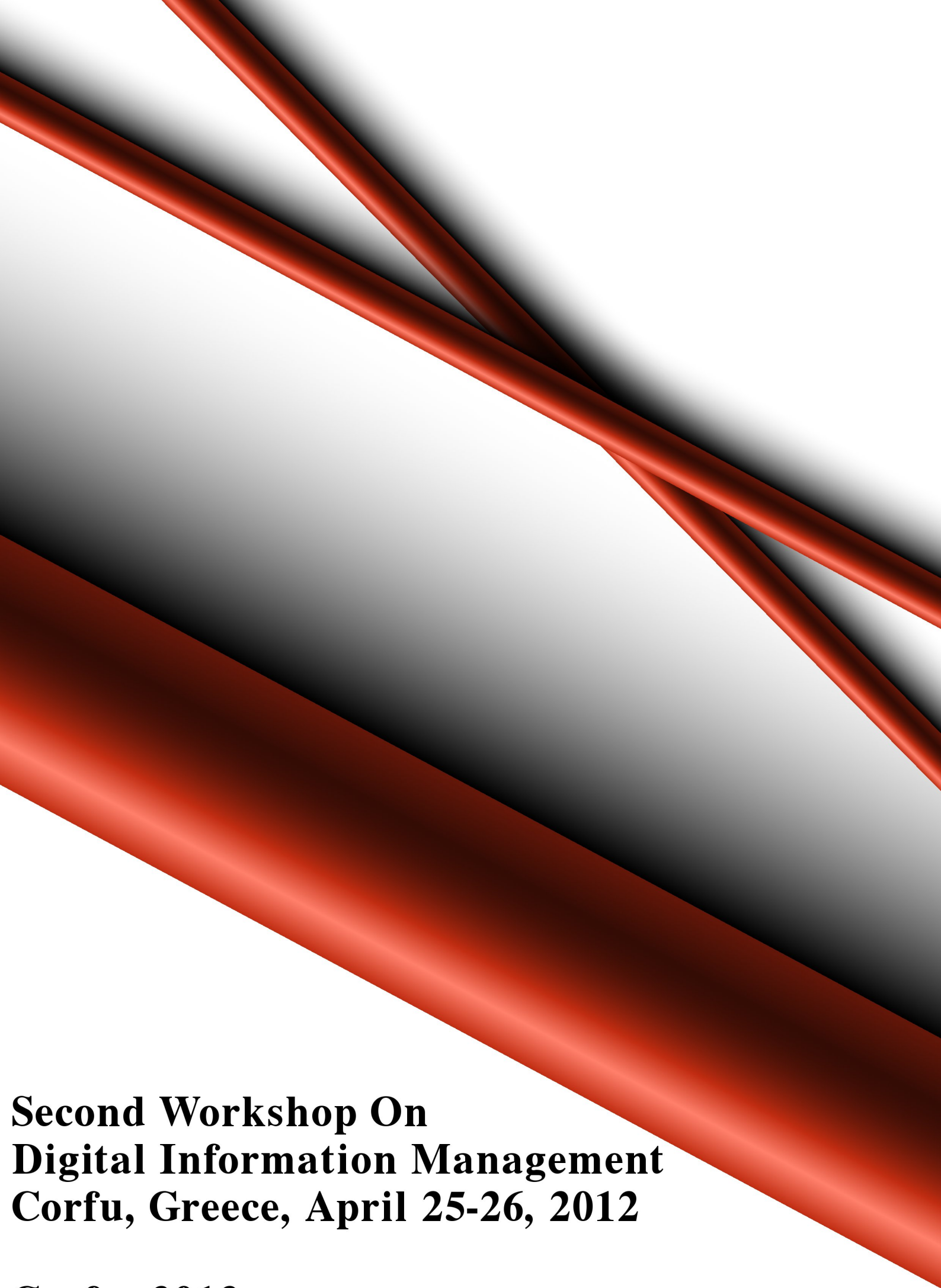


Second Workshop On Digital Information Management
Corfu, Greece, April 25-26, 2012

Corfu, 2012

Proceedings of the

Second Workshop on Digital Information Management

April 25-26, 2012 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
- XML and semistructured data management
- Information Integration
    - Metadata interoperability
    - Semantic interoperability
    - Semantic Web and linked data
- Knowledge organization and management
    - Ontologies and conceptual modelling
- Social Networking in information contexts
- User studies
- Evaluation of Information Services and Systems

Workshop Chairs

Manolis Gergatsoulis
Christos Papatheodorou

Program Committee

Manolis Gergatsoulis
Sarantos Kapidakis
Christos Papatheodorou
Marios Poulos
Maria Monopoli
Michalis Sfakakis
Spyros Veronikis
Giannis Tsakonas

