x	
did/physloc		x		
did/unitdate		x		
did/langmaterial				x
did/unitid		x	x	
did/origination		x		

Table 2. The association of some EAD nodes with the semantic hierarchies.

The nodes that refer to the archive as a physical object have as their point of reference the *HPO* and, as a consequence, the E22 Man-Made Object class. These nodes are mostly part of the did wrapper element or they are part of the administrative and supplemental information for the archive. Illustrative examples are the creator of the archive (origination), its date of creation (unitdate), its

[1] Note that this attribute is an attribute of the archdesc and not of the c01-c12 and c.

physical location (physloc) etc. An example of an element associated with the *HPO* is the following:

- origination: This element provides information about the individual organization responsible for the creation, accumulation, or assembly of the described materials. The activities of creating, accumulating or assembling the archival material are all associated with its physical substance. Thus, its association with the *HPO* is obvious.

Furthermore, most of the administrative and supplemental information included in the archdesc, c01-c12 and c elements refers to the informational aspect of the archival material, which is expressed by the instance of the E73 Information Object. This information is provided from the archive and sometimes it comes up after its content analysis, such as the scope and content of the archive (scopecontent), its custodial history (custodhist) etc. What is more, certain subelements of the did wrapper element (such as the unittitle and abstract) refer to the E73 Information Object. For example:

- unittitle: This element declares the title of the archival unit, which is a name either given by the archivist or expressed by the archival unit. Thus, the unittitle is an information provided by the archival unit or by the archivist (after its context and content analysis), hence it is associated with the *HIO*.

The archive is also a linguistic object, since it can carry verbal or oral speech. For this reason, there are nodes that are associated with the *HLO*. Currently in EAD, there is only one element referred to this semantic hierarchy, the langmaterial, provided that this element includes a prose statement enumerating the language(s) of the archival materials found in the unit being described.

It is important to notice that - while analyzing the semantics of certain subnodes of the archdesc, c01-c12 and c elements we conclude that they are associated with more than one of the four hierarchies and this fact arises from their semantics. For example, the unitid element defines the identifier of the archival unit, which is a unique reference point for it or a control number, such as the accession number or the classification number, and sometimes it secondarily provides location information. Hence, this element refers to the descriptive unit as a physical object (when it identifies the archival unit to its accession or its location), nevertheless it is also information given by the archivist in order to uniquely identify the item. Thus, the unitid is associated both to the *HPO* and the *HIO*.

6.2 Selection of a CIDOC CRM class to map the elements/attributes and its association with the semantic hierarchy(ies)

In Section 6.1, we presented how an element/attribute is associated with the appropriate semantic hierarchy(ies) based on the semantics of this element/attribute. The next step that must be followed is to map this node to an appropriate class.

Then, this class must be connected to the appropriate node of the semantic hierarchy (i.e. the node that corresponds to the archival component to which the node refers to) through an appropriate constructed CIDOC CRM path. This path consists of a single CIDOC CRM property; often it includes several properties and intermediate classes.

The presentation of the mappings of the EAD nodes is beyond the scope of this paper. Nonetheless, in the following paragraphs, some examples are presented to show the above mentioned paths. As you will see below, the relatedencoding is mapped to a CIDOC CRM path that includes several properties and intermediate classes, while the langmaterial is mapped to a CIDOC CRM path that consists of a single CIDOC CRM property"

- relatedencoding: The relatedencoding attribute includes values that define the descriptive encoding system to which the EAD elements can be mapped and, as already mentioned, it is associated with the *HDEA*. It is semantically mapped to the E55 Type, which is also semantically associated with the E32 Authority Document in order to define that the E55 Type instances are taken from an authoritative vocabulary named "relatedencoding". The EAD path (/ead/archdesc/@relatedencoding) is mapped to the following CIDOC CRM path: E31 Document → P106 is composed of → E31 Document → P2 has type → E55 Type → P71 lists in → E32 Authority Document{=relatedencoding}, declaring that the EAD documentation (E31 Document) consists of (P106 is composed of) the documentation of the archive (E31 Document), which has a specific type (P2 has type → E55 Type) and that this type is characterized (P71 lists in) as relatedencoding (E32 Authority Document{=relatedencoding}).
- langmaterial: This element encodes the language(s) in which the archive is written or expressed and it is mapped to an instance of the E56 Language, which comprises the natural languages. Based on its semantics, it is associated with the *HLO*. The EAD path (/ead/archdesc/did/langmaterial/language) is mapped to the following CIDOC CRM path: E31 Document → P106 is composed of → E31 Document → P70 documents → E22 Man-Made Object → P128 carries → E73 Information Object → P67 refers to → E33 Linguistic Object → P72 has language → E56 Language. This path expresses that the EAD documentation (E31 Document) consists of (P106 is composed of) the archive's documentation (E31 Document) that documents (P70 documents) a physical object (E22 Man-Made Object), which carries (P128 carries) an information object (E73 Information Object). Additionally, the archive is a linguistic object (P67 refers to → E33 Linguistic Object), which is expressed in one or more languages (→ P72 has language → E56 Language).

7 Discussion and related work

The key problem of integrating XML metadata schemas is an issue of great concern to the international research community. However, in most integration efforts no emphasis is given to the mapping of the semantics and of the documentation's targets of an XML metadata schema, even though these characteristics

shape the area of the metadata in archives, libraries and museums and that are indeed based on documentation policies followed for many years.

According to the literature, there are many XML (meta)data mapping to the CIDOC CRM ontology efforts, since this ontology is considered one of the most appropriate models in integration architectures. An example is the work of the *STAR* project [5], in which access to digital archaeological sources is enhanced through the mapping of them to an extension of the CIDOC CRM. Furthermore, the issue of mapping the Cultural Heritage metadata schemas to the ontology is also explored in the *BRICKS* project [15].

A well documented research proposal in relation to the mapping of the EAD semantics is presented in [16]. This mapping of EAD to CIDOC CRM ontology differs from the proposed mapping of our research work on the following points:

- this mapping refers to the first version of the EAD,
- the different semantic views of the archive and of the archival description are not defined and analyzed, hence not mapped to the ontology, and
- the EAD is considered as a format for describing the whole and there is no reference in mapping its hierarchical structure.

In general, the semantics of the metadata and of the information sources they describe are not taken into account while creating their mappings to an ontology. In [1] the mapping of the XML metadata schema of the Cultural Heritage domain to an ontology (which is similar to the CIDOC CRM) is proposed, nevertheless there is no reference to the importance of the metadata semantics.

The proposed mapping of the EAD to the CIDOC CRM ontology is targeted not only to capture the syntactic rules, but also to express the rich semantics of the EAD and of the information source it describes. The main goal is to be able to use this mapping in various integration scenarios that implement the CIDOC CRM as the mediated schema. It should be also noted that other mappings work of our team have been proposed, for schemas such as the DC and the VRA to the CIDOC CRM ontology (respectively presented in [12, 2]).

To conclude, we are currently working on the issue of the inheritance of information. Note that the inheritance of information between the hierarchically linked descriptive levels is one of the main characteristics of the archival description. Thus, specific techniques are needed in order to take into account this characteristic, during the mapping of an EAD document to the CIDOC CRM, otherwise considerable information may be lost.

References

1. B. Amann, C. Beeri, I. Fundulaki, and M. Scholl. Ontology-Based Integration of XML Web Resources. In I. Horrocks and J. Hendler, editors, *The 1st Int. Semantic Web Conference, Sardinia, Italy, June 912, 2002 Proceedings*, volume 2342 of *LNCS*, pages 117–131. Springer, 2002.
2. C. Kakali and I. Lourdi and T. Stasinopoulou and L. Bountouri and C. Papatheodorou and M. Doerr and M. Gergatsoulis. Integrating Dublin Core metadata

for cultural heritage collections using ontologies. In *Proc. of the Int. Conference on Dublin Core and Metadata Applications (DC 2007), Singapore, 27 - 31 August*, pages 128–139, 2007.

3. CIDOC CRM Special Interest Group. Definition of the CIDOC Conceptual Reference Model, version 5.0.2. Technical report, January 2010.
4. I.F. Cruz and H. Xiao. The Role of Ontologies in Data Integration. *Journal of Engineering Intelligent Systems*, 13(4):245–252, 2005.
5. D. Tudhope and C. Binding and K. May. Semantic interoperability issues from a case study in archaeology. In S. Kollias and J. Cousins, editor, *Semantic Interoperability in the European Digital Library, Proc. of the 1st Int. Workshop SIEDL 2008, associated with 5th ESWC*, pages 88–99, 2008.
6. T. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.
7. International Council on Archives. Committee on Descriptive Standards. *ISAD(G): General International Standard Archival Description*. ICA, 2nd edition, 2000.
8. Library of Congress. Encoded Archival Description Tag Library Version 2002. `http://www.loc.gov/ead/tglib/index.html`, 2002.
9. Library of Congress. Encoded Archival Description: Version 2002. `http://www.loc.gov/ead/`, 2002.
10. Library of Congress. Library of Congress Encoded Archival Description Best Practices. `http://www.loc.gov/rr/ead/lcp/lcp.pdf`, 2008.
11. M. Doerr. The CIDOC CRM: An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*, 24:75–92, 2003.
12. M. Gergatsoulis and L. Bountouri and P. Gaitanou and C. Papatheodorou. Mapping Cultural Metadata Schemas to CIDOC Conceptual Reference Model. In S. Konstantopoulos and S. Perantonis and V. Karkaletsis and C.D. Spyropoulos and G. Vouros, editor, *Artificial Intelligence: Theories, Models and Applications*, volume 6040 of *LNCS*, pages 321–326. Springer, 2010.
13. M.J. Fox and P.L. Wilkerson. *Introduction to Archival Organization and Description: Access to Cultural Heritage*. Getty Publications, 1999.
14. N. Noy. Semantic Integration: a Survey of Ontology-Based Approaches. *SIGMOD Record*, 33(4):65–70, 2004.
15. P. Nussbaumer and B. Haslhofer. CIDOC CRM in Action: Experiences and Challenges. In L. Kovacs, N. Fuhr, and C. Meghini, editors, *Research and Advanced Technology for Digital Libraries. ECDL 2007, Budapest, Hungary, September 16-21, 2007. Proc.*, volume 4657 of *LNCS*, pages 532–533. Springer, 2007.
16. M. Theodoridou and M. Doerr. Mapping of the Encoded Archival Description DTD Element Set to the CIDOC CRM. Technical Report 289, June 2001.

Mapping VRA Core 4.0 to the CIDOC CRM ontology*

Panorea Gaitanou and Manolis Gergatsoulis

Database & Information Systems Group (DBIS),
Laboratory on Digital Libraries and Electronic Publishing,
Department of Archives and Library Science, Ionian University,
Ioannou Theotoki 72, 49100 Corfu, Greece.
{rgaitanou, manolis}@ionio.gr

Abstract. In this paper, we present an effort to semantically map VRA Core 4.0, a cultural heritage metadata schema describing visual resources, to CIDOC CRM. This work is based on a semantic integration scenario, where CIDOC CRM acts as a mediation schema. More specifically, each element of the schema (along with its subelements and attributes) is mapped to the equivalent CRM path (represented as a sequence of classes and properties). The mapping is formally described using a Mapping Description Language (MDL), which explicitly defines semantic rules from the source schema to the target schema.

1 Introduction

Managing cultural heritage resources is a rather complex process, in which a range of sciences and scientists (computer scientists, information scientists, archives scientists, museologists, historians, etc.) are involved. Cultural heritage institutions are challenged to handle the information and knowledge dissemination in such a way that the needs and demands of various user groups are efficiently met. Within this framework, cultural heritage institutions (otherwise called "memory institutions") use various metadata schemas for the documentation of cultural collections, that facilitate access and retrieval to cultural information via the web. The complexity of the cultural information imposes the development of several different metadata standards (such as DCMI, VRA Core 4.0, EAD, Spectrum etc.), which exhibit significant diversity. This heterogeneity often results in data exchange failure, as the end user cannot access an integrated information system and retrieve the desired information. In order to address all the aforementioned issues and achieve a unified and standard-independent access to the relative information, it is necessary to integrate all these schemas. One

* This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

of the most important and continuously evolving methods implemented in the interoperability domain is the ontology-based integration [7]. Ontologies provide the means for defining common vocabularies, representing the domain knowledge, while at the same time facilitating knowledge sharing and reuse among heterogeneous and distributed application systems.

The basic component of an information integration system is the mapping of the various metadata schemas to a schema or a core ontology, acting as a mediation schema, so that (meta)data integration is successfully accomplished. In the integration scenario proposed by our research group [9], the CIDOC CRM ontology is used as a mediation schema, to which different metadata schemas (such as MODS, DC, MARC, EAD etc.) are mapped.

In this paper, we present a mapping methodology of the VRA Core 4.0 schema to the CIDOC CRM ontology. This methodology, which is based on a path-oriented approach, is formally defined using a Mapping Description Language (MDL), which defines semantic rules from the source schema to the target schema. In the proposed methodology, each element of the schema (with its subelements) is represented as a VRA path (expressed in XPath form) and is then semantically translated to an equivalent path of classes and properties of CIDOC CRM. It is important to note that the mapping procedure focuses on the restricted version of the VRA Core 4.0, which imposes controlled vocabularies and type lists. Thus, each attribute assigned to an element/subelement of the metadata schema may generate different semantic paths on the ontology, depending on the value it takes each time, and produces a plethora of conceptual expressions for the same element/subelement. The use of several global attributes provided by the VRA Core 4.0 schema makes the mapping procedure even more complicated, by generating additional semantic paths on the ontology.

2 Mapping VRA Core 4.0 to CIDOC CRM

2.1 Brief description of the VRA Core 4.0

VRA Core 4.0 [10] is a metadata schema for the cultural heritage community, initially developed by the Visual Resources Association's Data Standards Committee. Currently, it is hosted by the Network Development and MARC Standards Office of the Library of Congress (LC) [5] in partnership with the Visual Resources Association. VRA Core 4.0 provides guidance on describing works of visual culture, collections, as well as images that document them. Therefore, it allows for three broad groups of entities, which are works, images, and collections. A work may represent a painting, sculpture or other artistic product. An image is a visual representation of a work that can come in a wide range of formats, and include various image formats (such as JPEG, GIF, TIFF) or could include physical photographs, slides, etc. Finally, a collection represents a group of works or images.

VRA Core 4.0 contains 19 elements (`work`, `agent`, `culturalContext`, `date`, `description`, `inscription`, `location`, `material`, `measurements`, `relation`,

rights, source, stateEdition, stylePeriod, subject, technique, textref, title and worktype) and several optional global attributes (dataDate, extent, href, pref, refid, rules, source, vocab, xml:lang), which are applied additionally to any element or subelement, when necessary. Two XML Schema versions have been proposed for the VRA Core 4.0. An *unrestricted version*, which specifies the basic structure of the schema and imposes no restrictions on the values entered into any of the elements, sub-elements, or attributes, and a *restricted version*, which extends the unrestricted one by imposing controlled type lists and date formats.

Example 1. In this example we present a fragment of a simplified VRA document, describing a textual manuscript of the 18th century, taken from
http://www.vraweb.org/projects/vracore4/example017.html.

```
<?xml version="1.0" encoding="UTF-8" ?>
 <vra>
  <work id="w_4" source="Core 4 Sample Database (VCat)" refid="4">
   <agentSet>
    <agent>
     <name vocab="ULAN" refid="500017255"
           type="personal">Jefferson, Thomas</name>
     <dates type="life">
      <earliestDate>1743</earliestDate>
      <latestDate>1826</latestDate>
     </dates>
     <culture>American</culture>
     <role>author</role>
    </agent>
   </agentSet>
   <measurementsSet>
    <measurements type="height" unit="cm">75.56</measurements>
    <measurements type="width" unit="cm">62.23</measurements>
   </measurementsSet>
   <stylePeriodSet>
    <stylePeriod vocab="LCSAF"
                 refid="85041401">Eighteenth century</stylePeriod>
   </stylePeriodSet>
   <techniqueSet>
    <technique vocab="AAT" refid="300053162">calligraphy(process)</technique>
    <technique vocab="AAT" refid="300054698">writing(process)</technique>
   </techniqueSet>
   <titleSet>
    <title type="popular" xml:lang="en">Declaration of Independence</title>
   </titleSet>
   <worktypeSet>
    <worktype>manuscript (document genre)</worktype>
   </worktypeSet>
  </work>
 </vra>
```

2.2 The CIDOC CRM ontology

The *CIDOC Conceptual Reference Model* (CIDOC CRM) [3], which emerged from the CIDOC Documentation Standards Group in 1999, is a formal extensible ontology, which aims at providing a conceptual representation of cultural heritage domain, promoting semantic interoperability and integration. It is an object-oriented model comprised of a class hierarchy of 86 named classes interlinked by 137 named properties. CIDOC CRM defines the complex interrelationships between objects, actors, events, places, and other concepts used in the cultural heritage domain [2].

A *class* (also called *entity*), identified by a number preceded by the letter "E" (e.g. E1 CRM Entity, E2 Temporal Entity), groups items (called *class instances*) that share common characteristics. A class may be the *domain* or the *range* of *properties*, which are binary relations between classes. Properties are identified by numbers preceded by the letter "P" (e.g. P2 has type (is type of) with domain the class E1 CRM Entity and range the class E55 Type). A property can be interpreted in both directions (active and passive voice), with two distinct but related interpretations. A *subclass* is a class that specializes another class (its *superclass*). A class may have one or more immediate superclasses. When a class A is a subclass of a class B then all instances of A are also instances of B. A subclass inherits the properties of its superclasses without exception (*strict inheritance*) in addition to having none, one or more properties of its own. A *subproperty* is a property that specializes another property. A sample of CIDOC CRM properties is shown in Fig. 1.

Property Id & **Name**	**Entity - Domain**	**Entity - Range**
P1 is identified by (identifies)	E1 CRM Entity	E41 Appelation
P2 has type (is type of)	E1 CRM Entity	E55 Type
P4 has time-span (is time-span of)	E2 Temporal Entity	E52 Time-Span
P14 carried out by (performed)	E7 Activity	E39 Actor
P58 has section definition (defines section)	E18 Physical Thing	E46 Section Definition
P108 has produced (was produced by)	E12 Production	E24 Physical Man-Made Thing

Fig. 1. A sample of CIDOC CRM properties.

2.3 The Mapping Description Language (MDL)

The proposed mapping method between the metadata schemas and CIDOC CRM is based on a path-oriented approach. A mapping from a source schema to a target schema transforms each instance of the source schema into a valid instance of the target schema. Hence, we interpret the metadata paths to semantically equivalent CIDOC CRM paths. As we are interested in metadata schemas, which are based on XML, the paths in the source schemas are based on XPath [11],

in fact they extend the XPath *location paths* with *variables* and stars (meaning *data transfer*). The syntax of the MDL *mapping rules* is given bellow in EBNF:

(R1) R ::= Left '−−' Right
(R2) Left ::= A_{Path} | V_{Path}
(R3) A_{Path} ::= ϵ | '/' R_{Path}
(R4) R_{Path} ::= L | L '*' | L '{' V_l '}' | L '*' '{' V_l '}'
(R5) V_{Path} ::= '$' V_l '/' R_{Path} | '$' V_l '{' V_l '}'
(R6) Right ::= E_t | E_e '→' O | '$' V_c '→' O | '$' V_p '→' P_p '→' E_{t55}
(R7) O ::= P_e '→' E_t | P_e '→' E_e '→' O
(R8) E_e ::= E | E '{' V_c '}'
(R9) E_t ::= E | E '{' V_c '}' | E '{=' String '}'
(R10) E_{t55} ::= E55 | E55 '{' V_c '}' | E55 '{=' String '}'
(R11) P_e ::= P | P '{' V_p '}'

The terminals used in these rules have the following semantics:

- L: it represents an XPath *location path*.
- V_l: it represents the *location variables*, which are used to declare the "branches" of the XML trees (XPath paths).
- V_c: it represents the *class variables*. The class variables are used to declare that a class can be the starting point of one or more CIDOC CRM paths.
- V_p: it represents the *property variables* The property variables are used to declare that a property can be the starting point of a new CIDOC CRM path, which - in this case - it is a property of a property linking the property that the variable represents to an instance of the E55 Type class.
- E: it represents the identifier of the class.
- E55: it represents the identifier of the class E55 Type.
- P: it represents the identifier of the property.
- P_p: it represents the identifier of the property of a property.
- *String*: it represents a string.

3 Mapping VRA Core 4.0 elements to equivalent CIDOC CRM paths

VRA Core 4.0 is an XML-based standard, therefore we use the XPath to locate VRA elements/attributes. A *VRA path* is a sequence of VRA elements and subelements, starting from the schema root element `vra` separated by the slash symbol (/). For instance, the path `/vra/work/titleSet/title` denotes the title of a work being described. A *CRM path* is defined as a chain in the form entity-property-entity, such that the entities associated with a property correspond to the property's domain and range. VRA Core defines three basic top elements: `work`, `collection` and `image`. In the context of a VRA Core 4.0 record, a work is defined as a physical entity that exists, existed in the past, or may exist in the future. It might be an artistic creation, such as a painting or a sculpture, a performance, a building or other construction, etc. Therefore, we associate each `work` element in a VRA document with an instance of the class `E24 Physical Man-Made Thing`, which comprises all persistent physical items that are purposely created by human activity.

In the following paragraphs, we present the mapping of the `agent` element of the VRA Core 4.0 schema to CIDOC CRM. The `agent` (including its subelements and attributes) is a representative element of VRA Core 4.0 and its mapping presents significant diversity and complexity. The methodology applied to this mapping can be used to map the other elements of the VRA Core 4.0 as well.

3.1 Mapping the `agent` element and its subelements

The `agent` element denotes a person, group or corporate body that has contributed to the production or creation of the work being described. It contains the following five subelements: `name`, `culture`, `dates`, `role` and `attribution`. Each one of them provides a part of the `agent` element:

- The `name` subelement specifies the names and appellations, assigned to an individual, group or corporate body. A `type` attribute is assigned to this subelement, with possible values `personal`, `corporate`, `family`, or `other`.
- The `culture` subelement refers to the nationality or culture of the person, group, or corporate body that participated to the work being described.
- The `dates` subelement, which contains two additional subelements, namely the `earliestDate` and the `latestDate`, refers to the dates associated with the agent. A `type` attribute is also assigned to this subelement (with possible values `activity`, `life` and `other`).
- The `role` subelement denotes the specific role of the individual, group or corporate.
- The `attibution` subelement defines a characteristic or a specific attribute related to the agent.

Mapping the ***`agent`*** *element:* It is easy to see, by examining the semantics of the CIDOC CRM classes, that the appropriate class of CIDOC CRM to map the `agent` element of VRA is the class `E39 Actor`. The instances of `E39 Actor` corresponding to each specific agent need to be related to the instance of `E24 Physical Man-Made Thing` representing the work being described, in order to express that an agent "contributed to the production or creation of the work being described". However, as CIDOC CRM is event-centric, it does not provide properties to directly relate the instances of these two classes. Instead, these instances can be related indirectly, though an event (instance of the class `E12 Production`) during which the object was created. In this way, the work being described (i.e. the instance of the class `E24 Physical Man-Made Thing`) is related through the property `P108B was produced by` to this event. Additionaly, this event should then be related to the instances of the class `E39 Actor` (representing the agent), through the property `P14 carried out by`. In this way, a CIDOC CRM path of the following form is created:

`E24 Physical Man-Made Thing` → `P108B was produced by` →
`E12 Production` → `P14 carried out by` → `E39 Actor`

which semantically corresponds to a VRA path of the form:

/vra/work/agentSet/agent

We should note that, in case there are more than one agents (i.e. more than one agent subelements of the element agentSet), different subpaths of the form:

→ P14 carried out by → E39 Actor

will be rooted to the (same) instance of E12 Production to relate it with the different agents (instances of E39 Actor) that took part in this production event.

Mapping the name subelement of agent element: The name subelement, which identifies the name of an agent, is mapped to an instance of the class E82 Actor Appellation and is linked to the corresponding instance of the class E39 Actor through the property P131 is identified by. In this way, the CIDOC CRM path, which semantically corresponds to the VRA path:

/vra/work/agentSet/agent/name

becomes:

E24 Physical Man-Made Thing → P108B was produced by →
E12 Production → P14 carried out by → E39 Actor →
P131 is identified by → E82 Actor Appellation

Mapping the type attribute of the name subelement: In VRA an attribute named type is assigned to the name element. This attribute is quite remarkable given that it determines if an agent is a person (when the value of type is personal), a corporate or an organization (when the value of type is corporate), a family (when the value of type is family), or none of the above (when the value of type is other). To map the attribute type in CIDOC CRM, we have investigated two different approaches:

First approach: A first approach to map the type attribute in CIDOC CRM is to employ the class E55 Type and link instances of this class (of the values personal, corporate, family or other respectively) to the corresponding instances of the class E39 Actor through the property P2 has type. In this case, the following CIDOC CRM path will be created:

E24 Physical Man-Made Thing → P108B was produced by →
E12 Production → P14 carried out by → E39 Actor [→ P2 has type → E55 Type] → P131 is identified by → E82 Actor Appellation

which semantically corresponds to the VRA path:

/vra/work/agentSet/agent/name[@type]

Notice that in this approach the value of the type attribute is given as value of the instance E55 Type.