Website

Eleftherios Kalogeros
Giannis Tsakonas

Additional Information
Webpage: http://dlib.ionio.gr/workshop2012/

Table of Contents

Developing a Metadata Model for Historic Buildings

Describing and Linking Architectural Heritage

Michail Agathos, Loukia Ventoura and Sarantos Kapidakis

Ionian University, Department of Archives and Library Science
Laboratory on Digital Libraries and Electronic Publishing
Ioanni Theotoki 72, 49100, Corfu

agathos@ionio.gr,ventouraloukia@gmail.com,
sarantos@ionio.gr

Abstract. This work presents the formulation of a meta-model for architecture heritage, a top-level data model that tries to encompass the common aspects of all the specifications in the domain of build heritage. The model incorporates definitions of some of the essential concepts that represent the underlying conceptualisation of the information contained in an architectural work. Finally the article presents ArMOS (Architecture Metadata Object Schema), a metadata schema underlying this model, aiming at reducing heterogeneity in the description of architectural works, especially historic buildings and structuring data so that it can be interlinked and become more useful.

Keywords. Architecture, Architectural Heritage, Conceptual Model, Metadata Standards, Meta – Model, Monument Inventories, Immovable Monuments, Semantic Interoperability.

1 Introduction

Monument inventories are the initial and most basic form of documentation that list the build heritage and describe their basic attributes, in their majority, host records of historic buildings, which form a rich store of information about the past, some of it unique. Despite the recognition for their necessity and their significance, the nature and structure of records of historic buildings have not been examined and researched from the aspect of metadata schemas like other material of our culture, and have remained until nowadays unexplored.

Ideally, a uniform metadata standard approach for historic buildings would ensure maximum interoperability for the encoding of information, but the diversity of the build 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.

In such a situation, the application of global conceptual models, metadata frameworks, or the definition of metadata mappings could solve the interoperability problems (Haslhofer et Klas, 2010). Historical background, constructional attributes, typological and morphological features represent information of a building. To depict such sophisticated information, it is important to recognize all different categories of the data, their attributes and their relationships and to conceptualize them in an integrated conceptual model. This approach will move us away from the traditional flat metadata descriptions of historic buildings, often including isolated and disconnected information in text-based record forms, which including pre-established entry fields in order to facilitate queries, into a fresh perspective on the structure and relationships of these records.

2 Linking Architectural Works

Architectural works such as historic buildings are complex works, consisting of multiple parts. A common cataloguing practice is to conceptually subdivide an architectural structure into multiple components, making several records for one building, including, for example, a record for the building as a whole, and additional records for each significant element (such as a chapel, portal, dome, and so on) and finally linking records together through whole-part relationships. The relationship between the whole and the parts, (e.g. monastery – church) also known as larger entity component or parent-child relationships, are intrinsic relationships. An intrinsic (direct) relationship is considered as an essential relationship between two works, as a part cannot be fully understood without its whole (the part inherits much of its information from the whole) and this type of relationship should always be recorded (CCO, 2006).

The built environment often involves architectural complexes in which each building is significant in itself, yet all are related in some manner (CCO, 2006). Two or more individual architectural works may have an informative relationship that could be considered as an extrinsic relationship. In the context of CCO, an extrinsic relationship is not essential; although doing so may be informative, the cataloger need not identify the extrinsic relationship during the cataloguing process. Instead, for the documentation of the built heritage, these extrinsic relationships must be identified and recorded, allowing us to understand continuities, similarities, and variations of our architectural heritage. This is supported by the fact that a historic structure is best understood in the context of similar structures, such as books in the context of the series (although this is an intrinsic relationship). In order to achieve this, in this work we precisely define the type of the relationship between individual architectural works (Table 1.).

Extrinsic relationships are generally temporal, conceptual, or spatial. A Temporal relationship may include architectural works for which, a typological or morphological examination has shown that one building is a predecessor or successor of the other. A conceptual extrinsic relationship can link structures, which could be considered as a source of inspiration, influence or variation one for the other. An extrinsic relationship can also be the result of a spatial association, such as two or more works intended to be seen together. However, other spatial relationships could also be

discovered and defined, concerning attributes of positioning, facing, proximity or visibility for architecture (Flanagan, 2011).

These connections between individual architectural works allow us to "wander" through the evolution of architecture. Moreover all these different types of extrinsic relationships should be reciprocal so that a search for a specific architectural work can lead to one other.

Table 1. Relationship Types in Architecture

Relationship Type	*Reciprocal relationship type*
Intrinsic Relationships	
is part of	larger context for
formerly part of	formerly larger context for
Temporal Extrinsic Relationships	
successor of	successor is
predecessor of	predecessor is
designed after	designed before
Conceptual Extrinsic Relationships	
influenced from	influence for
is inspired from	inspiration for
variation of	variation for
Spatial Extrinsic Relationships	
is similar to	similar of

3 A Meta – Model for Architectural Composition

A domain model is a conceptual model identifying the entities we want to describe, the relationships between them and the attributes necessary to effectively describe the entities. It acts as a communication tool and should be understandable by technical and non- technical audiences (Diamantopoulos, et. al., 2011).