The notation [...] in the CIDOC CRM path is used to denote that a new branch is rooted on the E39 Actor class node.

Second approach: A second approach to map the type attribute in CIDOC CRM is to refine the mapping of the specific agent by replacing the class E39 Actor with an appropriate subclass of this class determined by the value of the type attribute. More specifically, if the value of the type attribute is personal, then the corresponding agent can be considered to be an instance of the class E21 Person. In this case, the CIDOC CRM path becomes:

E24 Physical Man-Made Thing → P108B was produced by →

E12 Production → P14 carried out by → E21 Person → P131 is identified by → E82 Actor Appellation

which semantically corresponds to the VRA path:

/vra/work/agentSet/agent/name[@type="personal"]

If the value of the type is corporate, the corresponding agent will be denoted as an instance of the class E40 Legal Body, while if the value is family then the corresponding agent will be considered as an instance of the class E74 Group.

Fig. 2 depicts the mapping of the agent element and its subelements when the type attribute has the value personal, while applying the second approach. In this figure, the upper part of each box indicates the VRA path mapped to the CIDOC CRM class shown in the lower part. The boxes are linked with arrows that represent CIDOC CRM properties, which appear as labels to these arrows. In case a property is used according to its inverse property name, it is characterized by the letter "B" as part of its name (e.g. P108B was produced by). The mapping of other subelements of the element agent, appearing also in Fig. 2 (that is the subelements culture, role and dates), will be presented in the following paragraphs. At this point, we should mention that the type attribute assigned to the name subelement exhibits a rather weak point of the VRA Core Schema, as it actually refers to the agent element to which we believe that is should have been assigned and not to the name subelement.

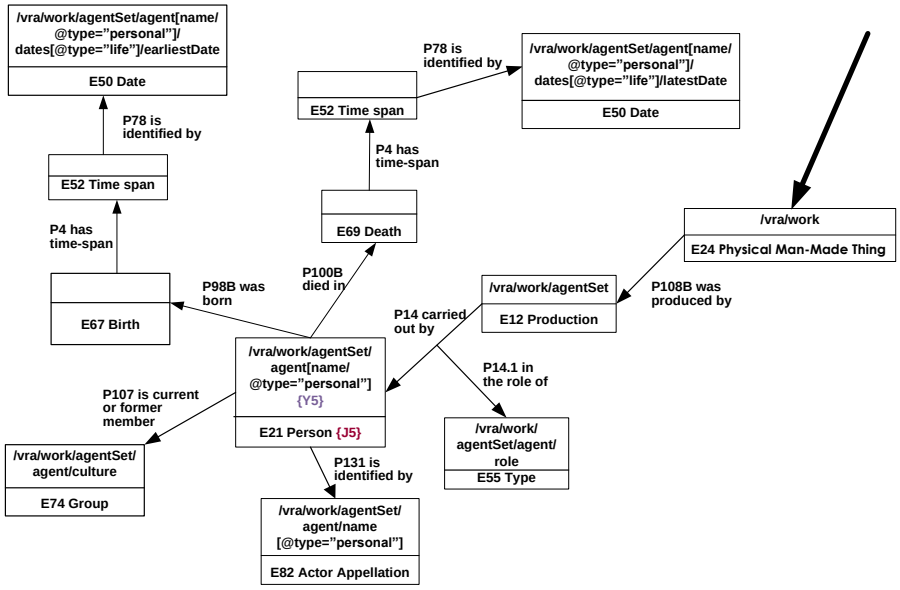

Fig. 2. The mapping of the value personal of the type attribute of the element agent.

Mapping the `role` *subelement of* `agent`*:* The `role` subelement, which identifies the role of an agent, is expressed in CIDOC CRM through the subproperty `P14.1 in the role of`, which actually links the property `P14 carried out by` to an instance of the class `E55 Type`. In this way, the CIDOC CRM path becomes:

`E24 Physical Man-Made Thing` → `P108B was produced by` →
`E12 Production` → `P14 carried out by` [→ `P14.1 in the role of` → `E55 Type`] → `E39 Actor`

which semantically corresponds to the VRA path:

`/vra/work/agentSet/agent/role`

Mapping the `culture` *subelement of* `agent`*:* The `culture` subelement, which identifies the nationality or culture of an agent, can be modelled as a membership of the agent to a group. This group is modelled in CIDOC CRM as an instance of the class `E74 Group`, which is related to the corresponding instance of the class `E39 Actor`, through the property `P107B is current or former member of`, resulting in the CIDOC CRM path of the form:

`E24 Physical Man-Made Thing` → `P108B was produced by` →
`E12 Production` → `P14 carried out by` → `E39 Actor` →
`P107B is current or former member of` → `E74 Group`

which semantically corresponds to the VRA path:

`/vra/work/agentSet/agent/culture`

Mapping the `dates` *subelement of* `agent`*:* The `dates` subelement is one of the most complex subelements to map, for three specific reasons: a) it contains a `type` attribute, with possible values `life`, `activity`, and `other`. Thus, it can define either the dates that span the known activity of an individual, group or corporate body, or the birth and death dates of a person (or even none of the above, by implementing the `other` attibute), b) it is strongly related to the `name` subelement, and more specifically to the value of the `type` attribute of the `name` subelement. For instance, if the `type` attribute of the subelement `name` is defined as `corporate`, then the value of the `type` attribute of the `dates` subelement can be either `activity` or `other`, denoting eg. the foundation dates of a corporate body, c) it contains two additional subelements, `earliestDate` and `latestDate`, which also define different semantic mappings. The following mapping of the `dates` refers to the case where the `type` attribute of the `name` subelement has the value `personal`, while the `type` attribute of the `dates` subelement gets the value `life`. The basic idea behind the mapping of the element `dates` (and its subelements) in this case is that the `earliestDate` subelement presents the birth date of an agent, while the `latestDate` subelement represents the date of his/her death.

Mapping the `earliestDate` *subelement (when* `@type="life"`*):* In order to map the `earliestDate` subelement, an instance of the class `E67 Birth` is created and related to an instance of the class `E21 Person`, through the property `P98B was born` (denoting the birth event of a person). Then, an instance of the class `E52 Time-Span` is linked to an instance of `E67 Birth`, through the property `P4 has`

time-span, and finally in order to denote the specific date of the birth event, an instance of the class E50 Date is linked to an instance of E52 Time-Span through the property P78 is identified by. Thus, the following CIDOC CRM path:

E24 Physical Man-Made Thing → P108B was produced by →
E12 Production → P14 carried out by → E21 Person →
P98B was born → E76 Birth → P4 has time-span →
E52 Time-Span → P78 is identified by → E50 Date

semantically corresponds to:

/vra/work/agentSet/agent/name[@type="personal"]
/dates[@type="life"]/earliestDate

Mapping the latestDate subelement (when @type="life"): In order to map the latestDate subelement, an instance of the class E69 Death is created and related to an instance of the class E21 Person, through the property P100B died in (denoting the death event of a person). Then, adding as before the path → P4 has time-span → E52 Time-Span → P78 is identified by → E50 Date, we get the following CIDOC CRM path:

E24 Physical Man-Made Thing → P108B was produced by →
E12 Production → P14 carried out by → E21 Person →
P100B died in → E69 Death → P4 has time-span →
E52 Time-Span → P78 is identified by → E50 Date

which semantically corresponds to the VRA path:

/vra/work/agentSet/agent/name[@type="personal"]
/dates[@type="life"]/latestDate

The mappings presented in this section are also shown in Fig. 2.

3.2 The mapping of the agent element expressed in MDL

MDL can be used to formally describe the mapping rules of the elements/ attributes of a source schema to equivalent paths of the target schema. Part of the mapping, containing the rules that map the VRA element agent and its subelements/attributes, is shown in Table 1, expressed in MDL. In this section, a brief analysis of the rules' semantics is presented. For example, Rule R1 states that the /vra/work is mapped to an instance of the class E24. R2 states that the agentSet corresponds to an instance of the class E12, which is linked to E24 through the binary relation P108B. Rules R3, R4, R5 describe the three different versions of the agent element, according to the three possible values of the type attribute of the name subelement, which correspond to the three different subclasses (E21, E40, E74), respectively. It is also important to note here that the variables Y5, Y10 and Y15 on the left part of the rules, as well as the variables J5, J10 and J15 on the right part, denote branching points, that indicate that more than one paths may extend the previous paths (see also Fig. 2). Rules R6, R7, R8, R9 and R10 can be appended to the Rule R3.

RuleNo	VRA paths	CIDOC CRM paths
R1:	/vra/work{X1}	E24{C1}
R2:	$X1/agentSet{Y1}	$C1→P108B→E12{J1}
R3:	$Y1/agent[name/@type="personal"]{Y5}	$J1→P14{$S2$} →E21{J5}
R4:	$Y1/agent[name/@type="corporate"]{Y10}	$J1→P14{$S3$} →E40{J10}
R5:	$Y1/agent[name/@type="family"]{Y15}	$J1→P14{$S4$} →E74{J15}
R6:	$Y5\|$Y10|$Y15/name*	$J5\|$J10|$J15→P131→E82
R7:	$Y5\|$Y10|$Y15/culture*	$J5\|$J10|$J15→P107→E74
R8:	$Y5\|$Y10|$Y15/role*	$S2\|$S3|$S4→P14.1→E55
R9:	$Y5/dates[@type="life"]/earliestDate*	$J5→P98→E67→P4→E52→ P78→E50
R10:	$Y5/dates[@type="life"]/latestDate*	$J5→P100B→E69→P4→E52→ P78→E50

Table 1. Mapping the VRA element `agent` to the CIDOC CRM using MDL.

4 Related work

There is quite an amount of research dealing with ontology-based integration. Amann et al. [1] propose a mechanism for the integration of cultural information resources, by mapping XML fragments to domain specific ontologies, such as CIDOC CRM. In this way, they define a mapping language, which provides a set of rules that describe these resources, relating XPath location paths to the concepts and roles of an ontology. Furthermore, they define a query rewriting algorithm which translates queries executed by users into queries expressed in an XML language and are afterwards sent to XML resources for evaluation. This approach is worth mentioning as it describes a mapping language quite similar to ours and also focuses on the significance of offering mechanisms for representing the semantics of XML data. In [4] XML data are transformed to a global ontology (using the OWL syntax), defining mapping rules that are also based in OWL. In this way, issues of synonymy and structure hierarchy are faced. This work shares common ideas with ours, as it transforms data to a global ontology, although the mapping rules defined in our MDL are not based in OWL syntax.

In [6], an effort is described to integrate the CIDOC CRM ontology in the core model of the BRICKS project. This integration has been accomplished through a mapping scenario applied between the source schemas and the CRM ontology, although a number of issues had to be resolved. Some of them refer to inconsistencies, which mostly originate from the abstractness of some concepts definitions of the CRM [8]. This approach provides mappings that are implemented in spreadsheets, without defining a formal mapping methodology.

5 Conclusions

The mapping methodology presented in this paper is part of an ontology-based metadata integration scenario, where CIDOC CRM acts as a mediating schema among several metadata schemas. More specifically, a semantic mapping from the VRA Core 4.0 standard to the CIDOC CRM ontology is presented.

Mapping VRA elements to CIDOC CRM paths proved to be a rather difficult and time-consuming activity, which required a deep and conceptual work. CIDOC CRM provides very rich structuring mechanisms for metadata descriptions and an abstract but fine-grained conceptualization for events, objects, agents, things, etc. Thus, the combination of this wide range of CRM classes and properties generated a large number of conceptual expressions that should be studied very carefully in order to select the semantically closest one to map to the metadata schemas. Furthermore, the mapping procedure encountered significant obstacles due to the plethora of conceptual expressions that should be aligned. The `type` attribute assigned to several subelements defined different semantic mappings, making mapping even more complex. Finally, it is essentiall to note that the `agent` element and all the related information to the work's production, include the class `E12 Production`, which reveals one of the main characteristics of CIDOC CRM, which is the event-based approach adopted.

Currently, we are investigating the transformation of queries among various cultural heritage metadata schemas and the CIDOC CRM ontology. Our next research steps include the definition of the reverse semantic mappings from the ontology to the VRA Core schema, in order to enrich the mapping procedure proposed by our research group.

References

1. B. Amann, C. Beeri, I. Fundulaki, and M. Scholl. Ontology-Based Integration of XML Web Resources. In *ISWC 2002*, volume 2342 of *LNCS*, pages 117–131. Springer, 2002.
2. C. Kakali and I. Lourdi and T. Stasinopoulou and L. Bountouri and C. Papatheodorou and M. Doerr and M. Gergatsoulis. Integrating Dublin Core Metadata for Cultural Heritage Collections Using Ontologies. In *DC-2007*, pages 128–139, 2007.
3. ICOM/CIDOC CRM Special interest Group. Definition of the CIDOC Conceptual Reference Model, Version 5.0.2, January 2010. Available at http://www.cidoc-crm.org.
4. P. Lehti and P. Fankhauser. XML Data Integration with OWL: Experiences and Challenges. In *SAINT 2004*, pages 160–170. IEEE Computer Society, 2004.
5. Library of Congress (LC). VRA Core: a Data Standard for the Description of Works of Visual Culture, 2011. Available at http://www.loc.gov/standards/vracore/.
6. C. Meghini and T. Risse. BRICKS: A Digital Library Management System for Cultural Heritage. *ERCIM News*, (61), April 2005.
7. N. F. Noy. Semantic Integration: a Survey of Ontology-Based Approaches. In *SIGMOD Record*, volume 33, 2004.
8. P. Nussbaumer and B. Haslhofer. CIDOC CRM in Action - Experiences and Challenges. In *ECDL 2007*, volume 4675 of *LNCS*, pages 532–533. Springer, 2007.
9. T. Stasinopoulou and L. Bountouri and I. Lourdi and C. Papatheodorou and M. Doerr and M. Gergatsoulis. Ontology-Based Metadata Integration in the Cultural Heritage Domain. In *ICADL 2007*, volume 4822 of *LNCS*, pages 165–175. Springer, 2007.

10. VRA Core Oversight Committee. VRA Core 4.0 Element Description and Tagging Examples, 2007. Available at http://www.loc.gov/standards/vracore/schemas.html.
11. World Wide Web Consortium (W3C). XML Path Language (XPath) 2.0. Available at http://www.w3.org/TR/xpath20/, 2007.

Developing a Formal Model for Mind Maps

Vasilis Siochos, Christos Papatheodorou
Database & Information System Group, Laboratory of Digital Libraries and Electronic Publications, Department of Archives and Library Sciences, Ionian University, Corfu, Greece
{vsiochos, papatheodor}@ionio.gr

Abstract. Mind map is a graphical technique, which is used to represent words, concepts, tasks or other connected items or arranged around central topic or idea. Mind maps are widely used, therefore exist plenty of software programs to create or edit them, while there is none format for the model representation, neither a standard format. This paper presents and effort to propose a formal mind map model aiming to describe the structure, content, semantics and social connections. The structure describes the basic mind map graph consisted of a node set, an edge set, a cloud set and a graphical connections set. The content includes the set of the texts and objects linked to the nodes. The social connections are the mind maps of other users, which form the neighborhood of the mind map owner in a social networking system. Finally, the mind map semantics is any true logic connection between mind map textual parts and a concept. Each of these elements of the model is formally described building the suggested mind map model. Its establishment will support the application of algorithms and methods towards their information extraction.

Keywords: mind map, knowledge organization, Web 2.0

1 Introduction

According to Buzan [1], the mind map is an expression of radiant thinking. It is used to represent graphically words, ideas, tasks, or other items linked to and arranged around a central key word or idea [2]. It is obvious that mind maps contain information, in the nodes, in the linked objects and in their structure. However, there are no formal rules on how to build a mind map, in order to express the creativity of the mind. Therefore a mind map differs from an ontology. Moreover up to today there is none standard model or at least a common file format for mind maps encoding followed by the variety of software helping the mind map development.

In order to apply information retrieval method or algorithms on mind maps, a formal model to define what exactly a mind map consists of, is necessary. Developing a formal mind map model, we propose the basic aspects of structure, content and social connections and plan the future semantics description of mind maps.

In the next section, related work about the use of mind maps in information retrieval is presented. Afterwards the basic aspects of the mind map model are presented, and finally the future directions on how we will describe the semantics of mind maps are discussed.

2 Related Work

Several ideas about the use of mind maps are currently under study. Beel,Geep and Stiller [3] explore whether data extracted from mind maps could be used to enhance information retrieval, denoting that the structure of mind map embeds a kind of semantic connections. Also a mind map can be used to define relations between documents linked in a mind map [4]. According to the researchers this process is similar to analyzing emails or other linked documents [5].

Furthermore, mind maps as a visualization tool can be used to enhance expert search document summarization, keyword based search engines, document recommender systems and determining word relatedness [3, 6]. Finally recently, mind maps have been used to model XML DTD's, XML schemas and XML documents [7].

3 The Mind Map Features

Definition: A mind map MM is a pair MM=<S,C>, where S is the structure and C is the content.

3.1 Structure

Definition 3.1.1: The structure S of a mind map is a 4-tupe S=<N,E,C,GC> where N are the nodes, E the Edges, C the clouds and GC the graphical connectors.

Definition 3.1.2: Each N_i belonging to the set of nodes N is a 5-tuple N=<T, nID, R, Frm>, where T is the node name, nID is the node ID, R are the resources, and Frm is a 7-tuple of numbers (denoting formatting values), Frm=<x_1,x_2,x_3,x_4,x_5,x_6,x_7> where x_i are the program defined values for each formatting values.

A node besides text can contain an image, an URI and LaTeX code. In the case of URI the node is a terminal node of the graph.

Definition 3.1.3: A resource R on a mind map is any text, image, URI, LaTeX and attribute value added on the mind map nodes. As mentioned, in case the resource is an URI then the node is terminal.

Definition 3.1.4: The attributes A is a pair, A=(a_i,b_i)$\subseteq$R, where a_i,b_i are user defined attribute-value pairs.

In some mind map software LaTeX is supported as content of the nodes. The tuple of an attribute can be used to add metadata or tags to a node. The metadata element can be assigned to a_i and the value to b_i.

Definition 3.1.5: The set of edges of a mind map E, is the 5-tuple E_i=<nID_i, nID_j, FmtCd, hid, EL>. nID_i, nID_j are the connected nodes IDs, FrmCd is the edge format

code, hid is a boolean parameter of hidden and EL is a relational operator value "is a" or "<>" different.

Generally the mind map's edges denote an undefined relation between two nodes. EL describes the option some software provides to assign a relational operator value to edges.

Definition 3.1.6: Each cloud Cl_i is member of the set of clouds Cl, and is defined as a connected subgraph of a mind map.

Definition 3.1.7: The Graphical Connectors set GC, is defined as a triple, $GC_i=<nID_i,nID_j,V>$, where nID_i, nID_j are the id's of the connected nodes and V a set of tags tagging a connector.

A graphical connector is a connection between two nodes, which belong to different subgraphs of the mind map. The graphical connectors do not imply a hierarchy between nodes and can be directed.

3.2 Content

The content C of a mind map is considered as a set of resources, which could be text, images, sound, video, hyperlink, spreadsheet, date and binary file. The content is attached to each node of a mind map. In some mind map software LaTeX is supported as content of the nodes.

Definition 3.2.1: Content C of a mind map MM is the set of all the resources R on the map.

3.3 Semantics

As a way of expressing radiant thinking, mind maps contain concepts connected in many undefined ways. In a formal model as described above, semantics can be defined between the concepts in the textual parts of mind map.

Definition 3.3.1: Semantics on a mind map is a function f: $K \rightarrow c$, where K is the powerset of the textual sets of the mind map and c is a concept.

The semantics of a mind map is an issue for further study, aiming to represent explicitly the knowledge (of a domain or a workflow) that a mind map carries. For this purpose the semantics of higher order logics will be studied and exploited in the proposed model. Figure 1 presents a mind map that concludes the concepts of the proposed model.

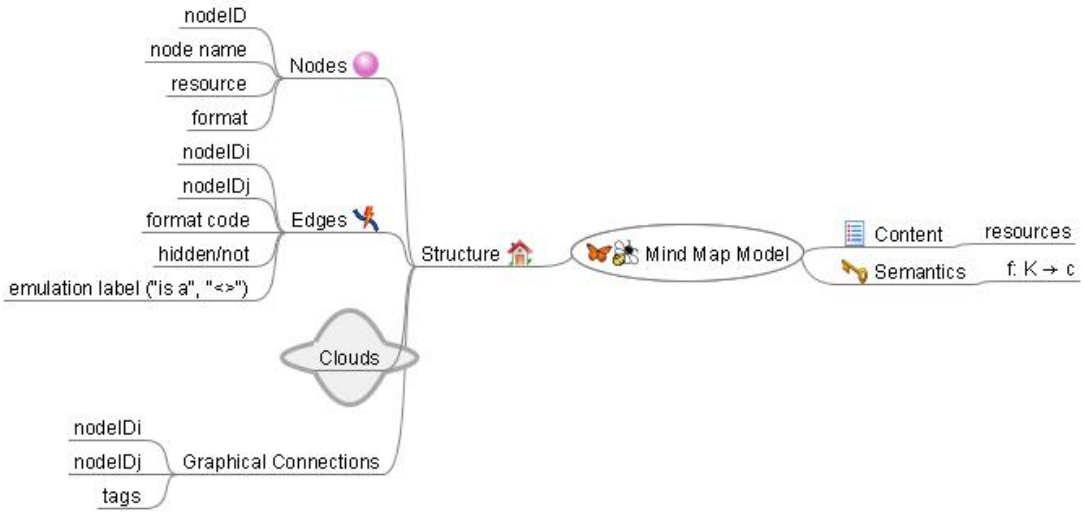

Fig. 1. The Mind Maps Features

4 Social Connections

According to the bibliography, mind maps can be used by social networking applications, to depict user interests, profiles and reflecting attitudes in performing tasks and workflows. Therefore there is the need for the definition of features that might affect the structural and content characteristics of a specific mind map, as well as its creation process.

In a mind map library a user can share his mind maps with other users, tag and organize them. Therefore, a user develops a folksonomy to tag his mind maps. This folksonomy might overlap with other users' folksonomies, reflecting their common interests.

Definition 4.1.1: User mind maps MM_u is the collection of the mind maps of user u.

Definition 4.1.2: User folksonomy Flk is the set of tags, $Flk_u=\{tag_1,tag_2,\ldots,tag_n\}$, the user tagged all the mind maps of his collection.

Definition 4.1.3: Mind map tags MM_{tags} is the set of tags, $MM_{tags}=\{tag_1,tag_2,\ldots,tag_n\}$, where $(tag,tag_i)\in A$ for i = 1,2,…,n, are the tags the users tagged the mind map nodes.

Definition 4.1.4: User's friends mind maps is the set $MM_{UF}=\{MM_1,MM_2,\ldots,MM_n\}$, where MM_i, i=1,2,…,n, are the mind maps of user's friends.

Definition 4.1.5: User's F_1 folksonomy expansion F_{1e} through the folksonomy of user F_2 is the set F_2-($F_1 \cap F_2$).

Definition 4.1.6: A user's U_1 recommended friends RU_u is the set of users $RU_U=\{U_1,U_2,\ldots,U_n\}$, where U_i, i=1,2,…,n, are the users with at least one similar mind map with the user U_1.

Definition 4.1.7: A user U_1, with folksonomy F_1, is a common friend to user U_2, with folksonomy F_2, if a user U, exists with folksonomy F_u, where $(F_1-(F_1 \cap F_U)) \cap (F_2-(F_2 \cap F_U)) \neq \varnothing$.

The crucial concept for the complete definition of the social features of a mind map and in particular the definition of the concept "recommended friend" is that of "mind map similarity". Even though the concepts "friend" and "common friend" denote the observed overlap between the folksonomies of two users, the similarity between two mind maps is a more general concept that incorporates the structural similarity of them as well as the semantic similarity of their content.

5 Conclusions and Further Research

Mind maps are becoming a popular tool for the representation of user interests, customs and tasks and therefore it is considered a suitable tool for defining user models. Hence, the proposed model aims to reveal and define the main characteristics of the mind map. The issue on which we will focus in the future is the integration of the mentioned features so that to derive a model for measuring the similarity between two mind maps. As mentioned, the first step for this direction is the study of the semantics of a specific mind map and how they could be compared with the semantics of a mind map collection.

References

1. Buzan, T. *The Mind Map Book.* Penguin Books, 1996.
2. Wikipedia. *Mind Maps*. http://en.wikipedia.org/wiki/Mind_maps, 2011.
3. Beel, J., Gipp, B., Stiller, J.O. (2009). Information Retrieval on Mind Maps – What could it be good for? In *Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom'09), Washington (USA), November 2009*, pp. 1-4. IEEE.
4. Beel, J., Gipp,B. (2010). Enhancing Information Search by Utilizing Mind Maps. In *Proceedings of the 21th ACM Conference on Hyptertext and Hypermedia. ACM, June 2010*.
5. Beel, J., Gipp, B., Stiller, J.O. (2009). Could Mind Maps Be Used To Improve Academic Search Engines? In *International Conference on Machine Learning and Data Analysis (ICMLDA'09)*, Lecture Notes in Engineering and Computer Science Vol. 2, pp. 832–834, Berkeley: International Association of Engineers (IAENG), Newswood Limited.

6. Theodore Dalamagas, Tryfon Farmakakis, Manolis Maragkakis, Artemis G. Hatzigeorgiou, FreePub: Collecting and Organizing Scientic Material Using Mindmaps, In *Proceedings of the Semantic Web Applications and Tools for Life Sciences Workshop (SWAT4LS'10), Dec 8-10, Berlin, Germany*, http://web.imis.athena-innovation.gr/~dalamag/pub/dfmh-swat4ls10.pdf
7. Bia, A., Muñoz, R., Gómez, J. Using mind maps to model semistructured documents. In *Proceedings 14th European Conference on Digital Libraries (ECDL 2010), September 2010, Glasgow, UK*, Springer LNCS 6273, pp. 421–424.

MXML Storage and the Problem of Manipulation of Context

Nikolaos Fousteris[1], Manolis Gergatsoulis[1], and Yannis Stavrakas[2]

[1] Database & Information Systems Group (DBIS),
Laboratory on Digital Libraries and Electronic Publishing,
Department of Archives and Library Science, Ionian University,
Ioannou Theotoki 72, 49100 Corfu, Greece.
{nfouster,manolis}@ionio.gr,
[2] Institute for the Management of Information Systems (IMIS),
R. C. Athena,
G. Mpakou 17, 11524, Athens, Greece.
yannis@imis.athena-innovation.gr

Abstract. The problem of storing and querying XML data using relational databases has been considered a lot and many techniques have been developed. MXML is an extension of XML suitable for representing data that assume different facets, having different value and structure under different contexts, which are determined by assigning values to a number of dimensions. In this paper, we explore techniques for storing MXML documents in relational databases, based on techniques previously proposed for conventional XML documents. Essential characteristics of the proposed techniques are the capabilities a) to reconstruct the original MXML document from its relational representation and b) to express MXML context-aware queries in SQL.

1 Introduction

The problem of storing XML data in relational databases has been intensively investigated [4, 10, 11, 13] during the past 10 years. The objective is to use an RDBMS in order to store and query XML data. First, a relational schema is chosen for storing the XML data, and then XML queries, produced by applications, are translated to SQL for evaluation. After the execution of SQL queries, the results are translated back to XML and returned to the application.

Multidimensional XML (MXML) is an extension of XML which allows context specifiers to qualify element and attribute values, and specify the contexts under which the document components have meaning. MXML is therefore suitable for representing data that assume different facets, having different value or structure, under different contexts. Contexts are specified by giving values to one or more user defined dimensions. In MXML, dimensions may be applied to elements and attributes (their values depend on the dimensions). An alternative solution would be to create a different XML document for every possible combination, but such an approach involves excessive duplication of information.

In this paper, we present two approaches for storing MXML in relational databases, based on XML storage approaches. We use MXML-graphs, which are graphs using appropriate types of nodes and edges, to represent MXML documents. In the first (naive) approach, a single relational table is used to store all information about the nodes and edges of the MXML-graph. Although simple, this approach presents some drawbacks, like the large number of expensive self-joins when evaluating queries. In the second approach we use several tables, each of them storing a different type of nodes of the MXML-graph. In this way the size of the tables involved in joins is reduced and consequently the efficiency of query evaluation is enhanced. Both approaches use additional tables to represent context in a way that it can be used and manipulated by SQL queries. Additionally to MXML storage, we propose techniques for context manipulation, as context is one of the major characteristics of MXML.

2 Preliminaries

2.1 Mutidimensional XML

In MXML, data assume different facets, having different value or structure, under different contexts according to a number of *dimensions* which may be applied to elements and attributes [7, 8]. The notion of "world" is fundamental in MXML. A world represents an environment under which data obtain a meaning. A *world* is determined by assigning to every dimension a single value, taken from the domain of the dimension. In MXML we use syntactic constructs called *context specifiers* that specify sets of worlds by imposing constraints on the values that dimensions can take. The elements/attributes that have different facets under different contexts are called *multidimensional elements/attributes*. Each multidimensional element/attribute contains one or more facets, called *context elements/attributes*, accompanied with the corresponding context specifier which denotes the set of worlds under which this facet is the holding facet of the element/attribute. The syntax of MXML is shown in Example 1, where a MXML document containing information about a book is presented.

Example 1. The MXML document shown below represents a book in a book store. Two dimensions are used namely `edition` whose domain is {`greek, english`}, and `customer_type` whose domain is {`student, library, teacher`}.

```
<book isbn=[edition=english]"0-13-110362-8"[/]
           [edition=greek]"0-13-110370-9"[/]>
  <title>The C programming language</title>
  <authors>
     <author>Brian W. Kernighan</author>
     <author>Dennis M. Ritchie</author>
  </authors>
  <@publisher>
     [edition = english] <publisher>Prentice Hall</publisher>[/]
     [edition = greek] <publisher>Klidarithmos</publisher>[/]
```

```
  </@publisher>
  <@translator>
     [edition = greek] <translator>Thomas Moraitis</translator>[/]
  </@translator>
  <@price>
     [edition=english]<price>15</price>[/]
     [edition=greek,customer_type in {student, teacher}]<price>9</price>[/]
     [edition=greek,customer_type=library]<price>12</price>[/]
  </@price>
  <@cover>
     [edition=english]<cover><material>leather</material></cover>[/]
     [edition=greek]
        <cover>
           <material>paper</material >
           <@picture>
             [customer_type=student]<picture>student.bmp</picture>[/]
             [customer_type=library]<picture>library.bmp</picture>[/]
           </@picture>
         </cover>
      [/]
  </@cover>
</book>
```

Notice that multidimensional elements (see for example the element `price`) are the elements whose name is preceded by the symbol `@` while the corresponding context elements have the same element name but without the symbol `@`.

A MXML document can be considered as a compact representation of a set of (conventional) XML documents, each of them holding under a specific world. For the extraction of XML documents holding under specific worlds the interested reader may refer to [7] where a related process called *reduction* is presented.

2.2 Storing XML data in relational databases

Many researchers have investigated how an RDBMS can be used to store and query XML data. Work has also been directed towards the storage of temporal extensions of XML [16, 1, 2]. The techniques proposed for XML storage can be divided in two categories, depending on the presence or absence of a schema:

1. *Schema-Based XML Storage techniques*: the objective here is to find a relational schema for storing a XML document, guided by the structure of a schema for that document [9, 13, 5, 15, 10, 3, 11].
2. *Schema-Oblivious XML Storage techniques*: the objective is to find a relational schema for storing XML documents independent of the presence or absence of a schema [13, 5, 15, 17, 10, 6, 4].

The approaches that we propose in this paper do not take schema information into account, and therefore belong to the Schema-Oblivious category.

3 Properties of MXML documents

3.1 A graphical model for MXML

In this section we present a graphical model for MXML called *MXML-graph*. The proposed model is node-based and each node is characterized by a unique "id". In MXML-graph, except from a special node called *root node*, there are the following node types: *multidimensional element nodes*, *context element nodes*, *multidimensional attribute nodes*, *context attribute nodes*, and *value nodes*. The *context element nodes*, *context attribute nodes*, and *value nodes* correspond to the element nodes, attribute nodes and value nodes in a conventional XML graph. Each multidimensional/context element node is labelled with the corresponding element name, while each multidimensional/context attribute node is labelled with the corresponding attribute name. As in conventional XML, value nodes are leaf nodes and carry the corresponding value. The facets (context element/attribute nodes) of a multidimensional node are connected to that node by edges labelled with context specifiers denoting the conditions under which each facet holds. These edges are called *element/attribute context edges* respectively. Context elements/attributes are connected to their child elements/attribute or value nodes by edges called *element/attribute/value edges* respectively. Finally, the context attributes of type IDREF(S) are connected to the element nodes that they point to by edges called *attribute reference edges*.

Example 2. In Fig. 1, we see the representation of the MXML document of Ex-

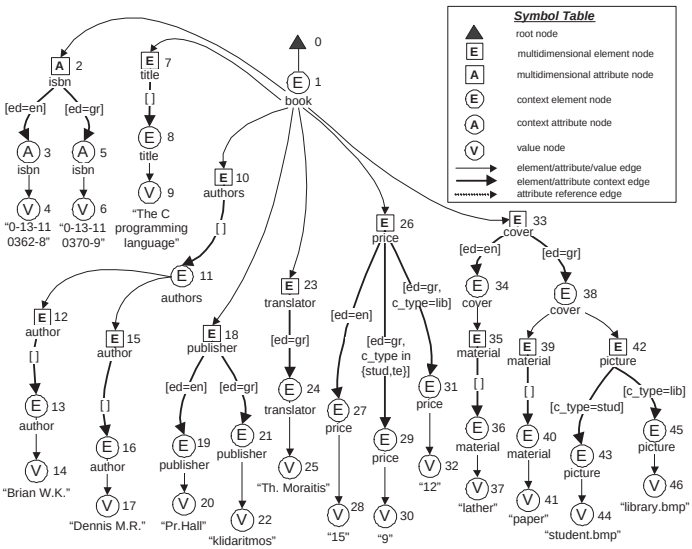

Fig. 1. Graphical representation of MXML (MXML tree)

ample 1 as a MXML-graph. Note that some additional multidimensional nodes (e.g. nodes 7 and 10) have been added to ensure that the types of the edges alternate consistently in every path of the graph. This does not affect the information contained in the document, but facilitates the navigation in the graph and the formulation of queries. For saving space, in Fig. 1 we use obvious abbreviations for dimension names and values that appear in the MXML document.

3.2 Properties of contexts

Context specifiers qualifying element/attribute context edges give the *explicit contexts* of the nodes to which these edges lead. The explicit context of all the other nodes of the MXML-graph is considered to be the *universal context* [ ], denoting the set of all possible worlds. The explicit context can be considered as the true context only within the boundaries of a single multidimensional element/attribute. When elements and attributes are combined to form a MXML document, the explicit context of each element/attribute does not alone determine the worlds under which that element/attribute holds, since when an element/attribute e_2 is part of another element e_1, then e_2 have substance only under the worlds that e_1 has substance. This can be conceived as if the context under which e_1 holds is inherited to e_2. The context propagated in that way is combined with (constraint by) the explicit context of a node to give the *inherited context* for that node. Formally, the inherited context $ic(q)$ of a node q is defined as $ic(q) = ic(p) \cap^c ec(q)$, where $ic(p)$ is the inherited context of its parent node p. $\cap^c$ is an operator called *context intersection* defined in [12] which combines two context specifiers and computes a new context specifier which represents the intersection of the worlds specified by the original context specifiers. The evaluation of the inherited context starts from the root of the MXML-graph. By definition, the inherited context of the root of the graph is the universal context [ ]. Note that contexts are not inherited through attribute reference edges.

As in conventional XML, the leaf nodes of MXML-graphs must be value nodes. The *inherited context coverage* of a node further constraints its inherited context, so as to contain only the worlds under which the node has access to some value node. This property is important for navigation and querying, but also for the reduction process [7]. The inherited context coverage $icc(n)$ of a node n is defined as follows: if n is a leaf node then $icc(n) = ic(n)$; otherwise $icc(n) = icc(n_1) \cup^c icc(n_2) \cup^c ... \cup^c icc(n_k)$, where $n_1, \ldots, n_k$ are the child element nodes of n. $\cup^c$ is an operator called *context union* defined in [12] which combines two context specifiers and computes a new one which represents the union of the worlds specified by the original context specifiers. The inherited context coverage gives the true context of a node in a MXML-graph.

4 Storing MXML in relational databases

In this section we present two approaches for storing MXML documents using relational databases.

4.1 Naive Approach

The first approach, called *naive approach*, uses a single table (*Node Table*), to store all information contained in a MXML document. Node Table contains all the information which is necessary to reconstruct the MXML document(graph). Each row of the table represents a MXML node. The attributes of Node Table are: `node_id` stores the id of the node, `parent_id` stores the id of the parent node, `ordinal` stores a number denoting the order of the node among its siblings, `tag` stores the label (tag) of the node or NULL (denoted by "-") if it is a value node, `value` stores the value of the node if it is a value node or NULL otherwise, `type` stores a code denoting the node type (CE for context element, CA for context attribute, ME for multidimensional element, MA for multidimensional attribute, and VN for value node), and `explicit_context` stores the explicit context of the node (as a string). Noted that the explicit context is kept here for completeness, and does not serve any retrieval purposes. In the following we will see how the correspondence of nodes to the worlds under which they hold is encoded.

Example 3. Fig. 2 shows how the MXML Graph of Fig. 1 is stored in the Node Table. Some of the nodes have been omitted, denoted by "....", for brevity.

Node Table						
node_id	parent_id	ordinal	tag	value	type	explicit_context
1	0	1	book	-	CE	-
2	1	1	isbn	-	MA	-
3	2	1	isbn	-	CA	[ed=en]
4	3	1	-	0-13-110362-8	VN	-
5	2	2	isbn	-	CA	[ed=gr]
6	5	1	-	0-13-110370-9	VN	-
7	1	2	title	-	ME	-
8	7	1	title	-	CE	[ ]
9	8	1	-	The C progr. lang.	VN	-
....	....	....	....	....	....	....
43	42	1	picture	-	CE	[c_type=stud]
....	....	....	....	....	....	....

Fig. 2. Storing the MXML-graph of Fig. 1 in a Node Table.

4.2 Limitations of the Naive Approach

The naive approach is straightforward, but it has some drawbacks mainly because of the use of a single table. As the different types of nodes are stored in the table, many NULL values appear in the fields `explicit_context`, `tag`, and `value`. Those NULL values could be avoided if we used different tables for different node types. Moreover, as we showed in Subsection 4.1, queries on MXML

documents involve a large number of self-joins of the Node Table, which is anticipated to be a very long table since it contains the whole tree. Splitting the Node Table would reduce the size of the tables involved in joins, and enhance the overall performance of queries. Finally, notice that the context representation scheme we introduced leads to a number of joins in the nested query. Probably a better scheme could be introduced that reduces the number of joins.

4.3 A Better Approach

In the *Type Approach* presented here, MXML nodes are divided into groups according to their type. Each group is stored in a separate table named after the type of the nodes. In particular *ME Table* stores multidimensional element nodes, *CE Table* stores context element nodes, *MA Table* stores multidimensional attribute nodes, *CA Table* stores context attribute nodes, and *Value Table* stores value nodes. The schema of these tables is shown in Fig. 3. Each row in these

ME Table

node_id	parent_id	ordinal	tag
7	1	2	title
10	1	3	authors
....	....	....	....

CE Table

node_id	parent_id	ordinal	tag	explicit_ context
1	0	1	book	-
8	7	1	title	[ ]
....	....	....	....	....
19	18	1	publisher	[ed=en]
21	18	2	publisher	[ed=gr]
....	....	....	....	....

MA Table

node_id	parent_id	ordinal	tag
2	1	1	isbn

CA Table

node_d	parent_id	ordinal	tag	explicit_context
3	2	1	isbn	[ed=en]
5	2	2	isbn	[ed=gr]

Value Table

node_id	parent_id	value
4	3	0-13-110362-8
6	5	0-13-110362-9
9	8	The C programming language
....	....	....

Fig. 3. The Type tables.

tables represents a MXML node. The attributes in the tables have the same meaning as the respective attributes of the Node Table. Using this approach we tackle some of the problems identified in the previous section. Namely, we eliminate NULL values and irrelevant attributes, while at the same time we reduce the size of the tables involved in joins when navigating the MXML-Graph.

5 Context Representation

In this section we present techniques that help us to store the context in such a way so as to facilitate the formulation of context-aware queries. Two approaches, for storing context in a Relational Database, are presented. The first, is a naive representation and the second one is called the Ordered-Based representation.

5.1 Naive Context Representation

For the Naive Context Representation technique, we introduce three additional tables, as shown in Fig. 4. The *Possible Worlds Table* which assigns a unique ID (attribute `word_id`) to each possible combination of dimension values. Each dimension in the MXML document has a corresponding attribute in this table. The *Explicit Context Table* keeps the correspondence of each node with the worlds represented by its explicit context. Finally, the *Inherited Coverage Table* keeps the correspondence of each node with the worlds represented by its inherited context coverage.

Example 4. Fig. 4, depicts (parts of) the Possible Worlds Table, the Explicit Context Table, and the Inherited Coverage Table obtained by encoding the context information appearing in the MXML-graph of Fig. 1. For example, the

Possible Worlds Table		
world_id	edition	customer_type
1	gr	stud
2	gr	lib
3	gr	te
4	en	stud
5	en	lib
6	en	te

Explicit Context Table	
node_id	world_id
1	1
1	2
1	3
1	4
1	5
1	6
....	....
5	1
5	2
5	3
6	1
6	2
6	3
6	4
6	5
6	6
....	....

Inherited Coverage Table	
node_id	world_id
1	1
1	2
1	3
1	4
1	5
1	6
....	....
5	1
5	2
5	3
6	1
6	2
6	3
....	....

Fig. 4. Mapping MXML nodes to worlds.

inherited context coverage of the node with `node_id=6` includes the worlds:

$w_1 = \{$`(edition, greek)`, `(customer_type, student)`$\}$,
$w_2 = \{$`(edition, greek)`, `(customer_type, library)`$\}$ and
$w_3 = \{$`(edition, greek)`, `(customer_type, teacher)`$\}$

This is encoded in the Inherited Coverage Table as three rows with `node_id=6` and the world ids 1, 2 and 3. In the Explicit Context Table the same node corresponds to all possible worlds (ids 1, 2, 3, 4, 5 and 6).

5.2 Ordered-Based Context Representation

According to the Ordered-Based Context Representation technique, we propose a scheme that reduces the size of tables and the number of joins in context-driven queries. The basic idea of this technique is that we achieve the total ordering of all possible worlds based on a) a total ordering of dimensions and b) a total ordering of dimension possible values. So, for k dimensions with each dimension i having m_i possible values, we may have $n = m_1 * m_2 * \ldots * m_k$ possible ordered worlds. Each of these worlds is assigned a unique integer value between 1 and n.

Example 5. In Fig. 5, we present how it is possible to order all possible worlds according to the dimensions and the dimension values of Example 1. In order

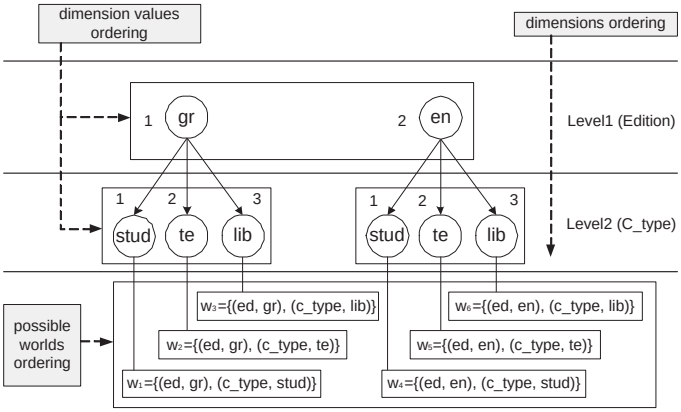

Fig. 5. Possible Worlds Ordering

to show this ordering, we use a forest of trees. As we can see, each dimension of the MXML document corresponds to a level in the forest. The ordering of these levels represents the ordering of dimensions. Also, for each level we can see the ordering of all possible values of the related dimension, under each node of the previous level. Each possible world can be produced by traversing a path from a root node of the forest to a leaf node of the corresponding tree. Finally, the order of the forest's leaves represents the total ordering of all possible worlds assigning a unique integer to each world $(w_1, w_2, \ldots, w_6)$.

Assuming that all possible worlds of a MXML document are totaly ordered, we define a vector of binary digits called World Vector.

Definition 1. *Given a total ordering of worlds* $W = (w_1, w_2, \ldots, w_n)$*, where* n *is the number of possible worlds, we define as* $V(c) = (a_1, a_2, \ldots, a_n)$ *the World Vector of a context specifier* c*, where* a_i *with* $i = 1, 2, \ldots, n$*, is a one bit value containing* 1 *if the world* w_i *is between the worlds represented by* c *or* 0 *if* w_i *is not included in the worlds represented by* c*.*

In Fig. 6 we can see how in general we can store dimensions' information to the Relational Database. One table (Table D) is used for storing ordered dimensions and one separate table D_i with $i = 1, 2, \ldots, k$ is used for storing the ordered values $d_{i,j}$ with $j = 1, 2, \ldots, m_i$ and m_i is the number of the different values of dimension D_i.

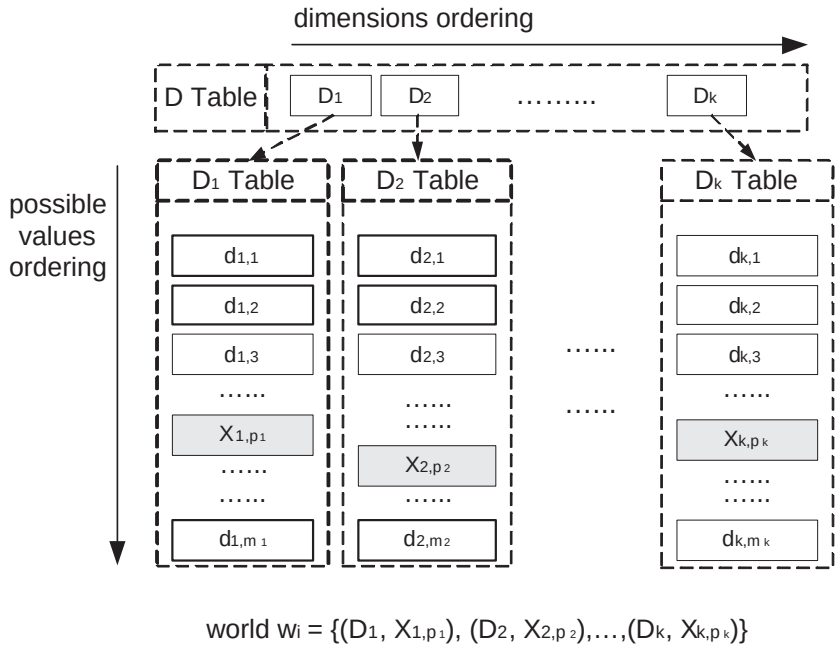

Fig. 6. Ordered-Based Representation in Relational Schema

5.2.1 Finding the position of a world in a World Vector A problem which arise when using the Ordered-Based Representation to represent worlds, is the problem of defining the position corresponding to a specific world in a world vector. Assuming that a context specifier contains the world w_i, shown in Fig. 6, we can find the bit-position i corresponding to this world in the world

D Table	
dimension_id	dimention_name
1	edition
2	customer_type

D_1 Table	
value_id	value
1	greek
2	english

D_2 Table	
value_id	value
1	student
2	teacher
3	library

Inherited Coverage Table	
node_id	world_vector
1	111111
2	111111
3	000111
4	000111
5	111000
6	111000
....	....

Explicit Context Table	
node_id	world_vector
1	111111
2	111111
3	000111
....	....
31	001000
....	....
43	100100
....	....

Fig. 7. Context Tables.

vector of the context specifier, using the following formula:

$$i = p_k + \sum_{j=2}^{k}[(p_{j-1} - 1) * (\prod_{w=j}^{k} m_w)]$$

Example 6. Fig. 7, depicts (parts of) the Explicit Context Table, and the Inherited Coverage Table obtained by encoding the context information appearing in the MXML-graph of Fig. 1. Also, we can see the contents of the tables D, D_1 and D_2 containing the ordering information for all possible worlds. For example, the explicit context of the node with `node_id=3` includes the worlds:

$w_1 = \{$`(edition, english)`, `(customer_type, student)`$\}$,
$w_2 = \{$`(edition, english)`, `(customer_type, teacher)`$\}$ and
$w_3 = \{$`(edition, english)`, `(customer_type, library)`$\}$

According to the ordering of Fig. 5, the bit-positions of these worlds in the world vector are 4, 5 and 6 respectively. As a result, the explicit context specifier of the node is encoded in the Explicit Context Table as one row with `node_id=3` and the world vector 000111.