In the architectural domain so far, many ontologies have been created for design tasks, but there is a need for a model with a fresh perspective, in order to modify the traditional and static conception of the history of architecture, which considers immovable monuments as finished and irremovable objects. In exchange, this new movable perspective of continuous transformation will allow us to understand the built environment in a better way (Casanova et.al. 2011).

The model described bellow is based on the ability to group buildings logically (as FRBR in bibliographic records), connecting them and to facilitate the discovery of all instances of a particular building type in a single search, while being able to distinguish between the different morphological and typological features and to navigate the user easily to the most appropriate.

A meta – model is the schematics used by an application to understand a metadata expression given the nature of terms and how they combine to form a metadata de-

scription, thus making it possible for a single standard, though expressed in several different formats, to still be understood in a uniform way by users and applications. Due to the abstraction of the model, it is possible to generate different representations of metadata schemas underlying this model. ArMOS metadata schema presented bellow is a valid instance of this model.

The ArMOS meta - model incorporates basic architecture theories for concepts such as morphology, typology, etc. In order to understand these sufficiently, we worked closely with the architecture community. To the best of our knowledge, there are so far no relevant modeling approaches for such upper - level architecture concepts proposed in literature. The model takes the form of a lightweight entity-relationship conveying the following semantics:

- *Architectural Composition:* constitutes the core concept (illustrated as an abstract entity) of the overall model. A higher level of abstraction - the conceptual content that underlies all the different building types or architectural design ideas. Architectural composition is the beginning of the existence of a structure (e.g. of a building) and is realized through morphology.
- *Morphology:* is also an abstract entity that encompasses all the features of the design and construction of a building. Morphology is the architectural form as support, and also its complements. It refers to features such as the relation between openings and solids, the expression of materials such as texture and own color and the functional applied ornamentation. Morphology does not refer so much to the decoration of the building (this is style or the rhythm) as to the elements of its general composition such as facades, plans, walls, windows, balconies etc. Morphology derives a new typology or derived from an existing applied typology. Morphology is composed of various morphological patterns embodied in a building.

- *Patterns:* in architecture, it is the capturing of architectural design ideas as archetypal and reusable descriptions (visual or textual) grouping objects by certain inherent structural similarities. Building that surround us embody such patterns.

- *Typology:* In typological science, the term typology can be understood as a term purely used to classify individuals within a group. In the field of architectural design, typology is considered as a rigorous method for analysis, organization, and classification of a variety of buildings into representative classes (Lawrence, 1994; Schneekloth & Franck, 1994). On the context of ArMOS the typology of a building is divided into three levels: The class level (e.g. educational building), the type level (e.g. high school) and the group level (e.g. two storeys L shaped). Typology is a comparative classification of dominant architectural solutions with objective and rational criteria. Typology also could be depicted in patterns or create new.

- *Exemplar(s):* is a representative example of a structure (e.g. an existing building), in which the various morphological or typological features have been embodied. Exemplars are associated with the core concept of the model – architectural com-

position - with various extrinsic relations, discussed in the previous section (such as: *similar*, *influenced*, *variation, preceding, succeeding*) allowing us to understand continuities, similarities, and variations among buildings.

- The entity defined as *place* in the model encompasses a comprehensive range of locations associated with an architectural work: terrestrial, historical and contemporary, geographic features and geo-political jurisdictions.
- *Period:* The entity encompasses single dates, historical periods or a range of dates, for a wide range of activities associated with the creation of an architectural work.
- *Agent*: encompasses individuals, group of individuals or organizations that are treated as entities only to the extent that they are involved in the creation or realization of an architectural work (e.g., as architect, engineer, artists, decorator, construction company e.t.c.).

The following figure (Fig.1) depicts a model for architectural composition that has maximal expressivity in the architectural knowledge domain using a minimal set of the above core concepts.