5.2.2 Finding the world corresponding to a bit in a World Vector

The opposite problem of finding the position of a world in a world vector is the problem of finding which world corresponds to a bit-position i of a world vector. In order to achieve this, we can use the algorithm represented by the flowchart shown in Fig. 8, using the notation of Fig. 5. The algorithm of Fig. 8 takes as input the i position of a world in a world vector. The output of the algorithm is a sequence of numbers $(p_1, p_2, \ldots, p_k)$. Each number p_i represents the position of a value among the ordered values of dimension D_i. Using this position, it is

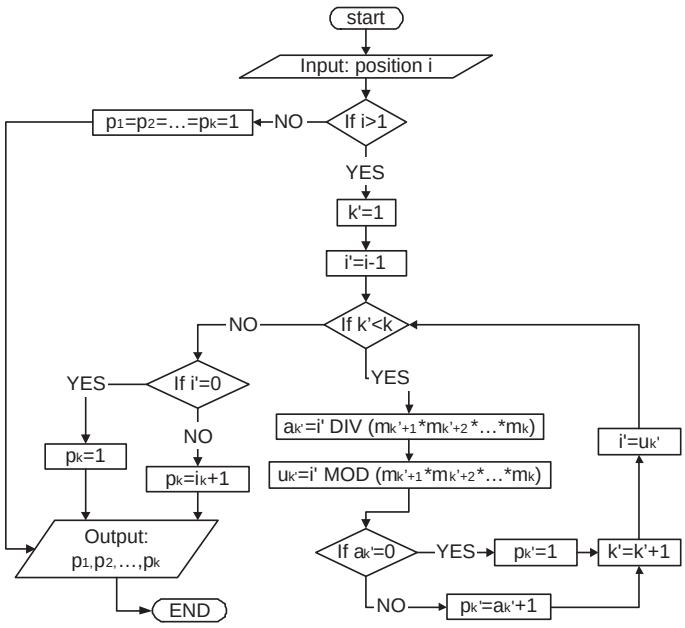

Fig. 8. Converting bit-position i of world vector to world

possible to find the value X_{i,p_i} of the dimension D_i from the appropriate table D_i of Fig. 5. A The set of pairs (D_i, X_{i,p_i}) represents the resulting world.

6 Querying MXML with Multidimensional XPath

In this section we present *Multidimensional XPath* (MXPath) as an extension of XPath used to navigate through MXML-graphs. In addition to the conventional XPath functionality, MXPath uses the inherited context coverage and the explicit context of MXML in order to select nodes in the MXML document. Similarly to XPath, MXPath uses *path expressions* as a sequence of steps to get from one MXML node to another node, or set of nodes.

In a MXPath, selection criteria concerning the explicit context are expressed through *explicit context qualifiers*. Selection criteria concerning the inherited context coverage are expressed through the *inherited context coverage qualifier*, which is placed at the beginning of the expression.

6.1 MXPath Syntax

An *MXPath expression* contains an *inherited context coverage qualifier* (or *icc qualifier* for short) followed by the *MXPath expression body*. The inherited context coverage qualifier is placed at the beginning of the expression and filters the

resulting nodes according to their inherited context coverage. The syntax of an MXPath expression is:

`[inherited_context_coverage_qualifier],MXPath_expression_body`

An MXPath expression may return either multidimensional nodes or context nodes. In what follows we brake down MXPath expressions, and specify each part separately.

6.1.1 Inherited context coverage qualifier The syntax of the *inherited context coverage qualifier* is:

`icc() comparison_op context_specifier_expression`

where `comparison_op` is one of the operators `=`, `!=`, `<`, `>`, `<=`, or `>=`. Note that it is easy to prove that for the inherited context coverages of the nodes in a path $r, n_1, \ldots, n_k$, from the root r of the MXML tree to a node n_k, it holds that $icc(n_k) \subseteq icc(n_{k-1}) \subseteq \ldots \subseteq icc(r)$. Thus $icc(n_k)$ denotes the worlds under which the complete path holds. The function `icc()` returns the icc of the current node, and, consequently of the currently evaluated path in MXML. This icc is then compared against the *context_specifier*, according to the comparison operator. The operator `=` tests for equality, `<` tests for proper subset, `>` for proper superset, etc. Note that it is actually the sets of worlds represented by the contexts that are compared. In case the comparison returns *false*, the current path is rejected and not considered further. If the inherited context coverage qualifier is omitted in an MXPath expression, the default is implied: `icc() >= "-"`, which evaluates always to *true*.

6.1.2 MXPath expression body *MXPath expression body* corresponds to (conventional) XPath expressions. As in XPath, in MXPath we also have two types of expression bodies, namely the *absolute* and the *relative*. An absolute MXPath expression body is a relative one preceded by the symbol "`/`" which denotes the root of the MXML tree. `MXPath_expression_body` is composed by one of more *MXPath steps* separated by "`/`". Thus, the syntax of a relative `MXPath_expression_body` is of the form:

`MXPath_step_1/MXPath_step_2/.../MXPath_step_n`

6.1.3 MXPath steps There are two types of MXPath steps, namely, the *Context MXPath steps* which return context nodes, and the *Multidimensional MXPath steps* which return multidimensional nodes. The syntax of a Context MXPath step is as follows:

`axis::node_test[pred_1][pred_2]...[pred_n]`

while the syntax of a Multidimensional MXPath step is as follows:

`axis->node_test[pred_1][pred_2]...[pred_n]`

Notice that, both types of MXPath steps contain an *axis*, a *node test* and zero or more *predicates*. The only difference is that in a context MXPath step the axis is followed by the symbol "`::`" which denotes that the step evaluates to

context nodes, while in a Multidimensional MXPath step axis is followed by the symbol "`->`" which denotes that the step evaluates to multidimensional nodes.

6.1.4 MXPath predicates In MXPath a *predicate* consists of an expression, called a *MXPath predicate expression*, enclosed in square brackets. A predicate serves to filter a sequence, retaining some items and discarding others. Multiple predicates are allowed in MXPath expressions. In the case of multiple adjacent predicates, the predicates are applied from left to right, and the result of applying each predicate serves as the input sequence for the following predicate. For each item in the input sequence, the predicate expression is evaluated and a truth value is returned. The items for which the truth value of the predicate is *true* are retained, while those for which the predicate evaluates to *false* are discarded. The operators (logical operators, comparison operators, etc.) used in MXPath predicates are those used in conventional XPath. MXPath predicates may also contain MXPath expression bodies in the same way as XPath expressions are allowed in conventional XPath predicates. Besides these syntactic constructs, *explicit context qualifiers* (or *ec qualifiers*) are also used in MXPath predicates. An ec qualifier may be applied in every step of a MXPath expression and filter the resulting nodes of the corresponding step according to their explicit context. Explicit context qualifiers are of the form:

```
ec() comparison_op context_specifier_expression
```

The function `ec()` returns the explicit context of the current node. Note that, the predicates assigned to a *context MXPath step* are applied to the context nodes obtained from the evaluation of this step. In the same way, if a MXPath step is a *multidimensional MXPath step*, predicates are applied to the resulting multidimensional nodes.

7 Ordered-Based Context Operations and Comparison

In this section we define how we can apply set operations and comparison among context specifiers when they are represented in Ordered-Based Context Representation.

We first demonstrate how the intersection and union of context specifiers is performed at the level of World Vectors.

Lemma 1. *Let c_1, c_2 be two context specifiers and b_1, b_2 the world vectors of c_1, c_2 respectively. Then the world vector b_3 of the context intersection $c_1 \cap^c c_2$ is obtained by applying the AND operation[3] to the corresponding bits of b_1 and b_2. Respectively, the world vector b_4 of the context union $c_1 \cup^c c_2$ is obtained by applying the OR operation[4] to the corresponding bits of b_1 and b_2.*

Example 7. Consider the context specifiers:
$c_1 = [edition = english]$, and

[3] For this bit-wise AND operation we will use the abbreviation AND_b.
[4] For the bit-wise OR operation we will use the abbreviation OR_b.

$c_2 = [edition = english, customer\_type = student]$.
As we have shown in Example 6 the world vector of the context specifier c_1 is $V(c_1)$=000111. Similarly, it is derived that $V(c_2)$=000100. Then we have:
$b_3 = V(c_1 \cap^c c_2) = 000111\ AND_b\ 000100 = 000100$
and
$b_4 = V(c_1 \cup^c c_2) = 000111\ OR_b\ 000100 = 000111$

It is also possible to compare two context specifiers using their world vectors. This is very useful when we are trying to transform MXML queries containing relevant conditions to SQL queries over a Relational Database. These conditions imply comparisons between the context specifiers which are stored with the MXML document in the relational schema, and the context specifiers which are used in the MXML queries. Similarly to AND_b and OR_b, in Lemma 2 we use the abbreviation XOR_b for the bit-wise XOR operation.

Lemma 2. *Let c_1, c_2 be two context specifiers and b_1, b_2 the world vectors of c_1, c_2 respectively. Then*

1. $c_1 = c_2$ *iff* $b_1 = b_2$, *alternatively* $c_1 = c_2$ *iff* $(b_1\ XOR_b\ b_2) = 0$
2. $c_1 \neq c_2$ *iff* $NOT(b_1 = b_2)$
3. $c_1 \geq c_2$ *iff* $(b_1\ AND_b\ b_2) = b_2$
4. $c_1 > c_2$ *iff* $((b_1\ AND_b\ b_2) = b_2)$ *and* $(b_1 \neq b_2)$.

Example 8. Consider the context specifiers:
$c_1 = [edition = english]$ and
$c_2 = [edition = english, customer\_type = student]$.
Calculating the world vectors of those two context specifiers we have $V(c_1)$=000111=b_1 and $V(c_2)$=000100=b_2. Then the expression $c_1 \geq c_2$ is *true*, as $(b_1\ AND_b\ b_2) = (000111\ AND_b\ 000100) = 000100 = b_2$ (see Case 3 of Lemma 2).

8 Discussion and motivation for future work

Two techniques to store MXML documents in relational databases are presented in this paper. The first one is straightforward and uses a single table to store MXML. The second divides MXML information according to node types in the MXML-graph and, although it is more complex than the first one, it performs better during querying. Additionally, we presented context representation techniques for storing context in a RDB. We also presented MXPath, which is an extension of XPath, in order to query MXML documents and finally, it was shown how we can perform operations and comparisons between context specifiers. Future work will focus on (a) algorithms for SQL translation of MXPath queries giving as the ability for experimental evaluation of the querying performance and (b) optimization of MXML storage using alternative indexing techniques for improving relational schema and query performance.

References

1. T. Amagasa, M. Yoshikawa, and S. Uemura. A Data Model for Temporal XML Documents. In *Proc. of DEXA 2000*, pages 334–344. Springer, 2000.
2. T. Amagasa, M. Yoshikawa, and S. Uemura. Realizing Temporal XML Repositories using Temporal Relational Databases. In *Proc. of the 3rd Int. Symp. on Cooperative Database Systems and Applications, Beijing, China*, pages 63–68, 2001.
3. P. Bohannon, J. Freire, P. Roy, and J. Simon. From XML Schema to Relations: A Cost-Based Approach to XML Storage. In *Proc. of ICDE 2002*.
4. A. Deutsch, M. F. Fernandez, and D. Suciu. Storing Semistructured Data with STORED. In *Proc. of ACM SIGMOD Int. Conf. on Management of Data*, pages 431–442. ACM Press, 1999.
5. F. Du, S. Amer-Yahia, and J. Freire. ShreX: Managing XML Documents in Relational Databases. In *Proc. of VLDB' 04*, pages 1297–1300. Morgan Kaufmann.
6. D. Florescu and D. Kossmann. Storing and Querying XML Data using an RDBMS. *Bulletin of the IEEE Comp. Soc. Tech. Com. on Data Eng.*, 22(3):27–34, 1999.
7. M. Gergatsoulis, Y. Stavrakas, and D. Karteris. Incorporating Dimensions in XML and DTD. In Proc. of *DEXA' 01*, LNCS Vol. 2113, pages 646–656. Springer, 2001.
8. M. Gergatsoulis, Y. Stavrakas, D. Karteris, A. Mouzaki, and D. Sterpis. A Web-based System for Handling Multidimensional Information through MXML. In Proc. of *ADBIS' 01*, LNCS, Vol. 2151, pages 352–365. Springer-Verlag, 2001.
9. M. Ramanath, J. Freire, J. R. Haritsa, and P. Roy. Searching for Efficient XML-to-Relational Mappings. In Proc. of *XSym 2003*, pages 19–36. Springer, 2003.
10. J. Shanmugasundaram, E. J. Shekita, J. Kiernan, R. Krishnamurthy, S. Viglas, J. F. Naughton, and I. Tatarinov. A General Technique for Querying XML Documents using a Relational Database System. *SIGMOD Record*, 30(3):20–26, 2001.
11. J. Shanmugasundaram, K. Tufte, C. Zhang, G. He, D. J. DeWitt, and J. F. Naughton. Relational Databases for Querying XML Documents: Limitations and Opportunities. In Proc. of *VLDB'99*, pages 302–314. Morgan Kaufmann, 1999.
12. Y. Stavrakas and M. Gergatsoulis. Multidimensional Semistructured Data: Representing Context-Dependent Information on the Web. In Proc. of *CAiSE 2002*, LNCS Vol. 2348, pages 183–199. Springer, 2002.
13. I. Tatarinov, S. Viglas, K. S. Beyer, J. Shanmugasundaram, E. J. Shekita, and C. Zhang. Storing and querying ordered XML using a relational database system. In *Proc. of the 2002 ACM SIGMOD Int. Conf. on Management of Data*, pages 204–215. ACM, 2002.
14. W3C CONSORTIUM. XML Path Language (XPath) 2.0. http://www.w3.org/TR/xpath20/, January 2007.
15. F. Tian, D. J. DeWitt, J. Chen, and C. Zhang. The Design and Performance Evaluation of Alternative XML Storage Strategies. *SIGMOD Record*, 31(1):5–10, 2002.
16. F. Wang, X. Zhou, and C. Zaniolo. Using XML to Build Efficient Transaction-Time Temporal Database Systems on Relational Databases. In *Proc. of ICDE 2006*.
17. M. Yoshikawa, T. Amagasa, T. Shimura, and S. Uemura. XRel: a path-based approach to storage and retrieval of XML documents using relational databases. *ACM Transactions on Internet Technology*, 1(1):110–141, 2001.

Discovering Current Practices for Records of Historic Buildings and Mapping them to Standards

Michail Agathos, Sarantos Kapidakis
Ionian University, Department of Archives and Library Science
Laboratory on Digital Libraries and Electronic Publishing
Ioanni Theotoki 72, 49100, Corfu
{agathos, sarantos}@ionio.gr

Abstract. The existence of historic building records in "paper fiches" is a reality and constitutes a rich store of information about the past, some of it unique. In this paper we present the results of a survey aimed to discover the current practices and methods for recording historic buildings, mainly from services of the Greek public sector, which are responsible for the build heritage. At the same time the survey focuses on the various schemas, from the collected "paper fiches" that participants use for the documentation of immovable monuments as well as on metadata standards for architectural works and their ability to describe the collected elements of these forms.

Keywords: Historic Building Records, Immovable Monuments, Metadata Standards, Monument Inventories, Inventory Forms, Architectural Heritage Council of Europe, Greek Public Sector, Architectural Heritage.

1 Introduction

The investigation and documentation of the built heritage is central to our understanding of our historical evolution. Historic buildings, especially, form a conspicuous component of the urban and rural scene, and constitute a rich store of information about the past, some of it unique. These structures of our culture usually have documentation in form of so-called: paper fiches [1], inventory cards or forms, white cards, register cards and are dispersed in a number of various Greek public services and institutions.

In order to explore this type of documentation, that remained unexplored, we conducted a survey, from April 1, 2010 through March 15, 2011 involving a sample of 43 services of public sector (90%), mostly of the Greek Ministry of Culture and Tourism) and 5 non-profit organizations and institutions in Greece (see Appendix 2). Most of the participants working in the field of the built heritage having an important role on local level as their authorities refer to all matters concerning mainly the safeguard and protection of Hellenic heritage as the conservation, reconstruction, study and publication of the monuments. Objectives of this survey was to explore - at a national level - the methodology used for documenting historic buildings and generally immovable monuments, the existence of building records in "paper fiches" the degree of syntactic and semantic interoperability regarding their compilation methods, as well as to identify and highlight common descriptive needs among these organizations.

Participants were asked to complete a questionnaire, contained a total of 17 questions (close ended questions, open ended - completely unstructured, scaled questions: use of Likert items and Likert scale) and to return it with a completed example of their form (if used such a form). Among many interesting findings we collected 31[1] different forms including a total of 141 elements (see Appendix 1).

2 Exploring the Practices

Participants were asked if they compile or use forms in "paper fiches" for the recording of historic buildings and general for immovable monuments, research reveals that 31 Organizations (65%) produce or use such forms. About 77 percent (24 Organizations), said that forms had been produced by their own staff, while 23 percent (7 Participants) use forms from cognate services. The compiler is always a member of the staff, either archaeologist or Architect or a working group composed of archaeologists and architects.

We asked form the participants to mention the basic purpose and objective of these forms. The responses reflect their needs to record, inventory or identify immovable monuments located within the jurisdiction of the Organization, making thus a "local" inventory for "local" use, while institutions embrace research as a basic purpose.

The most basic question in this research was about the method of preparation of that forms. The participants were asked if had followed or advised a guidance or a standard for the preparation of their forms (without mention any particular), as an interesting finding from the 25 organisations responded to that question only 8 (26%) followed an official guidance or schema. Specifically 2 Organizations prepared their form based to CIDOC–CRM (ISO 21127:2006)[2], 2 participants answered that followed general guidance's for recording historic buildings, another 2 use forms for international Organizations and Committees (UNESCO - DO.CO.MO.MO.) and finally 3 organizations followed specific guidelines of Hellenic Ministry of Culture and Tourism. The findings of this question was expected as there is no a legally binding standard for the built heritage recording in Greece.

Moreover Organizations were asked to rate, whether the elements recorded on these forms satisfy their needs. A likert scale (from 1 - 10 with 10 being the highest) revealed a moderate satisfaction (mean: 5,33) with no variation in satisfaction level, while only 28 percent of those responding to the question declare satisfied with the recorded elements (rating more than 7).

Furthermore, research gave space to participants to record their needs for additional elements that they would like to be included in their forms: The most common requirements was for elements that will record: documents related to the buildings, correspondence with other services, regular photography, marking on digital maps, recording of dimensions, analysis on materials, information about conservation and restoration status, interventions, delimitation of buffer zones. Not quite as many, but

[1] All the Participants keep in store a total of 900,000 forms.

[2] European Centre for Byzantine and Post Byzantine Monuments, Minister of Culture and Tourism - 13th Ephorate of Byzantine Antiquities.

still a large number of organizations asked for: Land Registry info, documents of ownership titles, drawings, description of decoration and recording of morphological elements.

A disappointing finding of the survey, was that just over half of these forms (52%) are available only to officials, and only 48 percent of this information is available to the public, as a result, persons requiring information on particular buildings have a limited access on their heritage status, and related data.

Although all of these records co - exist in digital and print format, 20 organizations (65%) register these forms in a computer system and only 35 percent of these exist only in print format. As a follow-on from the above question, participants were asked if they had developed a relevant application in order to register these forms, a small number of responses (13) showed that public services create and maintain their own computerised record systems, their own "local" systems. Specifically 9 participants said that they have created a local database system, another 3 use web applications and 1 participant indicate "other" application, without specifying any particular.

At this point it is worth to comment that, there is no lack of computerised heritage documentation system[3] in Greece, but public sector lacks the financial resources to maintain these information systems and there is a shortage of staff and of essential skills. This is a common problem, as 95 per cent of all cultural heritage institutions in Europe in 2002 were not in the position to participate in any kind of digital cultural heritage venture (Mulrenin: 2002) [2].

Furthermore organisations were asked if they produce digital content relative to historic buildings, more than half of the respondents (53%) replied positive: This is mainly: photographic material, drawings, scanned maps/plans, and in a small percentage: orthophotograpies - digital orthophoto mosaic, topographic backgrounds, Excel files, .doc, e.t.c). After being informed for the existence of this digital content, participants were asked again about the format of this content (see Table 1).

Table 1. Formats of digital content

JPEG/TIFF	42%
db Files	29%
cad Files	11%
xml	7%
xls	7%
Doc	4%

Finally, one of the most interesting statistics in this survey was that 46 participants (96%) thought that there is a need for encoding and standardization for information in the domain of immovable monuments, however only 2 (4%) thought that encoding of such information is not feasible and would be difficult to standardized.

The survey also contained a section for general comments. The following comment highlights that: "*The documentation, with a systematic way, is the basis of any serious*

[3] "POLEMON" is the official information system of Hellenic National Archive of Monuments and was designed to meet the needs of the various units and services of the Hellenic Ministry of Culture providing an integrated set of tools for Monuments and Collections Management.

scientific research, but also the basis for monitoring the history and interventions for the protection of any historic building. Unfortunately, this approach is not addressed with the expected serious way, of the protection bodies[4]*"*.

The most frequently voice requests (5 respondents) suggested the creation of a common schema for immovable monuments. The following comment is representative:"*It would be desirable to have a form common to all, in which will be recorded in addition to the historical and architectural data and maintenance data, response and recovery. Occasionally there were some attempts with no avail so far"*.

Also there were also a small number of comments that demonstrated that:"*Historic buildings – monuments, appears a set of unique characteristics, therefore, a coding would be quite limited only to few general elements"*.

3 Studying the Various Schemas

As mentioned bellow each organization prepares and uses its own form. The lack of a binding common schema for common building types has as a result same building types being described with a different element set (schema) each time. Moreover the study on 31 collected forms (one from each service) shows that a substantial majority of the participants record a minimum amount of information. Number of elements varies from one from to another: Specifically 90 percent of these forms are optical records[5] [3] (up to 5-6 elements) complemented by the minimal information necessary to identify the location of the building, its type, its legal status and some general characteristics. Description at this level is limited to the exterior of the building with some exceptions, where there are very significant internal or decorative features. Forms with a fuller description are limited. Finally there is a great discrepancy between the data recorded by the surveyed services and the recommended[6] by the Council of Europe element set of Core Data Index to Historic Buildings and Monuments [4] as well as the Principles for the recording of Monuments, Groups of Buildings and Sites as expressed in the 11th ICOMOS General Assembly in Sofia. As a result, forms do not include some information crucial for successful protection and management of historical buildings.

Specificity and exhaustivity is another major issue for these records. As emerged from the study, there is a terminological confusion, as organizations do not use a controlled list of terms for the various elements. Moreover elements of each schema even when used to describe the same concept, differ. In order to give a typical example organizations use many non equivalent terms (for example. Category/Typology/Type/Characterization) in order to describe the type of the building.

[4] Hellenic ICOMOS.

[5] According to English Heritage Recording Levels

[6] Recommendation R(95)3 of the Committee of Ministers of the Council of Europe to member States on co-ordinating documentation methods and systems related to historic buildings and monuments of the architectural heritage, Strasbourg (1995)

4 Reviewing the Standards

Since the 1960s, the Council of Europe has worked to protect and enhance the architectural and archaeological heritage, through the exchange of ideas and through developing guidelines and standards. Among their efforts is the design of two affined international standards for the documentation of the immovable cultural heritage: the Core Data Index to Historic Buildings and Monuments of the Architectural Heritage[7] (1992) and the International Core Data Standard for Archaeological Sites and Monuments[8] (1995). The standards define the core information (basic minimum categories) for documenting historic buildings, archaeological sites and monuments [5].

"Core information" may be defined as those categories of essential information or basic documentation (textual and pictorial) common to a broad array of documentation projects, whether manual or computerized, which make it easier to record, use, and exchange information. It has been described as an enabling mechanism that "represents a way of indexing, ordering and classifying information, independently of whether that information is on paper, card index, or database" [6]. The Dublin Core Metadata Element Set is an example of a such successful model of "core information".

The basic aim of the *CDI* (1992) is to make it possible to classify individual buildings and sites into 9 information groups (sections): Names and References, Location, Functional Type, Dating, Persons & Organizations, Building Materials and Techniques, Physical Condition, Protection/ Legal status and Notes [7]. These 9 sections are supported by sub-sections and a set of 45 data fields, some of which are mandatory. The *CDI* is designed to enable the compiler to make cross-references to the more detailed information about a building, including written descriptions and photographs; associated archaeological and environmental information; details of fixtures, fittings, and machinery installed within individual buildings; and the information on persons and organisations concerned with their history. The *CDI* has the potential not only to record individual buildings, but also to enable the compiler to relate a building to a larger site of which it may be a component or to the still larger ensemble of which it may form a part.

The International *CDS* (1995) aims to identify the categories necessary for documenting the immovable archaeological heritage. It consists of 7 sections: Names and References, Location, Type, Dating, Physical condition, Designation/Protection Status and Archaeological Summary [8]. These 7 sections contain sub sections, which in turn include a set of 52 data fields, some of which are mandatory.