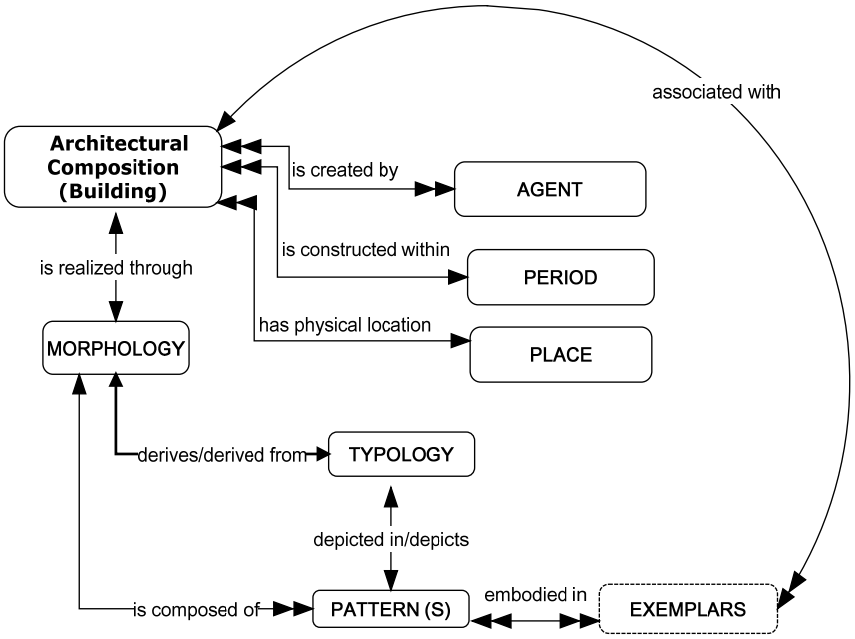


Fig. 1. The ArMos Meta - Model

4 Structuring ArMOS Metadata Schema

In our effort to formulate an harmonization profile for historic buildings that can map existing records in other schemas, with minimal loss of information, we started to study the metadata elements hosted in various heterogeneous schemas by monument inventories, used to describe different types of immovable monuments (Agathos-Kapidakis, 2011). Moreover in a later work, a series of crosswalks was defined among a derived schema from the above survey, and mature official metadata standards that are used to describe architectural works (Agathos, Kapidakis, 2012). The study allowed us to identify which mature metadata standards are useful in whole or part, for activities related to describing and managing architectural works. The results of these studies, gave us a new perspective for the method that these records must be compiled, from the aspect of metadata elements, in order to have semantics descriptions.

Many of the schemas studied from local and national monument inventories[1] accommodate similar descriptive types of metadata used for purposes of description, discovery and identification of historic buildings, including elements such as the name of building, the architect, the type of the building, its function etc. Moreover, each of the above schemas incorporates a different subset of administrative and structural metadata. According to institutional purposes, administrative metadata provide information to local authorities to help manage these structures for purposes of protection, restoration, conservation or planning, while structural metadata decompose the building into its various parts.

Given the diversity of the above metadata types, extensibility was a critical feature that was taken into account in the creation process of ArMOS. The schema that was created provides further possibilities for extensibility through the use of what is called extended data elements: elements of a data structure that is defined outside a standard and is permitted within an instance of the data structure providing a mechanism for extensions. ArMOS provides a high harmonization level for descriptive metadata for immovable monuments, while allows extension for other types of metadata (Table 2.)

Table 2. ArMOS Harmonization coverage on different types of metadata

Metadata Type	*Harmonization Level*
Descriptive	High
Structural	Medium – allows extensions
Administrative	Medium- allows extensions

Application profiling of metadata specifications in its simplest form supports the process of selection of a set of metadata elements from an element vocabulary, possibly extending the base element vocabulary as defined in the specification using lo-

[1] For the purpose of this research, we study the serialization of data of 11 national monument inventories recorded by UNESCO (sykes, 1984).

cally defined elements, and choosing a set of useful value vocabularies for use with these elements (Nilsson, 2008).

The above circumstances and the fact that most of the examined mature metadata standards, have been designed for general collection description of material of our culture (except *CDI*, which provides core information for historic buildings) resulted in a profile which is created by extending Core Data Index to Historic Buildings and Monuments and by taking elements from various mature metadata specifications, especially from CDWA (Categories of Description of Works of Arts). Moreover, in order to eliminate the absence of important elements for the description of these resources, ArMOS is supplemented by 'new' elements, locally defined, for which a namespace was created[2].

Data elements of the schema are grouped into 16 categories (Fig.2). The schema presented as a hierarchy, including aggregate data elements and simple data elements. In contrast to the property-value structure used by Dublin Core, ArMOS uses a hierarchical structure of elements-within-elements similarly to the LOM abstract model, so we could say that ArMOS belongs to the IEEE LOM family of specifications (Nilsson, 2008). Each element can be either a container element, thus containing other elements, or a leaf element, which holds a value of a certain data type. The top-level elements are called categories. ArMos harmonisation profile provides not only a set of data elements, but also a default pattern for the use of those data elements, a "base" application profile to which other community- or application-specific ArMOS application profiles should also conform.

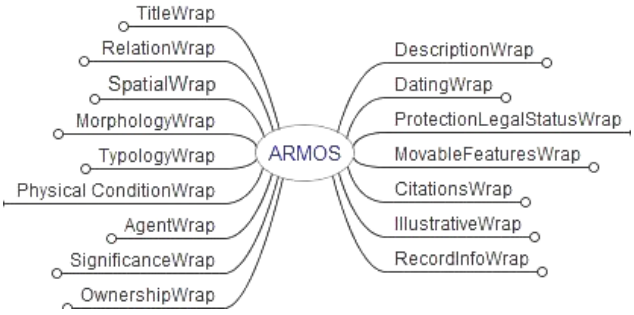


Fig. 2. Categories of ArMOS Profile

[2] Namespace: armos- `www.dlib.ionio.gr/standards/armos`

5 Overview of the Proposed Elements

All metadata terms of the proposed schema identified with URIs. At this sense a normalized documentation will be prepared[3] in which the above elements will identified as precisely as possible (*Principle of Appropriate Identification* such in the case of DCAPs), including enough description in order to be of optimal usefulness for the intended audience of the schema (*Principle of Readability*). Moreover the schema specifies a small set of vocabularies that should be used to provide values for the various metadata elements. An overview of the proposed elements of the two main categories *Morphology* and *Typology* to be included in the profile is provided in Table 3. The table also includes brief information about the cardinality, and if an element is mandatory or not.