The *CDS* has been designed to make it possible to record the minimum categories of information required to make a reasonable assessment of a monument or site. In addition, it makes it possible to provide references to further information held in databases, documentation centres, and elsewhere that may be necessary for the detailed understanding and care of individual monuments or sites or categories of monument or site.

MIDAS Heritage [9] is a data standard for information about the historic environment which was developed for use in the UK and Ireland and is maintained by the Forum on Information Standards in Heritage. It states what information should be

[7] For brevity's sake will be referred as *CDI*
[8] For brevity's sake will be referred as *CDS*

recorded to support effective sharing and long-term preservation of the knowledge of the historic environment. It consists of 9 Themes: the broadest level areas of interest, 16 Information Groups, these set the specific standard for what should be included in an entry covering a particular subject and 138 Units of Information the basic 'facts' or items that make up an entry. 'Monument' information group in MIDAS Heritage usage, among built, buried and underwater heritage of all dates and types, includes buildings (both ruined and in use). MIDAS Heritage can be used to plan the content of a new inventory, for example to support a new project. Alternatively it can be used to audit the existing content of an inventory, and identify any useful additional information that could be included. MIDAS is designed to be an 'open' standard, which can be applied in a variety of ways to different sorts of inventory records.

Realizing that there was a need in the art documentation and museum communities for a data structure standard specifically designed for describing unique works of art, architecture, and material culture, in the late 1990s the Getty Institute and the Art Information Task Force (AITF) developed CDWA an extensive set of metadata elements (includes 532 categories and subcategories) and guidelines, which can describe the content of art databases by articulating a conceptual framework for describing and accessing information about works of art, architecture, other material culture, groups and collections of works, and related images.

What was still missing were a "AACR for art objects" [10], a data content standard specifically for unique museum and special collections-type objects and built works, and a technical format or data interchange standard for expressing and exchanging metadata records about those kinds of works. CCO (Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images) was the response to this need, which designed specifically to deal with unique items of art, architecture, and material culture. Actually, CCO, which is based on a subset of CDWA, is a manual for describing, documenting, and cataloging cultural works and their visual surrogates. The primary focus of CCO is art and architecture, including but not limited to paintings, sculpture, prints, manuscripts, photographs, built works, installations, and other visual media and types of cultural works. CCO is concerned only with descriptive cataloging of objects in a Work Record.

The CDWA Lite[9] schema (2006), which corresponds to CCO, is a response to later needs. Is a distillation of the very ample, exhaustive set of elements and sub-elements of CDWA. The purpose of this schema is to describe a format for core records for works of art and material culture, based on the data elements and guidelines contained in the CDWA and CCO. Like VRA Core, CDWA Lite offers an XML format in which to store metadata about works of visual culture in accordance with CCO. CDWA Lite XML schema has a total of twenty-two top-level elements. It is OAI-harvestable, relatively simple, and much more appropriate for expressing metadata records for art and material culture.

The VRA Core 4.0 XML (2007) is a descriptive metadata standard for the description of culture works (paintings, sculptures, photographs, buildings etc) as well as the images that document them. It consists of nineteen elements and twenty-three subelements.

6 Rating the Standards

Finally we classified these 141 Elements in 14 Categories**:** Titles, Location, Functional Type – Use, Names and Roles, Dating, Building Parts Materials and Techniques, Conservation/Treatment History, Physical Condition, Protection – Legal Status, General Notes, Illustrative Material: Images/plans/Sketches and Record Info. In order to answer the question which metadata standard of the reviewed above, would cover better the elements of the collected forms, we focused mainly on three complex categories from the above: "Building Parts", "Protection - Legal Status", "Conservation- Treatment History". An exhaustive comparison of these categories with the elements of the above reviewed metadata standards allowed us a hierarchical rating according to coverage provided (Fig.1)

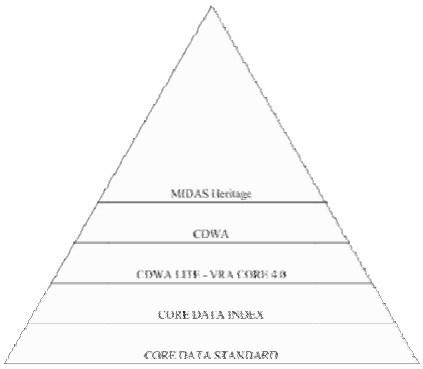

Fig.1. Hierarchical rating of the reviewed standards according to coverage provided.

MIDAS Heritage Standard is able to cover much of the collected elements. Specifically "Designation and Protection" information group of MIDAS allows for statements on whether the building is protected and, if so, the type of protection, the grade and the date at which it was granted. Moreover it is able to accommodate information's about the government body whish is responsible for the building, giving in parallel the relevant legislation with which the building is protected (Information Units: Statutory Name, Statutory Description, Protection Type, Protection Date, Protection Start/ End Date). Moreover the "Management Activity Documentation" information group covers a wide range of documentation for the significance of a building and the factors affecting its condition and survival. Last at not least, "Map Depiction", is a critical information group as include information to improve the understanding and use of spatial depictions of a building, which is a demand of the participants as described above. The various parts of the building could be described using the Information Unit "*component"* of the standard.

A shortcoming for Greek Forms is that MIDAS Heritage is aimed at planning the content of a new inventory, as is a set of closely integrated data standards, rather than one single stand alone standard. MIDAS has a three-level structure working from the broadest to the most specific (Information groups – Themes – Units of Information). User communities, who want to design any particular information system or dataset

based on MIDAS, have to develop first a shared compliance profile assisting them to develop a standard that meets their needs. The first step is to determine which Information Groups are relevant to the needs of the community, including these in the profile. Each Information Group includes a table which lists the requirement for Information Group entries to be qualified by entries in other Information Groups to create a full record. Moreover units of information for each Group can be assessed separately.

The Category "Conservation/Treatment History" of CDWA covers much of the collected elements that concerns procedures or actions that a building has undergone for repair or conserve. Description for the legal status and protection of a building is limited to "Legal Status" subcategory (one field), that allows for general statements as "public property" "scheduled property" "registered property" etc. Specific parts of the building could be described using "Materials/Techniques Extent" subcategory.

VRA CORE 4.0 as CDWA Lite provides the same level and method of description for these records. There are no equivalent elements to accommodate information for the conservation / treatment history or legal protection of a building. An additional shortcoming is that the various structural parts of the building (roofs, windows e.t.c) can be described in VRA CORE via the global attribute *extent* for CDWA Lite via the sub-element <cdwalite: extentMaterials Tech>. This is a shortcoming of the standards as for the complexity of the various parts may be required more sophisticated elements. On the other hand, VRA Core 4 is uniquely able to capture descriptive information about works and images, and indicate relationships between the two, using the same set of elements to describe both a building and its image(s)

CDI from the other recommends only two data fields to record the various parts of a building: "Main Materials and Structural Techniques" sub - section, for the main walling material, excluding partition walls and Covering Materials to record the main roofing material. In CDI there is no equivalent category for conservation or treatment history concerning the building and the elements for legal Information and legal protection are quite limited. Moreover Core Data Index is unable to cover Measurements, as there is no equivalent sub-section in the standard. Furthermore with the *CDI* and *CDS*, we can provide archival and bibliographic information or illustrative material about the building, only with references to external information held in databases, documentation centres, and elsewhere, enabling the compilers to conceptualise the route from microcosm to macrocosm and allowing the users of the information to make the same connections [5]. Data fields, which will be able to accommodate internal information as a map showing the building and its immediate curtilage or locality, a sketch ground plan and a photograph, would be desirable[10].

[10] Technical Co-operation and Consultancy Programme of the Council of Europe suggests a slightly expanded version of the *CDI*, with additional recommendations for the sections Physical Condition and Notes, as and a new section called Illustrations. It therefore goes a little beyond the officially agreed recommendation R (95) 3 of the Council of Europe.

7 Conclusion

It is recognised that these "local" practices and needs for the documentation of historic buildings, described above, will vary from organisation to organisation and country to country, and that each will define its own specific requirements since the diversity of the European heritage and the differences in national inventorisation traditions, and policies are such that the production of an international standard or recommendation would be neither feasible nor desirable [11]. Nevertheless standardization will help moderate this chaos, especially with the help of metadata standards that focused on works of architecture. Many of the metadata schemas described above, must be evolved and changed in order to stay aware of more global standards initiatives as methods of recording sites and buildings and of defining their significance have been developed to a high level of sophistication over recent decades.

The above described concepts are intended as a starting point about the maintenance and expansion already existing metadata schemas for historic buildings or the creation of a new harmonized profile.

References

1. Martens, B., A. Brown, et al. (2005). A Method Proposed for Adoption of Digital Technology in Architectural Heritage Documentation. *Computer Aided Architectural Design Futures 2005*, Springer, Netherlands: 73-82.
2. Mulrenin, A. (ed.). *The DigiCULT Report: Technological landscapes for tomorrow's cultural economy. Unlocking the value of cultural heritage*. Office for Official Publications of the European Communities, Luxembourg (2002).
3. English Heritage. *Understanding Historic Buildings - A guide to good recording practice – Part 1*. English Heritage Publishing (2006).
4. Council of Europe. *Core data index to historic buildings and monuments of the architectural heritage: Recommendation R(95)3 of the Committee of Ministers of the Council of Europe to member States on co-ordinating documentation methods and systems related to historic buildings and monuments of the architectural heritage*. Council of Europe, Strasbourg (1995).
5. Council of Europe: *Guidance on inventory and documentation of the cultural heritage*. Council of Europe, Strasbourg (2009).
6. Thornes, R.: *The value of core information. In: Protecting Cultural Objects: A preliminary survey*. J. Paul Getty Trust, California (1995), http://icom.museum/objectid/prelim/part1/part108.htm.
7. Council of Europe: *Documenting the Cultural Heritage*. Council of Europe, Strasbourg (1998), http://icom.museum/objectid/heritage/index.html.
8. Council of Europe. *Core data standard for archaeological sites / Fiche d'indexation minimale pour les sites archéologiques*. Council of Europe, Strasbourg (1999).
9. English Heritage, Data Standards Unit, National Monuments Record Centre. *MIDAS a Manual and Data Standard for Monument Inventories*, 3rd ed. English Heritage, Data Standards Unit, National Monuments Record Centre (2003).
10. Baca, M. CCO and CDWA Lite: Complementary Data Content and Data Format Standards for Art and Material Culture Information. *VRA Bulletin. 34*, 69-75 (2007).
11. Council of Europe. Architectural Heritage: Inventory and Documentation Methods in Europe. In *Proceedings of a European colloquy organized by the Council of Europe and the French*

Ministry for Education and Culture—Direction du patrimoine, Nantes, 28–31 October 1992. Strasbourg: Council of Europe, http://www.coe.int/t/dg4/cultureheritage/heritage/resources/Publications/Pat_PA_28_en.pdf.

12. Sykes, M.H. *Manual on Systems of Inventorying Immovable Cultural Property*. Unipub, Lanham (1984).

Appendix: 1. Elements of the Forms

1. Titles	
Elements	*Frequency*
Building Name	•••11
Local Appellation	•
2. Location	
Street - Road	•••
Commune	•••
Prefecture	•••
Location	•
Area	•
Municipal Department	•
Settlement	•
Within the Settlement	•
Outside the Settlement	•
Position of the Building in the area	•
Coordinates	•
Postal Code	•
Locality or Residential Section	•
Surroundings	
Characterization of the building ground	•
Exact Location	•
Cadastral Register Number	•
Block Number	•
Area Characteristics	•
Hellenic Statistical Authority Number	•
3. Functional Type / Use	
Primary Use	•••
Current Use	•••
Characterization	•
Proposed Use	•
Type of Monument	•
Category / Typology	•

4. Names and Roles	
Elements	*Frequency*
Owners/Ownership	••
Owner by floor	•
Architect	•
Donator	•
Constructor	•
Collaborators	•
Ownership status	•
5. Dating	
Date / Chronology	•••
Historical Period	•
Year of construction	•
Date of Recording	•
Construction Period	•
6. Conservation / Treatment History	
Conservation Status	••
Contemporary Interventions	•
Restoration Proposal	•
Conservation Works	•
Conservation Status	•
7. Physical Condition	
Deformation	
Additaments Denaturations	•
Dampness	•
Structural Status	•
Current Status	•
Attritions	•

[11] Repetition frequency: ••• big repeatability , •• moderate repeatability, • small repeatability

8. Building Parts: Materials&Techniques	
Elements	Frequency
Roof	••
Coloration	•
Frames	•
Masonry	•
Staircase	•
Balcony	•
Floors	•
Decoration	•
Technique	•
Ceiling	•
Soffit	•
Building Shell	•
Structure System	•
Building Shell	•
Type of folding shutter	•
Rails	•
Morphological Elements	•
Morphological Status	•
Construction	•
Bedrock	•
Coating	•
Inscriptions	•
Painting	•
Sculpture	•
Architecture	•
9. Measurements	
Number of Floors	•••
Building Area	•
Building Coefficient	•
Basement Area	•
Ground Floor Area	•
Floor Area	•
Building Site Area	•
Number of Entrances	•
Building Dimensions	•
Number of Houses	•

10. Protection / Legal Status	
Elements	Frequency
Gazette	•••
Number of Ministerial Decision – Statute Number	•••
Proposed Protection by	•••
Protection Body	•••
Gazette Title	•
Type of Declaration	•
Under Declaration	•
Grade of Protection I.P.C.E.	•
Characterization Date	•
Declaration Type	•
Ministerial Decision Date	•
Proposal of conservation	•
Grade of Protection	•
Inspected by	•
Buffer Zone (A or B)	•
Zone Borders / Delimitation	•
11. General Notes	
Comments	
Historical Facts	
Oral evidence	
Estimation / Appraisal	
Description of the Monument	
Approvals - autopsies	
Assessment Degree	
Artistic Value	
Building Assessment	
12. Related References	
Sources / Bibliography	•••
Folder Number	•
Film Number	•
Slide Number	•
Documents / Correspondence	•
Sources / Bibliography	•

Building Height	•
13. Illustrative Material: Images/plans/Sketches	
Elements	Frequency
Photography	•••
Scale	••
Map Extract	•
Extract of Cadastral Map	•
Area Map	•
Sketch	•
Sketch ground plan	•
Sketch ground plan of Floors	•
Sketch ground plan of East Aspect	•
Sketch ground plan of South Aspect	•
Sketch ground plan of West aspect	•
Sketch ground plan of Roof	•
Sketch ground plan of Basement	•
Topographical Plan	•
Sketch of South Aspect	•
Architect Drawings	•
Scale	•
Plan Dimensions	•
Plan Inscription	•
Plan Material Status	•
	•
Record Change Date	•
Checked by:	•
Compiler	•
Building Number / Record Number	•
Compilation Date	•

14. Record Info	
Elements	Frequency
Building Number	••
Record Number	••
Compiler	•
Record Change Date	•
Compilation Date	•
Checked by:	•

Appendix 2: Participants

We are grateful to all participants, who took time out of their busy schedules to participate in the study

Public Sector
General State Archives
General State Archives - District of Corfu
Municipality of Heraklion - Old Towh Office

Municapility of Corfu - Old Towh Office
Hellenic Statistical Authority

Ministry of Infrastructure, Transport and Networks - Depended Services:
Technical Chamber of Greece - Regional Department of West Crete
Technical Chamber of Greece - Regional Department of Eteoloakarnania
Technical Chamber of Greece - Regional Department of Corfu
Ministry of Environment Energy & Climate Change - Archive of traditional and listed buildings
Ministry of Maritime Affairs Islands and Fisheries - Secretariat General for the Aegean and Island Policy
Ministry of Finance, Real Estate Service (District of Corfu)

Hellenic Ministry of Culture and Tourism – Dependent Services:
National archive of Monuments.
Directorate of Cultural Buildings and Restoration of Contemporary Monuments - Department for the Study of Modern Monuments.
Directorate of Modern and Contemporary Architectural Heritage.
Directorate of Topography, Photogrammetry and Land Register.
3rd Ephorate of Byzantine Antiquities
6th Ephorate of Byzantine Antiquities.
9th Ephorate of Byzantine Antiquities.
10th Ephorate of Byzantine Antiquities.
11th Ephorate of Byzantine Antiquities.
13th Ephorate of Byzantine Antiquities.
14th Ephorate of Byzantine Antiquities.
15th Ephorate of Byzantine Antiquities.
16th Ephorate of Byzantine Antiquities.
19th Ephorate of Byzantine Antiquities.
22th Ephorate of Byzantine Antiquities.
25th Ephorate of Byzantine Antiquities.
26th Ephorate of Byzantine Antiquities.
Ephorate of Contemporary and Modern Monuments of Attica.
Ephorate of Contemporary and Modern Monuments of Crete.
Ephorate of Contemporary and Modern Monuments of Thessalia.
Ephorate of Contemporary and Modern Monuments of Central Macedonia.
Ephorate of Contemporary and Modern Monuments of North Aegean.
Ephorate of Contemporary and Modern Monuments of West Greece.
Ephorate of Contemporary and Modern Monuments of Hipirus.

Directorate of Prehistoric and Classical Antiquities
2nd Ephorate of Prehistoric and Classical Antiquities.
4th Ephorate of Prehistoric and Classical Antiquities.
16th Ephorate of Prehistoric and Classical Antiquities.
27th Ephorate of Prehistoric and Classical Antiquities.
29th Ephorate of Prehistoric and Classical Antiquities.
38th Ephorate of Prehistoric and Classical Antiquities.

National Gallery - Alexandros Soutzos Museum (supervised organization).

Non Government Organizations
European Centre for Byzantine and Post Byzantine Monuments
Hellenic Society for the Protection of the Environment and the Cultural Heritage
Hellenic ICOMOS (scientific committee)
Benaki Museum - Neohellenic Architectural Archives
Hellenic Society for the Protection of the Environment and the Cultural Heritage

The exploitation of social tagging in libraries

Constantia Kakali, Christos Papatheodorou
Database & Information Systems group, Laboratory on Digital Libraries and Electronic Publishing, Department of Archive and Library Sciences, Ionian University, Corfu, Greece
{nkakali, papatheodor}@ionio.gr

Abstract. Nowadays, many libraries have developed social tagging services, after the considerable use of social tagging and deployment as key components of Web 2.0. Another set of libraries have enriched the search and indexing services of their OPACs with the folksonomy of Library Thing. The evaluation of these metadata (folksonomies) and further their exploitation is one of our challenges. At the same time, we explore ways to define a methodology for the exploitation of user's vocabulary by the traditional indexing systems maintained by information organizations. Firstly, our research focused on the user acceptance for the OPACIAL an OPAC 2.0 with social tagging functionalities. The users' behavior was studied by qualitative evaluation using questionnaires and structured interviews. Social tags are then analyzed and categorized to identify the users' needs. After finding that a large number of tags consist new terms for the authority file of a Library, these tags were searched in other authority files. The research was completed by developing a methodology for social tagging evaluation and a proposal for developing policies to integrate social tags in their indexing processes. Moving to a new study, librarians - cataloguers assessed the value of the semantics of inserted tags and also investigated the possibility of using them for the subject indexing. Before the new experiment a new set of tags from LibraryThing's folksonomy had been added to the library. The experiment aimed to compare the two vocabularies and the participants recommended to develop the cooperation with users' communities in matters of terminology and apodosis of scientific terms.

1 Introduction

The Web 2.0 technologies offer to users the chance to create metadata by organizing their information resources. This metadata creation is implemented by adding uncontrolled keywords, named tags to the resources. The phenomenon is called social tagging or collaborative tagging and has grown in popularity firstly in social bookmarking sites like Delicious, CiteUlike, Flickr etc. The set of the tags introduced for a resource is called folksonomy, it could be presented as a tag cloud and express the users' vocabularies and needs.

Folksonomies are referred as a borderline case of knowledge organization systems (KOS) [1]. It is distinguished from other KOS as a flat system with many limitations, despite the democratic generation of users' literacy [2]. In contrast to traditional classification systems and thesauri, there is neither "authority control", nor selection criteria and instructions for tag generation and as a result many similar tags are

generated. The main disadvantages of folksonomies are their flat structure and inherent ambiguity of tags, which raises polysemy and synonymy problems. Usually the tags are appeared in singular and plural form concurrently, while different users apply to the same tags different meanings [3].

Recently the social tagging has been proved useful in various information organizations as museums, libraries and archives. Libraries have been taking the advantage of folksonomies to allow users to organize personal information spaces, provide tags to supplement existing controlled vocabulary and develop on line communities of interest [4]. Many pioneer libraries launch new catalogues (OPAC) or web-based applications that are inspired by the technologies of Web 2.0. The new systems, usually called OPAC 2.0, are either open source software, such as VuFind, Scriblio, AFI-OPAC 2.0 and SOPAC, or proprietary applications, such as Aquabrowser Encore and Primo. They all provide a set of key features, such as folksonomies (user keywords, tagging) and search terms recommendations, as enhanced means of supporting users' search strategies. Other libraries have enriched the indexing and search services to their lists by linking the social web application cataloging: Library Thing. LibraryThing (http://www.librarything.com/), a social cataloging site, allows among other social tagging and annotations in bibliographic records, which are used for organizing personal collections of users.

Given that an increasing number of libraries develop social tagging systems in parallel to their traditional services to develop structured and controlled knowledge organization systems, a key issue concerns the impact of social tags to the subject indexing. This study is focusing on the alignment of the two different approaches and present two different experimental studies on the use and the value of social tags in a library environment. The paper aims to propose a policy for the exploitation of social tagging system by information scientists in libraries.

2 Related work

Immediately after their development social tagging systems were been researched and studied by various categories of scientists. Information scientists aimed to compare the classical thematic indexes to the vocabularies used in tagging systems.

Lin, Beaudoin, Bui, and Desai [5] compared social tags with automatically extracted terms from resource titles and descriptors from MeSH, in order to check the adequacy of three keywords sets (tags, term titles, and thesaurus terms) regarding indexing quality. The comparison showed that only the 11% of tags match the MeSH terms and this was due to the different goals of the controlled vocabularies and social tagging. They also investigated how tags could be categorized to improve the searching and browsing effectiveness. Margaret Kipp [6], in her analysis on tags of CiteULike resources, compared the vocabularies of users, authors and cataloguers, and showed that user tags are related to the author keywords and cataloguers subjects, and the majority of tags were broader or new terms. Moreover the study of Al-Khalifa and Davis [7] showed that the folksonomy tags overlap significantly with the human generated keywords in contrast to the automatically generated. Voss [8] explored the similarities and differences between Wikipedia, folksonomies and traditional

hierarchical classification systems (e.g. Dewey Decimal Classification) and he concluded that Wikipedia's category system constitutes a thesaurus based on a special combination of social tagging and hierarchical subject indexing.

Most of the researchers that studied folksonomies agree to a positive role in libraries in parallel with the heavy controlled indexing systems, despite their differences. Yi and Chan [9] investigated the relation of the LCSH and social tags selected from Delicious. The study of the tags distribution over LCSH concluded that LCSH "may greatly enhance the collaborative tagging systems information control process" and "it is possible to connect collaborative tagging systems with OPACs or digital libraries". Next year, Yi [10] examined ways of predicting relevant subject headings from the social tags of resources, using 5 different similarity metrics (tf-idf, CoS, Jaccard, mutual information, iRad).

Thomas, Caudle and Schmitz [11] performed a comparison of social tags with LCSH. They report an effort of the librarians of the Cataloging Department, Auburn that compares the social tags and LCSH assigned to a sample of ten books in problematic subject areas across a sample of libraries. The analysis followed a combination of tag classification criteria mentioned by Golder and Huberman [12] and Kipp [6].

LibraryThing content has been used by several tag analysis experiments and innovative systems. According to [13], the comparison of LibraryThing's tags against their equivalent LC subject headings showed that the number of LC headings varied from book to book, but on average there existed more tags than headings. Smith [14] and Bartley [15] explored the relationship between folksonomy and subject analysis in a study of LibraryThing tags and (LCSH) associated with the same documents, and her results showed that the tags identified latent subjects. Bartley [15] in similar research showed that the majority of tags are overlap with MARC fields of the records (245: Title & 600: Subject fields). Pera, Lund and Ng [16] designed EnLibS, an online library system that aims to take advantage of the keyword similarity searching and folksonomy datasets to reduce the need for complicated search strategies and knowledge of LCSH terms. Finally Lawson [17] compared the 31 top-level subject divisions and the tags from Amazon.com and LibraryThing associated with a sample of 155 books and she claimed that social tagging enables librarians to partner with users to enhance subject access.

Heymann and Garcia-Molina [18] compared social tags and LCSH and found a large degree of overlap, but also differences in the usage of common terms by users and professionals. Rolla [19], analyzing 45 entries with subject headings and social tagging, found that in general the user tags cover the scientific domains, but a large percentage are personal and without value for information retrieval. Lu, Park and Hu [20] analyzed the similarities and differences in systems, highlighting the value of the tags as additional and complementary to the subject indexing. Even for the type of digital objects such as photographs, Stvilia and Jorgense [21] found that the half of tags they had examined from Flickr is not included on TGM and LCSH.