Table 3. The proposed elements of ArMOS for Categories Morphology and Typology.

Proposed Elements	*Namespace*	*Requirement*	*Cardinality*
Morphology Wrap	armos	M	N-R
Architectural Composition	armos	O[4]	N-R[5]
Style-Period	cdwa	O	R
Measurements Set	armos	O	N-R
Dimensions Description	cdwa	O	R
Dimensions Extent	cdwa	O	R
Dimensions Type	cdwa	O	R
Dimensions Value	cdwa	O	R
Dimensions Unit	cdwa	O	R
Dimensions Qualifier	cdwa	O	R
Shape	Cdwa	O	R
Buildings Materials & Techniques Set	armos	M	N-R
Main Materials StructuralTechniques	cdi[6]	M	R
Covering Material	cdi	M	R
Façade Material	armos	O	R
Color Facade	armos	O	R
Extent	armos	O	R
Technique	armos	O	R
Decorative Details Set	armos	O	N-R
Internal Decoration	armos	O	R
External Decoration	armos	O	R

[3] The details of each element and guidelines for adding content will be available from `http://dlib.ionio.gr/standards/armos`

[4] Optional (O) / Mandatory (M)

[5] Repeatable (R) / Non – Repeatable (N/R)

[6] Core Data Index to Historic Buildings Core Data Index to Historic Buildings and Monuments of the Architectural Heritage

Architectural Art Sculpture	armos	O	R
Decoration Technique	armos	O	R
Immovable Features Set	armos	O	O
Inscription Transcription or Description	cdwa	O	N-R
tablets	armos	O	R
coats of arms	armos	O	R
Murals	armos	O	R
Construction Element Set	armos	O	R
Windows type	armos	O	R
Roof type	armos	O	R
Door features	armos	O	R
Stairways type	armos	O	R
Dome	armos	O	R
Typology Wrap	armos	M	R
Building category class	armos	M	N-R
Building category type	armos	M	R
Typological group	armos	M	R
Current use	armos	M	N-R
Future proposed use	armos	O	R

In ArMOS, the global element "pattern" is the linchpin that links the described architectural work with one or more other architectural works (Fig.3). *Pattern* element can be applied to any data element hosted in the categories of morphology and typology. A *Pattern* may have a visual representation - an image – or/and a textual description. The "relation type" element of ArMOS describes the type of relationship between the two works. The list of the recommended relation types discussed above (Table 1) describes a multitude of relationships. In the example below (Fig.3) Booker T. Washington Highschool is associated with Trenton Highschool in Michigan due to their common typological group (two storeys L shaped).

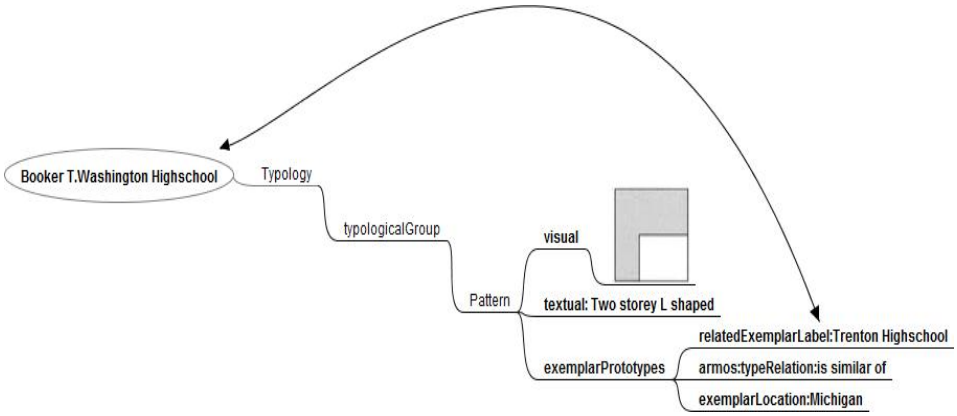


Fig. 3. Connecting two individual architectural works through the global element *pattern*

6 Conclusion

With the functionality that the information technology offers for information sharing, the benefits of creating cultural heritage information networks are clear. These include the enabling of common access to inventories created and managed by diverse organizations and memory institutions (Bold, 2009, p.12). However, metadata schemas relevant to historic buildings have not yet achieved until now the milestone of formal standardization. ArMOS schema that is consistent with the semantic web recommendations will help to achieve greater interoperability on the metadata practices and descriptions for these types of resources, supporting Semantic Web services. Our next steps include testing the application profile with concrete examples on historic building records from various monument inventories in order to evaluate the overall effectiveness, identifying areas in the schema requiring attention and revisions in order to meet the specific and concrete needs of these stakeholders.