3 The OPACIAL system – preliminary study

Few years ago, a new OPAC 2.0 was developed by the Panteion University Library, Athens, Greece. The added-value features of OPACIAL include tagging functionalities, folksonomy-based navigation to the library material, as well as tag searching. Moreover OPACIAL provides user annotations, ranking functionalities and use of reference tools. The users are able to annotate and rank each resource (on a 1 to 5 scale) and to export a record to external social networking sites by using a social networking site aggregator, like Socializer. A significant feature of OPACIAL is the integration of OPAC records with the ones of the University's digital repository, named Pandemos and also deployed by the Library. Thus, for each OPAC record the user is capable to retrieve similar digital objects. Recently, Opacial has been enhanced and every user of the library can develop and maintain its own personomy.

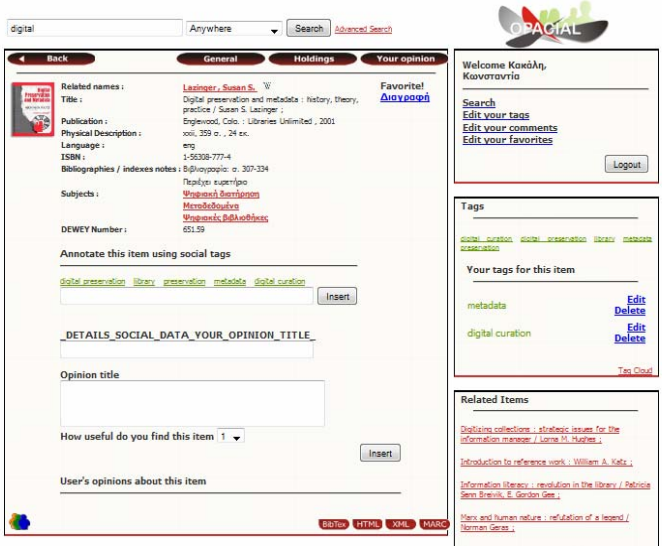

Fig. 1. OPACIAL's folksonomy management

OPACIAL has been evaluated by a technology acceptance experiment [22], in which twenty users (post graduate students and faculty members) used all its functionalities for a week, inserted more than 500 tags and finally were interviewed to assess the system usability and usefulness. The aims of this user study were to identify (a) the importance of social tagging for the users' information seeking process, (b) the difference in information search effectiveness between the use of tags and subject headings of the library catalog, and (c) the accessibility of the new services. During the empirical study a critical mass of tags was inserted by the participants, feeding the present research with valuable content.

The evaluation criteria were (a) relevance: how relevant items to the user needs returns the tagging functionality, (b) reliability: could the tags guide the users queries, (c) format: is the integration of OPAC records with object from the digital library helpful, (d) timeliness: the tags awareness, (e) learnability: how easy to learn tagging application, (f) navigation: how easy is the navigation, (g) Information architecture, (h) aesthetics. The first four criteria correspond to the usefulness concept, while the rest correspond to usability.

One of the important findings of the interviews was that the users in general consider tagging functionalities useful, as well as usable in their technological portrayal. Therefore they judged positively the new services, especially in comparison to the previous system, which was not regarded as satisfactory, despite the high level quality of the subject headings of the Library. Their general satisfaction grade was above average in the 7-point Likert scale, while the usefulness of tag introduction and search via tags functionalities recorded an average of 5.47. After experimenting with OPACIAL, the users rated the reliability of searching using tags with an average of 6.37. Referring to the second study aim users seemed prefer to use both the tags and the Library subject index. Specifically the users' view on the tags was that they play a complementary role to the existent subject index. Some of them used tags, either to describe precisely some OPAC records, or to correct wrong subject terms featured in them. Their preference was expressed by an 89.5% agreement on the assistive presence of both subject headings and social tags in their desktops. However, they were skeptical to browse the tag cloud and they were afraid of its constantly expanding size. Based on this remark a social tag searching functionality was added to the system. Concerning the tag introduction functionality the users suggested that domain experts should be allowed to add tags in order to create folksonomies and to suggest bibliographic lists for user communities. Finally, regarding the usability the general finding was that users found interaction with OPACIAL quite satisfactory and the level of accessibility quite high.

4 Tag analysis methodology

The results of the technology acceptance experiment provided an insight for the subject indexing process. This lead to a new objective, which was articulated as (a) the development of a policy for deciding the impact of the user community vocabularies to the local authority file development, and (b) the possibility of converging the user-based and the expert-based subject indexing approaches. For this purpose a tag analysis study was conducted considering several aspects of the tagging behaviour expected in this setting. The activities of the presented research could be grouped in concrete stages, formulating, thus, a methodology for the analysis and comparison of the two indexing approaches.

The methodology is shown in Figure 2 and its stages are briefly described as follows:

1. In Pre-procession stages the tag collection is delimitated. The collection can be defined by some criteria such as the time, the taggers (group of users), or a particular domain of the total collection. In our case, the collection was defined

by the tags of the first users who were the participants of our experiment we had referred above.

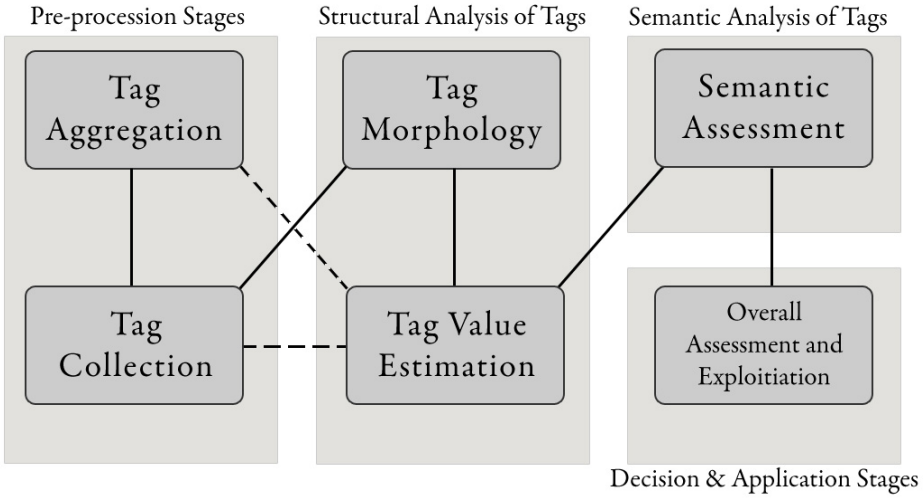

Fig. 2. Methodology of Tag Analysis

2. In the stages of Tag's structural analysis, in Morphology stage we began with a lexical analysis for grammatical forms. In the same stage, a significant activity is the study of the distribution of the tags over the bibliographic records. The interesting is that as the number of subject headings per record increases, the number of tags decreases (Fig. 3). This result confirms the assumption that tagging plays a complementary and enhancing role to weak subject descriptions.

   Continuing the search with the stage of Tag's value estimation, we aimed to emerge the similarities and differences between the tags and the descriptor terms in the authority file. A significant indicator which supports the behavior analysis held on this stage is the percentage of tags which already exist in the authority file. The 46.2% (269 tags) of the total amount of tags is not present in the existing authority file.

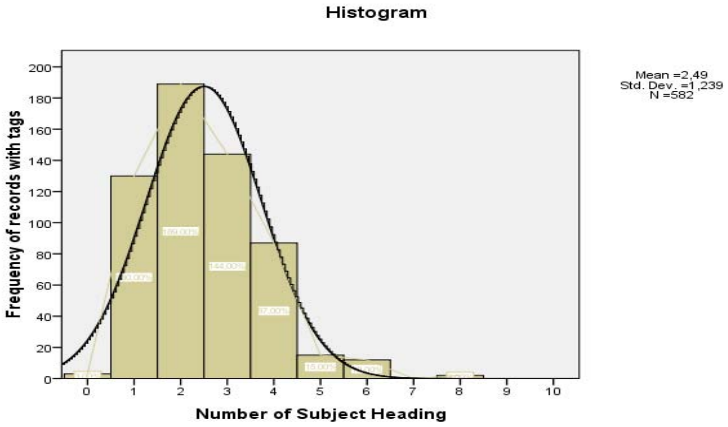

Fig. 3. Distribution of the number of subject headings over the tagged records

A following step is to categorize the user behaviors and to discriminate the purposes for social tagging. During the particular case study the one by one assiduous examination of the 582 tags emerged the following classes of tagging purposes and their frequency:
(a) 2.1 % of the tags correct the thematic description of a record, to propose more accurate terms for the particular bibliographic records.
(b) 80% of the tags enhance / refine the thematic description of a record, manifested by two partial behaviors:
(bi) uses terms that belong in the authority file as tags but not for the particular record,
(bii) adds new terms, disjoint from the authority file descriptors.
(c) 3.8 % are new terms, disjoint to the local authority terms, expressing new concepts or synonyms, denoting both correction and enhancement.

Moreover there exists a group of tags (13.4%) that does not contribute to the precise expression of the subject of the documents since they are the same with the most of subject terms of the corresponding records. It is crucial to mention that these classes were confirmed by the interviews with the taggers.

3. In the stage of Semantic Assessment was examined the semantic value of tags that are not included in local authority file of the library. For this purpose five systems were selected, namely the Library of Congress Authorities (LCSH), Greek National Documentation Centre (NDC) Thesaurus, Thesaurus of Social Sciences Index Terms (SSIT), Wikipedia and WordNet, and this selection is based on three criteria: coverage, language, relevance.

   The 269 “missing tags” were searched in these KOS either as a preferred or non preferred term (in Wilipedia as an article or proposed article, and any term in Wordnet) and the lexical overlap was very high in Wikipedia (61.7%), as Table 1 shows. For example, the tag “modernity” is not an authorized term in LCSH. Moreover the tag “social ontology” exists as a term in Wikipedia in some articles for social scientists, but there is not yet an article for it. Although its large size,

LC authorities cover the 34.6% of the 269 "missing tags" (28.3% in main entries and 6.3% as non-preferred terms). The main reason for this impressive coverage percentage might be the frequent update of the user-based KOS, which follows closely the vocabulary evolution of the scientific communities.

A next step in the same stage is the investigation of the semantic relation between the folksonomy tags and the local authority file terms. We formulated for each of the tagged of 245 bibliographic records ek a set of pairs (ti, sj) corresponding to all possible combinations of the tags (ti) and the subject headings (sj) used for the thematic description of a particular record. This procedure generated totally 1420 pairs, 1125 of which being unique. For each tag ti the records ei that include in their description both the tag ti and the descriptor sj were retrieved by the mentioned KOS.

Table 1: Number of tags that exist in other KOS (percentages inside parentheses)

	LCSH Authority	**NDC Thesaurus**	**SSIT Thesaurus**	**Wikipedia**	**WordNet**
Exist	76 (28.3)	26 (9.7)	35(13.0)	166 (61.7)	26 (9.7)
Not exist	176 (65.4)	229 (85.1)	234 (87.0)	66 (24.5)	243 (90.3)
Exist as non preferred	17 (6.3)	14 (5.2)	-	37 (13.8)	-
Total	269 (100.0)	269 (100.0)	269 (100.0)	269 (100.0)	269 (100.0)

Example: for the pair "archetype" - "Symbolism (Psychology)", it is found in LC authorities that the subject heading "Archetype (Psychology)" has an associative relation with the subject heading "Symbolism (Psychology)". The search of each relation opposed the full records of LC authorities, the Greek and Social Sciences thesauri, the WorldNet synsets for the tag ti ("archetype") and finally the "See also" terms occurring in the article entitled by the tag ti ("archetype").

The search showed that the majority of the pairs are not correlated in any KOS (60.6%). Once more, Wikipedia includes the majority of the correlated pairs, 28.8% of the total pairs were found.

The derived results could be explained by observing the significant differences in the philosophy and practices between social tagging and subject description.

4. The following stage, Overall Assessment and Exploitation considers two aspects of information management: (a) the micro decision making level, which focuses on particular actions and tasks regarding the inter-relations of tags and headings, and (b) the macro decision-making level, which outlines the vision of the information organization and the framework of its activities. The micro-level decisions includes the assessment and the performance of particular corrective actions on the local authority file, while the macro-level focuses on the policy development issues on social tagging by the information organization.

Concluding, the development of a policy for the exploitation of social tagging is equivalent to the establishment of a Library 2.0 environment in an information organization grounded on the concept of user collaboration and the design of collective information services.

5 Developing micro-level policies

These promising results triggered the design of a new experiment, which aimed to survey the subject cataloguers' opinion concerning the impact of the user community vocabularies to the local authority file evolution and the definition of a policy to converge the user-based and the expert-based subject indexing approaches.

A representative sample of 30 socially tagged bibliographic records was selected, which carried 72 subject headings, 66 being unique. The corresponding tags were gathered, totally 540, 120 being from OPACIAL and 420 from LibraryThing. The bibliographic records along with the corresponding subject headings and the associated tags were presented in a tabular form (Table 2 presents a part of the data).

Table 2. A sample of tagged records

Bibliographic Record	**Subject Headings**	**Tags**
Author: Weber, Max (1864-1920), Roth, Guenther (Editor), Wittich, Claus (Editor). Title: Economy and society: an outline of interpretive sociology / Max Weber; edited by Guenther Roth and Claus Wittich Publication: Berkeley, Calif. : University of California Press, c1978	Sociology Economics	19th century 20th century Europe Germany Verstehen Weber bureaucracy class structure economic sociology economics economy german history interpretation knowledge philosophy political economy political science political theory politics religion social theory society sociological theory sociology state the state theory world history Αξιολογική Ελευθερία Γερμανοί Φιλόσοφοι Κατανόηση Κοινωνιολογία

Then the Panteion University Library's subject librarians (9 cataloguers) were interviewed in order to (a) compare the expressive power of the local and the LibraryThing tags and (b) assess the semantic value of both the local and the LibraryThing tags, with respect to the corresponding subject headings that describe thematically the selected records. The focus of the discussion was on whether the tags correct, enhance or refine the subject description of the selected documents.

The findings of this study provide a great opportunity to the library staff to reconsider and evaluate the organizational schemes of subject indices, and to renew their content by adding new terms or relations. In particular the study addressed that the tags express directly the evolution of a scientific domain and the library should (a) create new subject descriptors, (b) substitute the current subject headings with more appropriate ones and (c) create references between the subject descriptors of the local authority file.

Concerning the results of the research, the interviews proved that OPACIAL has more representative and accurate tags than LibraryThing. In particular, the cataloguers "vote" for the 60% of OPACIAL tags are useful and more precise and 40% for LibraryThing. This finding is explained by the fact that OPACIAL serves a scholar community that uses a specialized vocabulary; on the other hand LibraryThing is a general-purpose collaborative cataloguing service.

All librarians confirmed that in general the tags enrich the subject description of the documents and they found a significant number of tags that are identical to authority records but not used for the thematic description of the particular records. This opinion was confirmed by the fact that only 21 tags were the same with the subject description of the selected documents, while the majority of the tags, 355 out of 540, are identical to the subject descriptors of the library authorities.

Indicative examples of this analysis are given in Table 2. The 2 subject headings of the record are included in the tag cloud. The tag cloud consists of 34 tags and 28 of them belong to the local authority. The evaluation of the tag cloud revealed that 11 of the tags could be used in the subject description of the record, while 2 of them are new terms.

Finally the librarians found that several tags constitute either new concepts or neologisms, or alternative translations of terms to the Greek language and admit that social tagging could help them to approach the user's way of thinking and help them more effectively as well as to observe the communities terminology evolution.

Regarding the macro-level of the library policy, two librarians proposed the creation of a wiki to enhance the collaboration of subject cataloguers and the faculty members for the disambiguation of the inserted tags, the apodosis of subject descriptors in the Greek language and in general the improvement of the library authorities.

6 A methodology for enriching library authorities

Given the mentioned analysis an interesting summative question for assessing the value of the social tags is whether they improve the information seeking performance. This investigation needs the adoption of the precision and recall metrics, probably modified by a new definition for the set of relevant returned records. Besides, another issue is the definition of a criterion for the incorporation of a social tag in the thematic description of a bibliographic record and to be added as new subject term in the library's authority file. For this purpose the following methodology is proposed:

(a) Examination of the overlap degree between the folksonomy and the authority file of the library. Examining the degree of overlap of social tags to the terms of

the authority file we intend to highlight the percentage of social tags, which represent new terminology for the subject description of resources of the library.
(b) Examination of the overlap degree between the folksonomy and the library catalog queries logs (searches based on the following indexes: subject, author, title, language, notes, publisher, series title, anywhere and the independent index based on social tags). The aim of this step is to define the percentage of queries that are new terminology for the subject description of resources of the library.
(c) Examination of the relevance degree between a social tag and the thematic description of an annotated bibliographic record with this particular tag. The relevance measurement arises as a combination of metrics, originated by two approaches:

(i) the social aspect, in which the popularity (the frequency) of the tags applied to a record is taken into account. This estimation could be based on well-known social tagging systems, such as Library Thing, in which the number of users who have applied it to annotate a resource accompanies each tag.
(ii) the content aspect, in which the frequency of a term generated by automatic indexing systems is taken into account. In this aspect the generated index terms that are common with the tags of a bibliographic record will be selected. This estimation could be exploit known automated indexing sources, e.g. Google books. Moreover the tf-idf metric could be used in this case, instead of measuring the frequency of each index term.

7 Conclusions and further research

As a matter of fact several open issues there exist to obtain a policy for the activation of users to collaborate for the generation of a Library 2.0 environment. First of all comparative user studies should be organized and performed so that to investigate in depth the hypothesis that users prefer an information services capable to integrate the social and the traditional knowledge organization approaches. Moreover significant research should be made on the convergence of tags of a folksonomy and other knowledge organization systems in order to fulfill the user's trend who demands such integration. Given this hypothesis significant effort should be made for the incorporation of the folksonomy tags in the ideas of information organizations. The work is laborious and demands the cooperation of both the users and subject cataloguers, as well as the exploitation of semantic web technologies and collaboration tools.

References

1. Stock, W.G. (2010). Concepts and Semantic Relations in Information Science. *Journal of the American Society for Information Science, 61(10)*,1951-1969
2. Quintarelli, E. (2005) Folksonomies: power to the people. In *ISKO Italy-UniMIB meeting. Milan*, Italy. http://www.iskoi.org/doc/folksonomies.htm.

3. Mathes, A. (2004). *Folksonomies - cooperative classification and communication through shared metadata. Report*. Graduate School of Library and Information Science, Illinois Urbana-Champaign. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html
4. Spiteri, L. F. (2006). The use of folksonomies in Public Library catalogues. *Serials Librarian 51*, 75-89.
5. Lin, X., Beaudoin, J. E., Bui, Y., & Desai, K. (2006). Exploring characteristics of social classification. In *Proceedings of the 17th ASIS&T Classification Research Workshop, Austin, Texas, USA.* http://dlist.sir.arizona.edu/1790/01/lin.pdf.
6. Kipp, M.E.I. (2006). Complementary or discrete contexts in online indexing: a comparison of user, creator, and intermediary keywords. *Canadian Journal of Information and Library Science 30(3)*. http://dlist.sir.arizona.edu/1533/01/mkipp-caispaper.pdf
7. Al-Khalifa, H.S., & Davis, H.C. (2007). Exploring the value of folksonomies for creating semantic metadata. *International Journal on Semantic Web and Information Systems, 3(1)*, 13-39.
8. Voss, J. (2006). *Collaborative thesaurus tagging the Wikipedia way*. http://arxiv.org/abs/cs.IR/0604036.
9. Yi, K., Chan, L. M. (2009). Linking folksonomy to Library of Congress subject headings: an exploratory study. *Journal of Documentation 65(6)*, 872-900.
10. Yi, K. (2010). A Semantic Similarity Approach to Predicting Library of Congress Subject Headings for Social Tags. *Journal of the American Society for Information Science, 61(8)*,1658-1672.
11. Thomas, M., Caudle D., Schmitz C. (2009). To tag or not to tag? *Library Hi Tech 27(3)*, 411-334.
12. Golder, S., & Huberman, B. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science, 32(2)*, 198-208.
13. Mendes, L.H. Quinonez-Skinner J., Skaggs D. (2008). Subjecting the catalog to tagging. *Library Hi Tech 27(1)*, 30-41.
14. Smith, T. (2007). Cataloguing and you: Measuring the efficacy of a folksonomy for subject analysis. In *18th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research, Milwaukee, Wisconsin, USA*. http://dlist.sir.arizona.edu/2061/01/Smith%5FUpdated.doc.
15. Bartley P. (2009). *Book Tagging on LibraryThing: How, why, and what are in the tags?* Available at: http://www.asis.org/Conferences/AM09/open-proceedings/papers/28.xml.
16. Pera, M.S., Lund, W., Ng, Y-K., (2009). A sophisticated library search strategy using folksonomies and similarity matching. *Journal of the American Society for Information Science and Technology, 60(7),* 1392-1406.
17. Lawson, K. (2009). Mining social tagging data for enhanced subject access for readers and researchers. *The Journal of Academic Librarianship 35(6)*, 574-82.
18. Heymann P, Garcia-Molina H. (2009). Contrasting Controlled Vocabulary and Tagging Do Experts Choose the Right Names to Label the Wrong Things? In *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09.*
19. Rolla, BPJ. (2009). User Tags versus Subject Headings Can User-Supplied Data Improve Subject Access to Library Collections? *Library, 53(3)*:174-185.
20. Lu, C. Park J-r., Xu, (2010). User tags versus expert-assigned subject terms: A comparison of LibraryThing tags and Library of Congress Subject Headings. *Journal of Information Science, 36(6)*:763-779.
21. Stvilia B, Jörgensen C. (2010). Member Activities and Quality of Tags in a Collection of Historical Photographs in Flickr. *Journal of the American Society for Information Science, 61(12)*, 2477-2489.
22. Gavrilis, D., Kakali, C., & Papatheodorou, C. (2008). Enhancing library services with Web 2.0 functionalities. In *Proceedings of the 12th European Conference on Research and*

Advanced Technology for Digital Libraries (Aarhus, Denmark, September 14 - 19, 2008). Lecture Notes in Computer Science, vol. 5173. Springer-Verlag, Berlin, Heidelberg, 148-159.

Geographic collections development policies and GIS services: a research in US academic libraries' websites

Ifigenia Vardakosta, Sarantos Kapidakis
Laboratory of Digital Libraries and Electronic Publishing, Department of Archives and Library Science, Ionian University
{ifigenia, sarantos}@ionio.gr

Abstract. Management and analysis of geospatial data evolved into a rapid developmental field nowadays. Scientific researches debate that 80% of economic and political decisions internationally, include indirect or direct geographic information while this is also present in everyday life under various applications (GPS, in PDA's, in mobile phones). The digital libraries offer various tools, including open systems that can be used in order to organize and accommodate the retrieval of a variety of geospatial data. In addition, institutional arrangements have facilitate the access to geospatial data setting the geospatial information a promising field for libraries that want to offer a variety of new services to their users. In order to investigate the GIS services and whether the libraries hold a geospatial collection,(had also established a collection development policy for it, we systematically reviewed, in March 2011, 133 websites of US academic libraries. This paper aims at tracing those libraries that use GIS services in order to make their geospatial collection, (either developed by subscriptions or by their own sources), accessible to the end user. The following elements were examined in the current research: 1) How many libraries provide GIS services? 2) How many libraries provide collection development policy for their geospatial collections to their patrons? 3) What kind of information do they offer? 4) What kind of infrastructure do they provide to the public? 5) What services do they offer? (user education, assistance, remote access, guidelines for hardware/software). We also aim to compare the results of our survey with the results of previous surveys in the field while we parallel the libraries we research in our survey to ARL, UCGIS, and FRPAA lists. The majority of the examined libraries offer GIS services, but only 14% of them currently inform their users for their collection development policy. The types of information that these collections sustain varies (gazetteers, maps, geographical data sets etc), while most of the libraries provide information about their infrastructure (workstations, printers, scanners etc). The main desktop software for 58% of the reviewed libraries that mention it is ArcGIS. As little previous research has been conducted on the topic of geospatial collection development policies and GIS services, this study is exploratory. Although the timing and the fixed duration of the study limited the size of the sample and the depth of the investigation, sufficient data were collected. This paper seeks to examine the potential role of policies in geospatial services that libraries can offer, in a rapidly changing digital environment.

1 Introduction

Libraries are facing new challenges with the distribution of geographic information in digital form. Most important, libraries will be challenged to manage geographic information in a new way since the opportunities include the spread of geographical information science and spatial analysis across disciplines, the ability to present map information in more dynamic forms than previously possible, the increasing information query, the interpretation and display capabilities, and the access to more current information [1]. Furthermore, libraries are challenged either to find ways to provide information in this area, or scholars, students and others users will obtain the information they need from sources outside the library [2].

A geographic information system is an appropriate tool for libraries whose primary function is the management (storage and distribution) of information. In a way, most libraries use GIS when they store and manage atlases and maps. Libraries are now expanding that traditional usage by employing computer based, automated GIS capabilities. This expansion is spurred by the rapid development of computer hardware and software capabilities [1]. The advantage of having a unifying GIS platform in which users may combine otherwise disparate data sources is attracting an ever increasing number of users [3]. An academic library's homepage mainly functions as a public service, typically including digital reference, online interlibrary loan request forms and online information tutorials, to name a few [4]. Thus, it is reasonable to ask, however, whether there is a GIS services role in a library's homepage.

2 Literature Review

2.1 Collection Development Policies

GIS data collection development constitutes a core element of GIS services within libraries and information centres. As "collection development is a process that allows the identification of the strengths and weakness of the materials collection of a library in terms of users needs and the resources of the community [5], in the creation of GIS collection development policy, library professionals should consider the established collection development policy, needs of the GIS user community, and library infrastructure. Additionally, information professionals should examine the current and planned GIS activities in the institution, which will have strong influence on their form [6].

When making decisions regarding GIS data acquisition, the decision maker should consider cost, availability, license agreements and distribution policies, documentation, data structures, software and hardware [7]. The policies could differ from one organization to another in the sense that each one has its own requirements and priorities [8] but the most important and helpful for librarians is "to incorporate

elements of a need assessment into their workflow to help organize the various types of information elements they collect" [9].

2.2 GIS services

Geographic Information Systems (GIS) are designed to allow the management of large quantities of spatially referenced information about natural and man-made environments, covering areas such as public health, urban and regional planning, disaster response and recovery, environmental assessments, wetlands delineation, renewable resource management, automated mapping/facilities management, and national defence [10]. GIS platform in the library opens many new gateways and provides several opportunities to the libraries for contributing their share in planning and decision making in the area of handling geographic information, which they did not avail earlier. It is possible to answer a variety of queries put by patrons working in different fields [11].

3 Methodology

The websites of 133 US academic libraries funded by either the public or the private sector were examined in March 2011. We choose academic libraries because:

- academic libraries support a wide part of the society,
- they have more reliance on new technologies,
- of the quantity of US academic libraries,
- of their history in the implementation of GIS services [12].

The objectives of the review were: 1) How many libraries provide GIS services? 2) How many libraries provide collection development policy for their geospatial collections to their patrons? 3) What kind of information do they offer? 4) What kind of infrastructure do they provide to the public? 5) What services do they offer? (user education, assistance, remote access, guidelines for hardware/software). Specific information regarding the examined questions was recorded in Excel sheet.

4 Limitations of the research

Among the limitations of the study we include the specific library type and the geographic region as we examined only academic libraries' websites in the US. As the author was the sole researcher we can't exclude any amount of bias into the analysis.

5 Research Results

5.1 GIS in Academic Libraries services

95 out of the 133 libraries we examined appeared to have GIS services for their academic community while 17 out of 133 provide GIS services either cooperatively with an academic department or they offer such kind of services in a Center[1] or a Lab[2]. Overall, 21 out of 133 libraries did not offer GIS services at all.

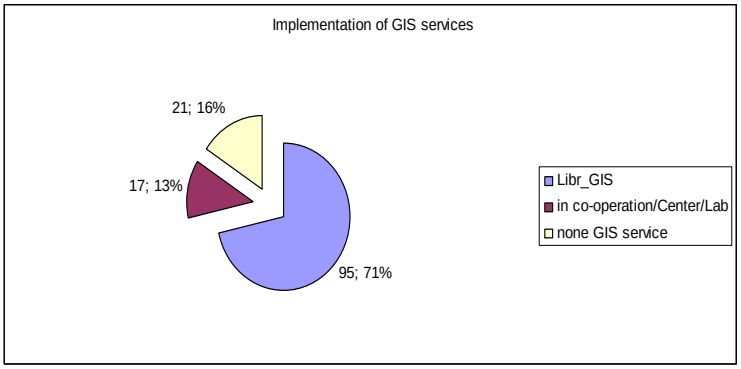

Fig. 1. Implementation of GIS services

5.2 Collection Development Policies

Of the 95 academic libraries that had established GIS services, only in 13[3] (14%) we located a geospatial collection development policy. The majority (82/95, 86%) did not have any information on their webpage about such policies.

[1] University of Cincinnati, http://www.gissa.uc.edu/
[2] University of Denver, http://www.du.edu/gis/dataresources.html.
[3] Colorado State Library, Duke University, Emory University Library, Portland State University Library, Stanford University-GIS at Branner, University of Connecticut, University of Georgia, University of Hawaii-Manoa, University of Iowa, University of Nebraska-Linkoln, University of Wieconsin-Madison, University of Wisconsin-Milwaukee, Western Michigan University.

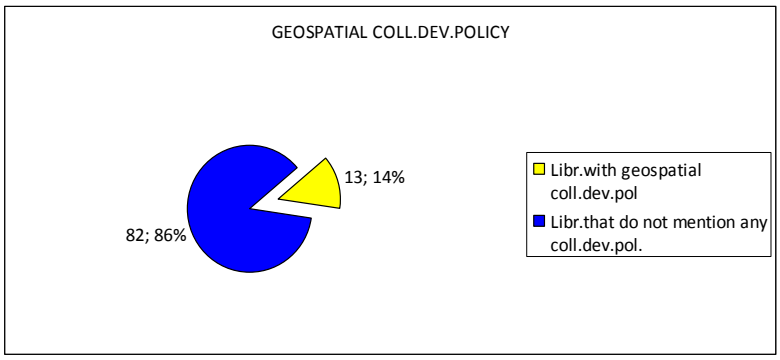

Fig. 2. Geospatial Collection Development Policy

5.3 Provided information

As most of the libraries (85/95) were members of the Federal Depository Library Programme (FDLP)[4], patrons have the chance to access a variety of Government's Information (local base data, national data sets, data from federal agencies, etc). In conjunction with ESRI, the majority of libraries provide data as tutorials as well. The growing use of electronic information is particularly obvious in the specific type of information sources and GIS Librarians have organized their websites in a form that provides access to several free electronic resources containing either national, local or international data in several topics (e.g. labor statistics, US Census Data, International Financial Statistics). Scanned historic maps, interactive maps, digital orthophoto files, satellite imagery, aerial photographs, aeronautical charts, atlases, gazetteers and thesauri, shapefiles, are some of the collections that patrons can find and use, for their educational or scientific purposes in a diverse variety among 95 libraries with GIS services. Of course, except for the above, more "traditional" collections are also available like journals, databases, books and dictionaries that cover GIS aspects.

5.4 Infrastructure

Hardware: 46 out of 95 libraries provide information on the infrastructure that can be used in the library and which contains: workstations, printers, scanners, plotters, GPS. We note that infrastructure availability varies from library to library, although there are libraries[5] that have the ability to offer all the above in order to cover their users needs.

[4] The Federal Depository Library Program (FDLP) was established by Congress in 1813 to ensure that the American public has access to its Government's information (http://www.fdlp.gov/home/about)

[5] Pennsylvania State University, Rice University-Fondren Library, Washington University in St. Louis

Software: As ESRI was partner in ARL GIS Literacy project it is not a surprise that 58% use ArcGIS. We also detect information about other software packages like Auto Cad, Idrisi/Erdas, SPSS, as well as open source software like GRASS, QuantumGIS, DIVA, MapWindow, GoogleEarth, GoogleEarthPro that were provided either for educational purposes or for developing specific applications.

5.5 Services

According to their websites, 51 % of the 95 libraries organize training programmes, while 77% offer assistance to the users (e.g. Ask a Librarian). A patron can find information about hardware/software that can be used in 44% of the GIS libraries, while guidelines for data/software use, provide only 16%. The majority of examined libraries (67%) provide data for local access.

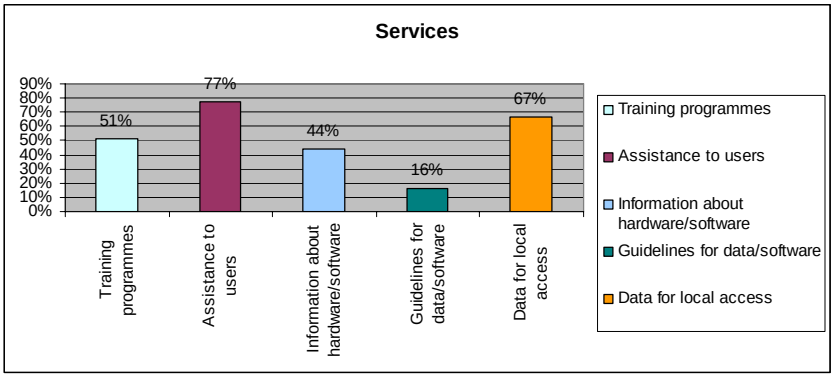

Fig. 3. Services

5.6 Other Findings

While searching websites for answering our 5 main research questions we record some additional interesting findings, useful for further research: Among 95 libraries offering GIS services and support to a community of users focusing on services such as information literacy, access, infrastructure, based on a static collection housed in a specific location accessible via the internet, it is rather interesting that 9 developed such services while didn't serve any familiar department (e.g. Geography, Environmental Science, Architecture etc.).[6] Besides data and infrastructure, personnel is the third important element for the whole service in order to be properly delivered to the public and 11 libraries adopted the term "GIS Librarian", recognizing that way that "in addition to the requisite skills needed by librarians in today's hybrid libraries,

[6] American University Library, Emory University Libraries, Georgetown University, Miami University, Michigan State University, New York University, Oklahoma State University, University of Pennsylvania, Washington University in St. Louis.

additional skill sets are basic for librarians who want to work with geospatial data" [12]. For our research, we proceed on examination of ARL GIS Literacy Project list and exclude all other kind of libraries except for academic ones (college and public libraries and libraries of Canada that we intend to research in future survey). Library of Congress and National Libraries, excluded as well as they were not in target group, in this initial examination. For the academic libraries left, we examined their websites and discovered that 58% were members of the initial ARL project and continue to supply their patrons with such kind of services.

In 1994, representatives of 34 US universities and other research organizations met in Boulder, Colorado and decided to establish an organization "dedicated to the development and use of theories, methods, technology, and data for understanding geographic processes, relationships, and pattern" [13]. We conclude that 46% out of 95 libraries we trace offering GIS services were members of the UCGIS as well.

Finally, as lately there has been a lot of discussion about the need to make research results accessible to a worldwide readership, and having in mind FRPAA[7], we discovered that 35% of libraries were in institutions that their presidents and provosts support it.

6 Previous researches

ARL conducted a survey in 1999, to examine the way the ARL libraries have organized their delivery of GIS in the years after GIS Literacy project began[8], and 64 institutions indicated that they provide GIS services (in 53/64 services administered by library). Kinikin and Hench [14] survey in small academic libraries, in 2002 indicated that 22 (out of 168 libraries which joint the research), support GIS services and proved that GIS services in academic libraries in the United States tend to differ, based on availability of GIS data, software, hardware and staff expertise. Kinikin and Hench [15] conducted in 2004 a follow-up survey of those libraries which had adopted GIS and they discovered that out of the eleven libraries which returned the survey, two have discontinued offering GIS services in their libraries.

Until the last decade, hardware and training costs were often prohibitive for all but the largest institutions. Larger and well-funded institutions have been able to overcome these barriers by hiring full-time staff to work with students and faculty, and to collect data and data sources as Gabaldon & Repplinger concluded in their

[7] FRPAA would require that 11 U.S. government agencies with annual extramural research expenditures over $100 million make manuscripts of journal articles stemming from research funded by that agency publicly available via the Internet (http://www.arl.org/sparc/advocacy/frpaa/index.shtml)

[8] By the early 1990s libraries were receiving large quantities of government documents, but many of them, lacked the system components necessary to allow the information to be used more effectively. ARL in partnership with Environmental Research Institute, Inc. (ESRI), launched the GIS Literacy Project in 1992. Member libraries were invited to send one or two of their librarians to ESRI for free training in using that company's software, which was also furnished free of charge (ARL, 199).

research among 103 institutions in two consortia in the United States in 2006. Sorice [17] in her master thesis argues that "no current resource exists that lists academic libraries providing GIS services, so the first challenge was identifying potential candidates before undertaking any evaluation". She examined how academic libraries present GIS services on their websites and identified potential barriers that the websites may pose to users. In this way she identified 35 out of 69 academic libraries from the ARL/GIS Roster and then she chose 6 eligible websites for content analysis. Finally, Good [18] in his research concluded that approximately 90% of academic libraries in the United States developed GIS services.

Table 1 indicates that the percentage of GIS implementation varies in the last decade although we cannot proceed in any reliable comparison among these previous surveys because of the differences in the methodology and the way they treated the libraries they targeted. Nevertheless we can argue that our research comes to a point: geospatial collections are active part of those academic libraries that are willing to offer high quality services to their patrons.

Table 1. Researches for GIS Implementation in US libraries

Research	**Percentage of GIS implementation in libraries**
ARL (1999)	64/72 (89%)
Kinikin & Hench (2005)	22/138 (20%)
Kinikin & Hench (2005a)	9/11 (82%)
Gabaldon & Repplinger (2006)	31/103 (31%)
Sorice (2006)	35/69 (51%)
Good (2009)	~90% in academic libraries
Our research (2011)	95/133 (72%)

7 Discussion

The paper aims to identify the percentage of libraries offering geospatial collections through GIS services and the percentage of those libraries that established collection development policies. As we conclude, despite the fact that GIS technology and services are popular within the university research environment, only a very small amount of libraries have developed collection development policies. In an academic environment, collection development policies need to support teaching, research, and applications [6]. Collection development is not what it used to be. It has changed considerably in the last ten years by changes in publishing, scholarly communication, technology, and budgeting. Developments in these areas have redefined what a library collection is, how it is acquired, and how it is used. All the above are obvious, but the degree of the resistance to change and the addiction to the status quo, in collection management organization and procedures in academic libraries belies our acknowledgement of the obvious. This is reflected in the structure most commonly employed in academic libraries, a structure that has been in place for two decades or more [19]. For policies to be fully effective, users must understand them. Information about policies, the levels of GIS services users can expect from academic libraries and

what kind of GIS resources are available need to be clearly communicated to users through GIS services websites [17].

Our findings demonstrate a recent contribution to the field while also raise some questions for further research. The investigation of developed policies in college and public libraries enhanced by active GIS services, as bibliography refers[9], would offer more insights in the way geospatial collection development policies affect those established services.

Fig. 4. Academic Libraries with Geospatial Collection Development Policies

8 Conclusions

As the recent socioeconomic trends and the convergence of telecommunication technologies have had significant effects on geospatial information spread [9] digital libraries can offer a more varied information experience to the community of online users. GIS services in academic libraries represent an evolution of traditional information services and undeniably offer a holistic learning environment. For this achievement to be better accomplished, defined policies should be followed.

References

1. Cox, A., Gifford, F. (1997). An overview to GIS. *The Journal of Academic Librarianship, November*, 449-461.

[9] New York Public Library, Franklin Park Public Library (http://www.franklinlibrary.org), Middlebury College Library and information Services (Macfarlane and Rodgers, 2008)

2. Boisse, J., Larsgaard, M. (1995). GIS in Academic libraries: a managerial perspective. *The Journal of Academic Librarianship, July*, 288-291.
3. Shawa, T.W. (2006) Building a system to disseminate digital map and geospatial data online. *Library Trends, 55(2)*, 254-263.
4. Wang, J., Gao, V. (2004) Technical services on the net: where are we now? A comparative study of 60 websites of academic libraries. *The Journal of Academic Librarianship, 30(3)*, 218-221.
5. Evans, E. (1987) *Developing library and Information Center*. Libraries Unlimited: Littleton, Co.
6. Longstreth, K. (1995) GIS collection development, staffing and training. *The Journal of Academic Librarianship, July*, 267-274
7. Florance, P. (2006) GIS collection development within an academic library. *Library Trends, 55(2)*, 222-235 .
8. Sanchez Vignau, B.S., Meneses, G. (2005) Collection development policies in university libraries: a space for reflection. *Collection Building, 24(1)*, 35-43.
9. Abresch, J. et al. (2008) *Integrating GIS into Library Services: a guide for academic libraries*. Hershey: Information Publishing Company.
10. Hanson, A., Heron, S. (2008) "From print formats to digital: describing GIS data standards". In Abresch, J. et al. *Integrating GIS into Library Services: a guide for academic libraries*. Hershey: Information Publishing Company.
11. Phadke, D.N. (2006) *Geographical Information Systems (GIS) in Library and Information services*. New Delhi: Concept Publishing Co.
12. ARL (1999) *The ARL GIS Literacy Project. Spec Kit 238*. http://www.eric.ed.gov/PDFS/ED429609.pdf.
13. Mark, D.M. (1999) *Geographic information science: critical issue in an emerging cross-disciplinary research domain*, [http://www.ncgia.buffalo.edu/GIScienceReport.html.
14. Kinikin, J.N. and Hench, K. (2005) Survey of GIS implementation and use within smaller academic libraries. *Issues in Science and Technology Librarianship.*
15. Kinikin, J.N. and Hench, K. (2005) Follow-up survey of GIS at smaller academic libraries. *Issues in Science and Technology Librarianship, Summer*.
16. Gabaldon, C., Repplinger, J. (2006) GIS and the academic library: a survey of libraries offering GIS services in two consortia. *Issues in Science & Technology Librarianship, 48, Fall*, http://www.istl.org/06-fall/refereed.html.
17. Sorice, M. (2006) *An analysis of GIS services websites in academic libraries.* Master Thesis, http://etd.ils.unc.edu/dspace/handle/1901/303.
18. Good, H.N. (2009) Trend of GIS services in US academic libraries: from comparison of past surveys and current situation of the University of Pittsburgh. *Information Science & Technology Association, 59(11)*, 539-544.
19. Nabe, J. (2011) Changing the organization of collection development. *Collection Management, 36*, 3-16.

Information seeking behavior of Greek astronomers

Hara Brindesi and Sarantos Kapidakis
Laboratory on Digital Libraries and Electronic Publishing
Archive and Library Sciences Department, Ionian University
hbrinde@eugenfound.edu.gr, sarantos@ionio.gr

Abstract: This study examines three aspects of information seeking behaviour of astronomers in Greece including a) the importance they place in keeping up-to-date with current developments b) the methods they depend on for keeping up-to-date and c) the information sources they mostly use. We adopted an intradisciplinary approach in order to investigate similarities and differences in information seeking behaviour among astronomers when examining them as groups bearing different characteristics, including academic status, subfield-research area of astronomy, age, and affiliated institution. The analysis of our results a) revealed that although some similarities exist, there are significant variations in the behaviour of the different groups of our participants, and b) highlighted the need for deeper investigation of narrower subject communities within disciplines in order to acquire deeper understanding of their information seeking behavior.

Keywords: Information seeking behaviour, User studies

1 Introduction

Information seeking behavior studies have always been of the main concerns of librarians and information scientists. According to Wilson [1] "Information Seeking Behavior is the purposive seeking for information as a consequence of a need to satisfy some goal", and its "origins are found in work on the users of libraries and in readership studies in general".

Such studies aim at the evaluation of information collections [2], or the maximization of the efficiency of information services provided, specific to the field of study [3, 4, 5]. Furthermore, the study of the information behavior or habits of specific communities help detect users' habits and needs, hence making it possible to introduce the necessary instruction programs in information literacy, responding effectively to those communities' requirements [6, 7, 8].

Our area of study is the research related to information seeking behavior of astronomers. The last of the statements in the paragraph above, that is the information literacy programs, constitutes the main aim of our study. Moreover, we favor the concept of the domain-analytic paradigm in information science, which states that "the best way to understand information in IS, is to study the knowledge-domains as thought or discourse communities, which are parts of society's division of labor" [9]. Accordingly, we narrowed our research focus on astronomers, and particularly on Greek astronomers of the area of Athens, for in depth domain study and we detected

their habits and needs in order to introduce an information literacy program appropriate for their requirements.

This article presents part of the findings of the survey study which constitutes the first step of a PhD thesis. The main aim of this particular work is to examine three aspects of information seeking behaviour of Greek astronomers including a) the importance they place in keeping up-to-date with current developments b) the methods they depend on for keeping up-to-date and c) the information sources they mostly use. Furthermore, the study also uses an intradisciplinary approach in order to investigate similarities and differences in information seeking behaviour among astronomers with different characteristics, including academic status, subfield-research area of astronomy, age, and affiliated institution.

2 Literature review

Unfortunately, there is not much bibliography concerning the information seeking behavior of astronomers. What we have noticed is mainly studies about scientists in a general context, in which astronomers are included. For example, as Tenopir [10] mentions, "preferences of physicists are often studied, but astronomers are less often singled out for study".

In 1993, Ellis et al. [3] investigated the information seeking patterns of a group of social scientists, physicists and chemists using the grounded theory approach. The result of this study is the well known Ellis' model of information-seeking behaviour, which is constituted from five features for the information-seeking behaviour of the above mentioned group. The five features were: initial familiarization, chasing, source prioritization, maintaining awareness, and locating, and they were the same for everyone in the group regardless of their area of study.

However, as Hemminger [11] mentions, "when examining differences between subgroups most researchers have found specific differences. Hurd, Wheeler, and Curtis [12] found that chemists rely heavily on current journals. Mathematicians make more use of older material based on citation studies [13]. Physicists and astronomers have made more use of preprints due to the development of preprint servers (e.g., arXiv) in their field".

Characteristic example of Hemminger's remark is Brown's [14] study, who investigated astronomers, chemists, mathematicians, and physicists at the University of Oklahoma. The astronomers of her group, showed differences in their preferences, for example, as far as their visits to the library or the information sources they used is concerned. They made a lot of use of the library, in contrast to mathematicians, and they were more dependent on current journals, as well as on pre-print archives. In general, physicists and astronomers are heavy users of e-print archives [15], especially of the arXiv.org eprint archive, originally developed by Paul Ginsparg at the Los Alamos National Laboratory [10].

One of the most recent studies is that of Jamali and Nicholas [16]. The two researchers examined two aspects of information seeking behaviour of physicists and astronomers including methods applied for keeping up-to-date and methods used for finding articles. They concluded that "there are significant differences among

subfields of physics and astronomy with regard to information-seeking behaviour in terms of their reliance on different methods used for keeping up-to-date as well as methods used for finding articles."

As we mentioned above, there are not any information seeking behaviour studies with a focus on astronomers, investigating similarities or differences among them according to their main characteristics, as for example, their academic status, research area, age, or affiliated institution. This study aims to fill this gap.

3 Methodology

The population of our study was restricted to the area of Athens, so we came into contact with the 18 professors of the Department of Physics and Astronomy of the University of Athens, as well as the 41 researchers of the Academy of Athens and of the National Observatory. In our sample we also included the 25 PhD and the 22 MSc students of the University of Athens. The total number of people that constitute our population is 106.

Firstly, thirteen (13) face-to-face semi-structured interviews were conducted. The analysis of these interviews, as well as the study of the corresponding bibliography, helped us to set the online questionnaire, which was filled in by 71 recipients (68.8% response rate).

4 Main results

We present our results in three sections using simple descriptive statistics. The three sections are the following: a) Interest in keeping up to date with current developments, b) Methods used for keeping up-to-date, and c) Information sources usage.

As a general remark we could mention that the MSc students present differences in their behaviour in comparison to the other groups. So, we have conducted the analysis both with and without the answers of that particular group, particularly in the cases we had the feeling that we would come up with distorted results.

4.1 Interest in keeping up to date with current developments

In this section we present the results from the two relative questions we had included in our questionnaire. The first question was: "How important is rapid awareness of new papers for you?" The participants had to choose among the following rate of options (Not at all important/ A little important/ Somewhat important/ Quite important/ Absolutely important). The second question was: "How many hours a week do you spend for keeping up with current developments?"

The majority of the respondents to our research, as it might be expected, deem absolutely necessary keeping up-to-date with the latest papers, as 52.1% ticked the option "absolutely important" for the first question. Moreover, nobody (0%) chose the

option “not at all important”. The rest of the responses to that same question were the following: “a little important” (2.8%), “somewhat important” (18.3%), “quite important” (26.8%).