Bibliography

1. Agathos, M., and Kapidakis, S. Discovering Current Practices for Records of Historic Buildings and Mapping them to Standards. *Paper presented at the First Workshop on Digital Information Management.* 30-31 March 2011. Corfu, Greece, pp. 61-75. Available at: `http://hdl.handle.net/10760/15847` (Accessed 10 May 2012).
2. Agathos, M., & Kapidakis, SDescribing Immovable Monuments with Metadata Standards, International Journal on Metadata Standards, Semantics and Ontologies, (2012) (under publication – Acceptance date 20 Apr.2012)
3. Baca, M. et al. Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images, American Library Association, Chicago (2006).
4. Bold, J. (ed). Guidance on inventory and documentation of the cultural heritage, Council of Europe Pub., Strasbourg (2009.
5. Casanova et. Al. Evolutionary Processes, Morphology and Typology of Historical Architecture as a Line of Research: a Tool for Conservation. *Paper presented at 16th International Conference on Urban Planning.* Development and Information Society, 18-20 May 2011. Essen, Germany, pp. 285-292 (2011). Available at: `http://www.corp.at/archive/CORP2011_209.pdf` (Accessed 10 March 2012).
6. Diamantopoulos, N., et al. Developing a Metadata Application Profile for Sharing Agricultural Scientific and Scholarly Research Resources Metadata and Semantic Research, Springer Berlin Heidelberg (2011). 240: 453-466.
7. Flanagan, G. Conceptual requirement Validation for architecture design systems.. Thesis, California Polytechnic State University (2011).
8. Haslhofer, B., Klas, W. A survey of techniques for achieving metadata interoperability, *ACM Comput. Surv.*, vol. 42, no. 2, (2010). Available at: `http://www.cs.univie.ac.at/upload/550/papers/haslhofer08_acmSur_final.pdf` (Accessed 10 March 2012).

9. Nilsson, M., Baker, T., Johnston, P.: The Singapore Framework for Dublin Core Application Profiles (2008), http://dublincore.org/documents/singapore-framework/
10. Nilsson, M. Harmonization of Metadata Standards. Deliverable of the PROLEARN IST-507310 European Project, (2008) Available at: http://ariadne.cs.kuleuven.be/lomi/images/5/52/D4.7-prolearn.pdf (Accessed 14 March 2012).
11. Sykes, M. H. Manual on Systems of Inventorying Immovable Cultural Property, Unesco, Paris (1984).

Automatic Medical Document Generation via Spatial SNOMED Elements in Hysteroscopy

Anastasios Kollias[1]

Ionian University, Department of Archives and Library Science
Laboratory of Information Technology
Ioanni Theotoki 72, 49100, Corfu

tkodka@gmail.com

Abstract: In many medical examinations, specifically in endoscopic surgeries, the doctor's observation plays an important role in the final diagnosis. However, some of doctors use manuscripts in order to depict the medical endoscopic findings .Those manuscripts basically map the observations on Case sketches. The Hysteroscopy procedure follows this practice of degrees of information which nonetheless creates heterogenetic problems in the process of recording medical, digital information. The aim of this study is to solve the arising problems by using an automatic recording of Hysteroscopic findings. This can be achieved by combining the image processing technique with the SNOMED and CDA medical standards.

Keywords: SNOMED, Metadata, Image Processing, Hysteroscopy.

1 Introduction

It is obvious, by the problems arising in the medical sector due to the increasing role of modern technology in our everyday lives, that modern medical science needs to fully develop and exploit all forms of medical information originating from existing medical incidents [2] .In many medical examinations, specifically in endoscopic surgery, the doctor's observation plays an important role in the final diagnosis. However, these observations in the case of diagnostic endoscopic hysteroscopy and similarly in other medical examinations are not established as an objective diagnostic element [7]. Thus, many doctors use manuscripts in order to depict the medical endoscopic findings using mapping observation on Case sketch [6][8][1].This technique creates many problems which arise from the heterogeneity of practices which respectively leads in interoperable problems. Yet, the semantic organizing of documents, which is based on the extended markup language (XML), gives many solutions to the above mentioned problem [9]. The main interface between different medical computer systems is served by the international medical standards which were formulated after years of studying and processing medical terminology. SNOMED, the international

standard for medical terminology, is a complete, encoded system of medical information that is becoming the most recognized tool for exploiting medical knowledge. SNOMED achieves interoperability between different types of medical information systems and also defines rules and processes in order to solve the problem of interoperability between computer systems based on different principles, as its content is not purely medical [3]. Furthermore, the Clinical Document Architecture (CDA) model proposed by the international medical service organization HL7, realizes the correspondence between the different medical units and systems. Thus, using these properties it is possible that the conversion between SNOMED and CDA will be achieved and will lead to a generation of clinical documentation [10] via medical examinations such as Hysteroscopy produces.