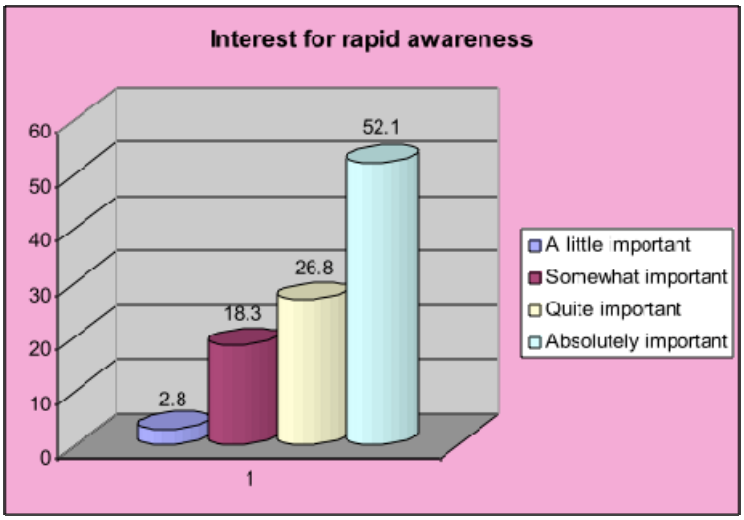

Fig. 1

Further analysis of our results revealed that levels of importance varied depending on the status of the respondents. Professors and researchers show greater interest in keeping in touch with current developments in comparison to PhD and MSc students. “Absolutely important” was the most popular answer among professors (75%), researchers (62.5%) and PhD students (52.4%). MSc students were the only group of which the majority of respondents ticked “somewhat important” (36.4%) and “quite important” (36.4%).

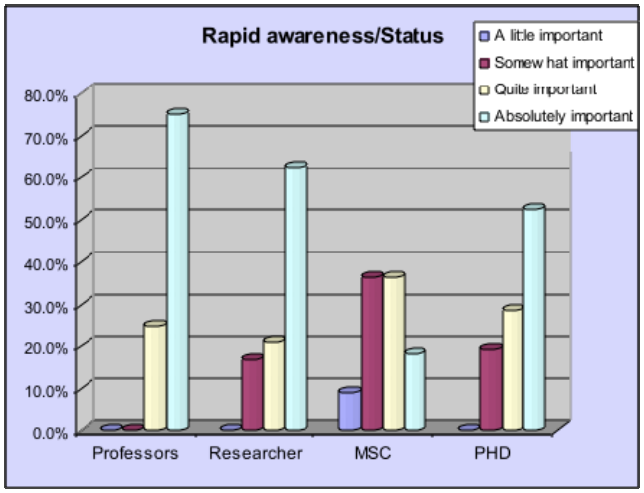

Fig. 2

We tried excluding the MSc students from the general results of that question and as a consequence we took different values, with a transposition towards "absolutely important", "not at all important" (0%), "a little important" (1.6%), "somewhat important" (14,5%), "quite important" (24.2%) and "absolutely important" (59.7%).

Jamali and Nicholas [15] had included a relative question in their research. We can have a comparison with our results, although unfortunately they don't give us any data concerning exclusively the astronomers. In their article 'Information-seeking behaviour of physicists and astronomers' they mention that "the majority of their respondents believed that it was important for them to keep up with the developments of their subfields. A quarter of the respondents considered keeping up-to-date as absolutely critical for their research. Fifty-five per cent ticked the option 'quite important'. Looking at the academic status of the respondents, it turned out that those who associated less importance with keeping up-to-date were more likely to be PhD students or research fellows."

Moreover, levels of importance varied when examining different subfields of astronomy (Fig. 3). 100% of astronomers in the subfields of cosmology as well as history and philosophy of astronomy expressed the view that keeping up-to-date is of quite to absolute importance to them. However, the rest of the participants valued keeping up to date less: that is, 90.5% of participants in space physics, 76.9% in stars, 70% in astrophysics, 66.7% in extragalactic astronomy, and 60% in dynamical astronomy stressed that keeping up to date is of quite to absolute importance.

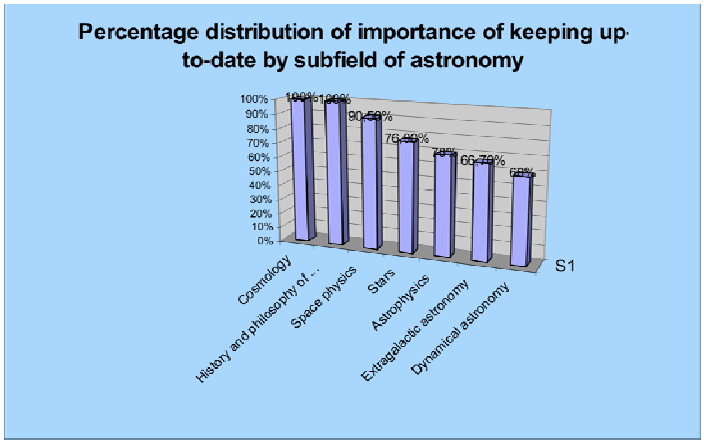

Fig. 3

Statistical manipulation of responses on the second question, i.e. "How many hours a week do you spend for keeping up with current developments?" revealed that astronomers in Greece spend on average 7 hours per week in keeping up to date (median and mode= 5, the minimum time they spend= 0 hours, the maximum= 30 hours).

Looking into the amount of hours per week astronomers in Greece spend on average keeping up to date varies according to their status (Fig. 4). Specifically, professors spend on average 9 hours per week, that is, more time than any other group. Researchers and PhD students spend 7 hours per week. MSc students present very low rates (Mean= 4 hours/week, mode=2).

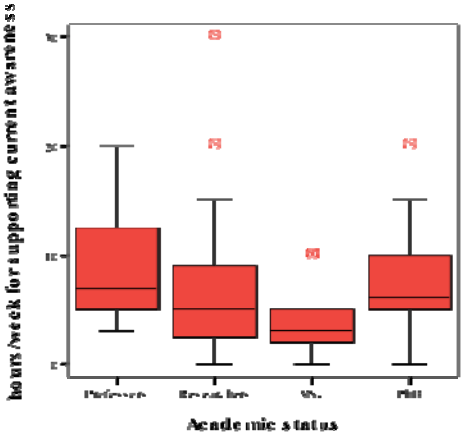

Fig. 4

These findings resemble the results of the former question "How important is rapid awareness of new papers for you?" where we found that professors show greater interest in keeping up with current developments, in comparison to any other group.

The amount of hours per week astronomers in Greece spend on average keeping up to date varies also according to the subfield of astronomy they work on (Fig. 5). Cosmologists spend the most amount of hours keeping up-to date than any other category (Mean= 14.33 hours/week). This category is followed by the subfield of History and philosophy of astronomy (Mean= 14 hours/week). In the figure below the differences in the time our respondents spend according to their research area are obvious.

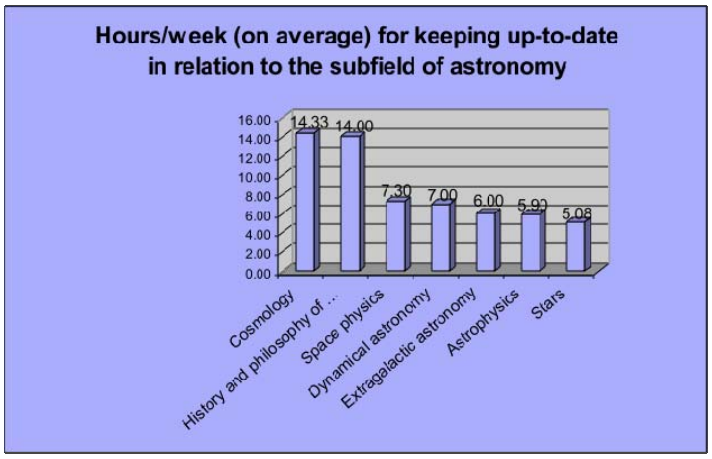

Fig. 5

These findings resemble the results of the former question "How important is rapid awareness of new papers for you?", where we found that astronomers in the subfields of Cosmology as well as History and philosophy of astronomy expressed the view that keeping up-to-date is of quite to absolute importance to them more strongly than any other category.

Differences in the levels of the time spent for keeping up-to-date are observed also among the respondents occupied in different institutions. As the figure below shows, researchers of the Academy of Athens dedicate the most time in comparison to the scholars of the other two institutions (Mean= 12.82 hours/week, median= 10, mode =10). The corresponding values for the University of Athens and the National Observatory are: Mean= 6.67 hours/week, median= 5, modes= 2.5 and Mean= 4.23 hours/week, median= 5, mode= 5, respectively.

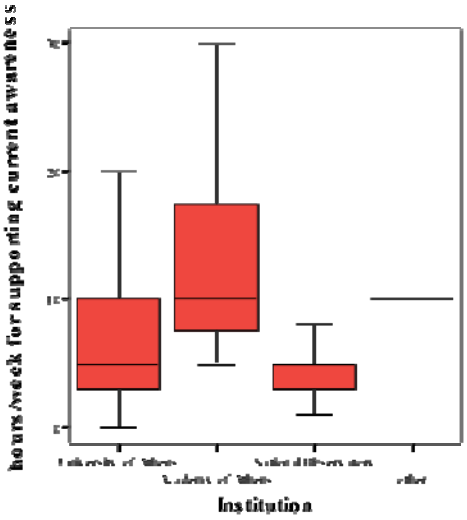

Fig. 6

The following figure shows that there are no great differences in keeping up-to-date as far as the various age groups are concerned, except for the groups 18-24 and 25-34 that seem to show lower interest in comparison to the rest.

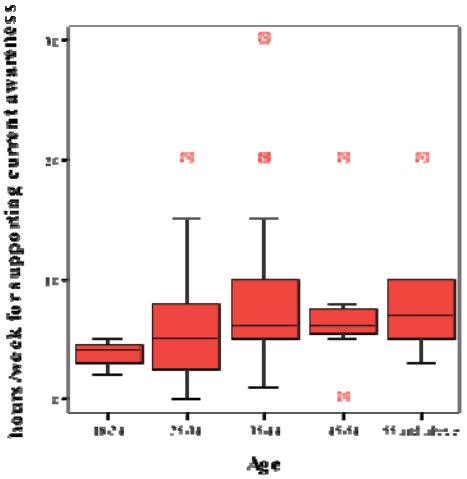

Fig. 7

4.2 Methods used for keeping up-to-date

In order to investigate this aspect of information seeking behaviour of Greek astronomers we included in our questionnaire the following question: "How dependent are you on each of these methods for keeping up-to-date with current

developments?" For each of the cited methods, the respondents had to choose among rated options ranging from "Not at all necessary" to "Absolutely necessary".

For the analysis of this question we used the percentage of the option "Quite necessary" additionally with that of the option "Absolutely necessary". The most popular methods that our respondents rely on for keeping up with the developments in their field are the conferences and their colleagues (Fig. 8). 81.2% of our respondents chose each of these methods in the relevant question. 71% chose the conduct of regular searches on the Internet, and 68.1% the seminars. Lower on this list are the regular browsing of ADS (63.8%), of arXiv (62.3%), and of e-journals (55.1%). It is interesting that the email alerts of ADS are not used heavily (31.9%), as, according to what the interviews showed us, astronomers prefer the regular browsing of the database. The same is true for the email alerts of the e-journals (31.9%). Less necessary are considered by our respondents the newsletters (24.6%), the classic browsing of printed journals (17.4%) and the publishers' catalogs (5.8%).

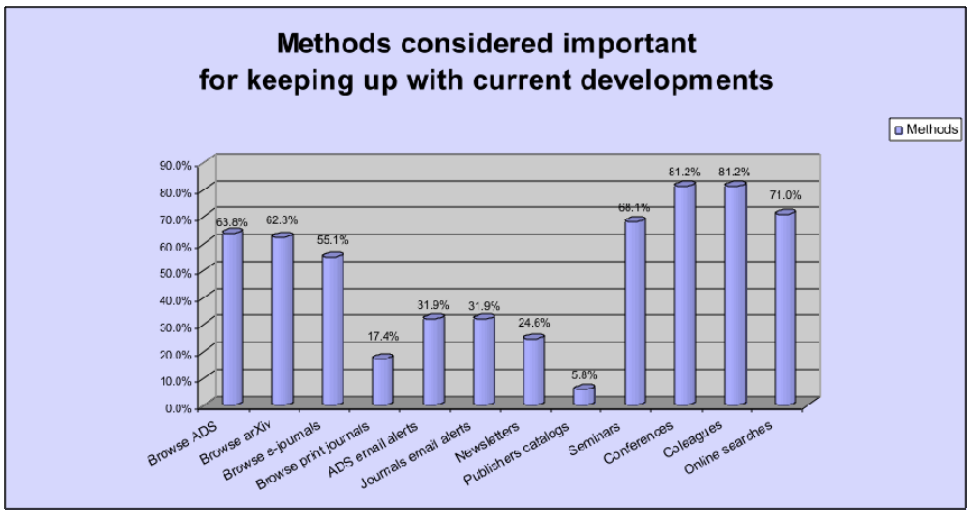

Fig. 8

The analysis of the results respectively to the research area of the participants is of special interest:

- The astronomers in the subfield of Dynamical astronomy don't use so intensely as the other groups the arXiv.org eprint archive, as well as the electronic periodicals for their keeping up-to-date. On the contrary they use more than the others the ADS email alerts. Furthermore, all the participants of this group chose the seminars as the most necessary method for their keeping up-to-date.
- All the participants in the subfield of Cosmology chose the conferences as the most necessary method for their keeping up-to-date. Moreover, they use ADS less than all the others.

- Astrophysicists cling more than all the other groups on ADS, as well as on the discussions with their colleagues following the scholars in the subfield of Extragalactic astronomy, who, in turn don't use any printed journals at all.
- The scholars of the research area of Space physics use more than all the others the email alerts of the e-journals, as well as the ADS email alerts, being second to the scholars of Dynamical astronomy.

We didn't observe remarkable differences as far as the age, the academic status and the institute that our participants are occupied.

4.3 Information sources usage

In order to investigate this aspect of information seeking behaviour of Greek astronomers we included in our questionnaire the question "How often do you use each of the following information resources for identifying the necessary information you need?" For each of the cited numbered sources, the participants had to choose among the following rate of options: Never/ once or twice a month/ 4-5 times a month/ 2-3 times a week/ Daily.

Apart from the answer to this question, we also asked our participants to mark the source (preceded by serial number) they consider primary source of information for their teaching, research, observations, keeping up with current developments, writing of articles, books, etc., personal updating, and their introduction into a subject area not well known. Our goal was to spot any differences in their preferences of sources respectively to their upcoming information needs.

For the analysis of the first question we used the percentage of the option "Daily" additionally with that of the option "2-3 times a week". By analyzing the question the results showed that (as it appears in the Fig. 9) the information sources mostly used (at least 2 to 3 times a week) are as follows: Google 88.20%, ADS 67.6%, websites 64.2%, electronic reference material 60.9%, ArXiv 58.6%, e-journals 55.40% and citations 54.3%.

Lower on this list are printed books (38.80%), electronic books (31.80%), Google Scholar (29.70%), colleagues recommendations (25.40%), library catalogs (22.10%), printed journals (21.70%), printed reference material (20.60%), databases for observations (17.10%), occupational meetings (conferences etc) (11.60%), ISI Web of Science (7,40%), and Web of knowledge (4,30%).

The main results concerning the usage of the information sources are the following (Fig. 9):

- ADS and Google is used by everyone in our sample, regardless of the subfield of astronomy our participants work on.
- Google Scholar is not used so often, especially if compared to the use of Google.
- Databases such as "ISI Web of Science" or "Web of Knowledge" are not so popular among the Greek astronomers.
- Wikipedia is being increasingly used.
- The use of printed material as well as of traditional libraries has been limited to a minimum, with the only exception of the printed books that are more popular than e-books.

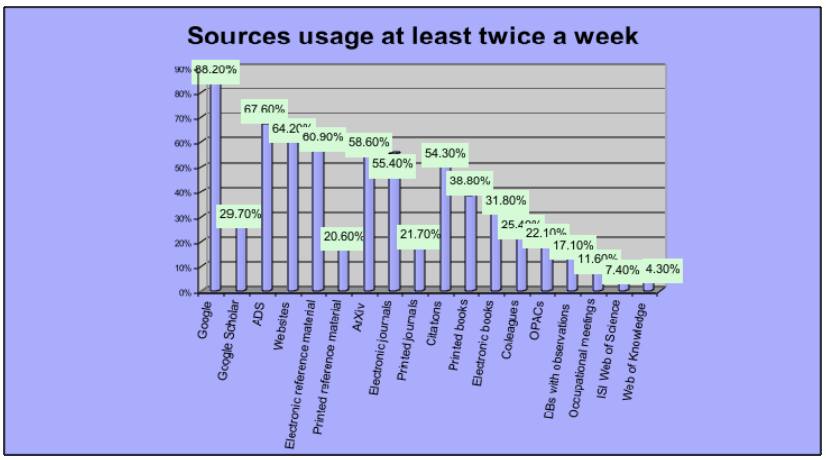

Fig. 9

The main results concerning the usage of the information sources in comparison to the status of the participants are as follows:

- Unlike the majority of astronomers, MSc students don't use ADS neither do they use arXiv.org database heavily, but they use mainly Google, reference material in electronic format and printed books.
- Journals and books in printed format, as well as Google Scholar are used mainly by professors.
- Books in electronic format are used mainly by PhD and MSc students.
- Citations are used heavily by researchers.

Differences were also observed while analyzing the results concerning usage of the information sources by the subfield-research area of the participants:

- The participants in the subfield of Dynamical astronomy don't use arXiv.org database so heavily, in comparison to the participants of the other subfields. Furthermore, Dynamical astronomy and History and philosophy of astronomy scholars use e-journals less often than the rest.
- Cosmologists use ADS less often than everyone. They equally cling on arXiv.org and e-journals as often as they cling on Google.

All of the above findings concerning subfield of Dynamical astronomy and of Cosmology resemble the findings of our former question "How dependent are you on each of these methods for keeping up-to-date with current developments?"

Moreover:

- Databases for observations and printed reference material are mostly used by the subfield of Stars.
- Websites are used less by the subfield of Dynamical astronomy.

The most remarkable results concerning the usage of the information sources in relation to the age of the participants are as follows:

- arXiv.org, as well as electronic library catalogs and electronic books are not used so much by the astronomers of 55 years old and above.

- The same age category uses mostly Google scholar and printed journals.
- The age category 18-24 uses mostly Google, websites, electronic reference material and printed books.

Furthermore, the differences among our participants concerning the usage of information sources in relation to their information needs, their academic status and the institute they are occupied are of special interest:

- Databases for observations (35.4%), as well as ADS (22.9%) were the tools of choice for all of our participants for support of observations data gathering. PhD students, in contrast to the other groups, indicated ADS (37.5%) to be the primary source for the same purpose. ADS database was chosen mainly by the University of Athens (90.9%). Researchers of the National Observatory (33.3%) chose Databases for observations (43.8%) and Websites (25%).
- For support of their research all of our participants indicated ADS (54.8%), arXiv.org (11.3%) and e-journals (9.7%) as primary sources of information. E-journals were chosen mainly by the PhD students.
- In order to keep abreast with current developments in their field the most important sources are arXiv.org (22%) and ADS (15.3%).
- For support of their teaching the participants chose arXiv.org (19.6%), websites (15.7%) and printed books (13.7%). 44% of professors chose arXiv.org in this question. Researchers of the National Observatory chose Google and printed books. Researchers of the Academy of Athens chose mainly Google.
- In order to discover information for writing books, articles etc. our participants use mainly ADS (39.3%), arXiv and Web of knowledge (9.8%), but also electronic library catalogs and e-journals (8.2%).
    - We tried excluding the MSc students from the general results of that question and as a consequence we took the following different values: ADS (43.6%), arXiv (10.9%), e-journals (9.1%), electronic library catalogues, websites and Web of knowledge (7.3%).
- For support of their personal information needs, websites (20%) and Google (18.3%) were mainly chosen. Lower on this list are arXiv.org (13.3%) and electronic reference material (8.3%). Electronic reference material is used only by researchers and PhD students.
    - When excluding MSc students our results are formed as follows: Google (19.6%), websites (17.6%), arXiv.org (11.8%) and electronic reference material (9.8%).
- Finally, in order to cover their investigation needs for a subject area not well known, our participants chose as the most important sources of information electronic reference material (24.6%), websites (19.7%) and then printed books, as well as ADS (8.2%). Printed books were chosen mainly by researchers of the Academy of Athens, and websites by PhD students.

5 Conclusions

The main aim of our study was the investigation of similarities and differences in information seeking behaviour among astronomers in Greece when examining them as groups with different characteristics, such as academic status, subfield-research area of astronomy, age, and affiliated institution. The analysis of our results showed that although some similarities exist, each of the above group has its own characteristics. This was confirmed through the analysis of all of the three aspects of information seeking behaviour we examined, that is a) the importance they place in keeping up-to-date with current developments b) the methods they depend on for keeping up-to-date and c) the information sources they mostly use.

For example, the majority of the respondents deem absolutely necessary keeping up with current developments. Complementary to that, and as far as the methods participants use for keeping up to date is concerned, there is high reliance on resources entailing human contact (e.g. seminars, colleagues, etc.) and informal communication. But, although there are such similarities, the levels of importance and of reliance varied depending on their status, research area, age, or affiliated institution.

Furthermore, as it happens with their colleagues from foreign countries, the astronomers in Greece highlight ADS (Astrophysics Data System) as their primary source of information. ADS is the well known NASA supported bibliographic database, which covers all the important literature for astronomers, and is freely available on the Web. In addition, everyone shows a preference to electronic sources of information versus the printed ones. But although there are general tendencies as far as the information sources usage is concerned, a lot of variations were observed when examining our participants as groups with different characteristics.

Concluding, our work revealed the need for deeper investigation of narrower subject communities within disciplines in order to acquire deeper understanding of the information seeking behavior of the users we study.

References

1. Wilson, T.D. (2000). Human information behaviour. *Informing Science*. http://inform.nu/Articles/Vol3/v3n2p49-56.pdf
2. Broadus, R.N. (1980). Use studies of library collections. *Library Resources & Technical Services, 24(4)*, 317–324.
3. Ellis, D. (1993). Modeling the information-seeking patterns of academic researchers: A grounded theory approach. *The Library Quarterly, 63(4)*, 469–486.
4. Kuhlthau, C.C. (1993). *Seeking meaning: A process approach to library and information services*. Norwood, NJ: Ablex.
5. Marchionini, G. (1995). *Information seeking in electronic environment*. Cambridge: Cambridge University Press
6. Hepworth, M., Wema, E. (2006) The design and implementation of an information literacy training course that integrated Information and Library Science conceptions of information literacy, educational theory and information behaviour research: a Tanzanian pilot study. *ITALICS, 5(1)*, www.ics.heacademy.ac.uk/italics/vol5-1/pdf/hepworth-evans-final.pdf

7. Pinto, M., Sales, D. (2007). A research case study for user-centred information literacy instruction: information behaviour of translation trainees. *Journal of Information Science, 33(5)*, 531–550.
8. Walker, J.R., Moen, W.E. (2001). *Identifying and Categorizing Information Seeking Behaviors in the Networked Environment: An Exploratory Study of Young Adults, School of Library and Information Sciences University of North Texas, Final Report.* http://home.swbell.net/walkerjr/ISBS/internetart.pdf
9. Hjørland, B. (1995). Toward a new horizon in information science (I.S.): Domain-analysis. In *ASIS 56's Annual Meeting. Columbus, Ohio, 25 October 1993*.
10. Tenopir, C., King, D. W., Boyce, P., Grayson, M., Paulson, K.-L. (2005).Relying on electronic journals: Reading patterns of astronomers. *Journal of the American Society for Information Science and Technology, 56*, 786–802.
11. Hemminger, B.M., Lu, D., Vaughan, K.T.L., Adams, S.J. (2007). Information seeking behavior of academic scientists. *Journal of the American Society for Information Science and Technology, 54(14)*, 2205-2225.
12. Hurd, J.M., Wheeler, A.C., Curtis, K.L. (1992). Information seeking behavior of faculty: use of indexes and abstracts by scientists and engineers. In *Proceedings of the 55th annual meeting on Celebrating change: information management on the move: information management on the move* (ASIS '92). American Society for Information Science: Silver Springs, MD, USA, 136-143.
13. Garfield, E. (1982). Journal citation studies. 36. Pure and applied mathematics journals: What they cite and vice versa. In *Essays of an information scientist. Vol. 5*, 1981–1982, pp. 484–492). Philadelphia: ISI Press.
14. Brown, C.M. (1999). Information seeking behavior of scientists in the electronic information age: Astronomers, chemists, mathematicians, and physicists. *Journal of the American Society for Information Science, 50(10)*, 929-943.
15. Jamali, H.R., Nicholas, D. (2008). Information-seeking behaviour of physicists and astronomers. *Aslib Proceedings: New Information Perspectives, 60(5)*, 444-462.
16. Jamali, H.R., Nicholas, D. (2009). E-print depositing behavior of physicists and astronomers: an intradisciplinary study. *Journal of Academic Librarianship, 35(2)*, 117-125.

List of Tutorials

Conceptual similarity: why, where, how
Michalis Sfakakis

Data exploration: representing and testing data properties
Spyros Veronikis

User studies: enquiry foundations and methodological considerations
Giannis Tsakonas