This study focuses on an Automatic Document Production (Summary) CDA by conventional recording (sketch of uterus) which locates the hysteroscopy findings, using encoded information based on the standard SNOMED CT. The aim of the study includes the international practice of gynecology objective and is designated by the respective office of hysteroscopy in the gynecological clinic of the Medical School of the University of Ioannina.

2. Methodology

2.1 Pre-Processing Stage

The Pre-processing stage is divided in the Graphical User Interface and the SNOMED Analysis.

2.1.1 Graphical User Interface

In accordance with the Introduction Part we created a graphical user interface for image editing [5] where the image is a pre-drawn sketch of a typical uterus. In more details, the user (doctor) can select regions of the original sketch and characterize them, depending on the findings, by using an included toolbox. The implementation of this took place using open source image editing components. The interface of this is depicted in Figure 1.

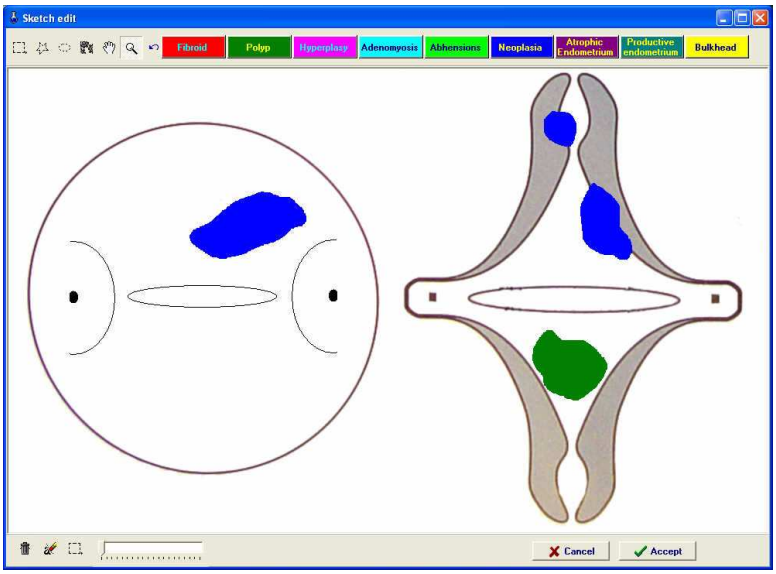


Fig. 1. The graphical user interface for uterus sketch editing

2.1.2 Analyzing SNOMED

An SNOMED analysis is performed in order to investigate the "spatial" division of the uterus SNOMED supports and the selected terms of this division to be used in the application. It should be noted that many of the terms used in it correspond to overlapping regions of the uterus. This required the selection of specific terms of the standard, and not all, so that no problems arise during the processing. Also, the colors used in the toolbox to define the kind of findings are matched with the corresponding terms of the findings according to the medical standards. After this, a standard sketch is created , in which each point was characterized by the area to which it belongs as above.(Figure 2).The enriched sketch in the background , is used as the pre-drawn image on the image editing by the doctor.

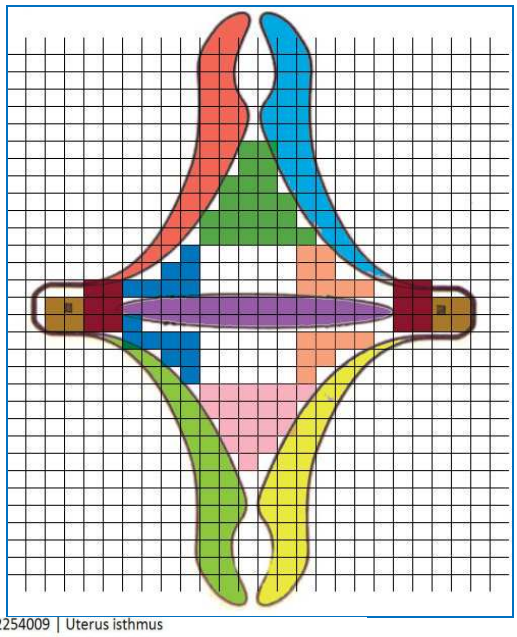


Fig. 2. The adaption of SNOMED in the graphical user interface panel

Furthermore, a connection between a medical application of hysteroscopy procedure management (parent application) [4] and the proposed method is attempted. Thus, some very important data such as the patient's age, the menstruation details, births number and the cause of the examination (see figure 5) are obtained in SNOMED codes form, by parent application.

2.2 Processing Stage

This stage is divided into six (6) following steps:

1. Image Analysis
2. Data grouping and clearing
3. Metadata editing
4. Application management

5. Searching
6. CDA Generation

2.2.1 Image Analysis

At this step a data structure is defined that contains two fields which describe the pixels. Thus, in this mapping two (2), a particular SNOMED code corresponds in the color code which is the doctor's remark type of finding and the position which describes the spatial pixel's position.

.

```
Pixel_Property=Record
                finding_code   : longint;
                area_code      : longint;
                end;
```

After that, a linear analysis of every pixel of the sketch registered by the doctor is performed. This procedure is implemented via the following logical steps which are depicted in the Figure 3.

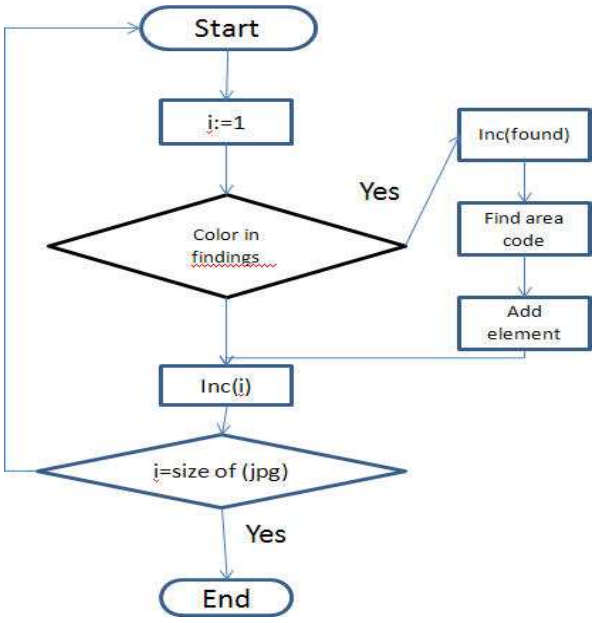


Fig. 3. Linear image analysis procedure

2.2.2 Data grouping and clearing

The array created at the first step, is sorted by the type of the area and then the data grouped by the type of findings in order to create totals. A statistical analysis of the sorted and grouped data is performed in order to clear the extreme values of them and to reject the obviously incorrect values. In more details, this procedure is depicted in figure 4

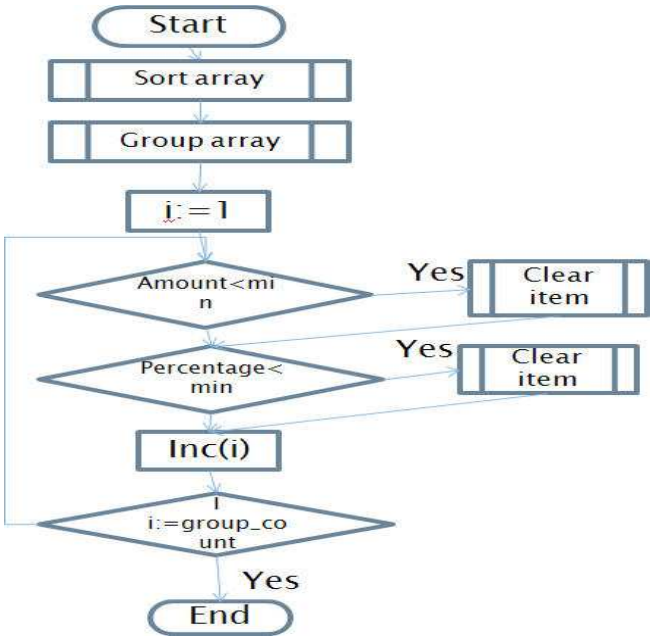


Fig 4. Clearing procedure

2.2.3. Metadata editing

The result of the above procedure return one or more records (at this time limit to three) with the values of three (3) variables (the area code of uterus, the type of finding and presentence of the overlapping area).

```
Image_Area      =….
Image_Find      =….
Presentence     =….
```

The final results (SNOMED codes and quantifiable), are submitted inside the JPG image in metadata form by using open source components together with the personal data which are gathered from the parent application. The form of these metadata is shown in details in figure 5

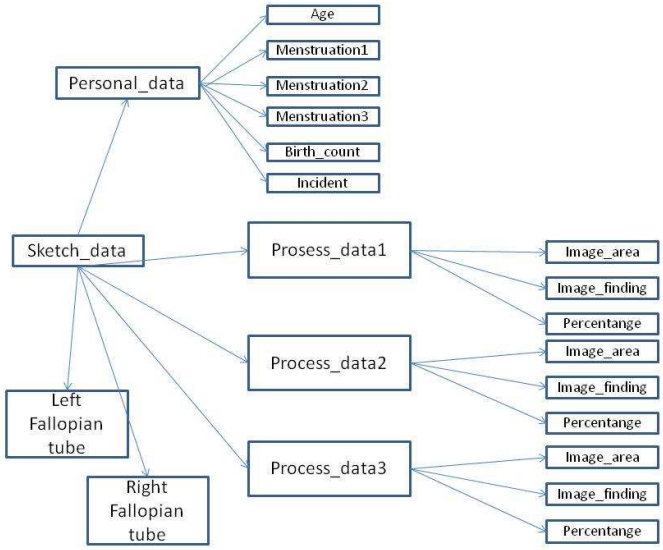


Fig. 4. Metadata structure

2.2.4 Application management

A library management application of enriched sketches is designed and implemented, to support the created database. In the process of sketches' registration in the database, the stored metadata record in tables of the application. Simultaneously, the application creates index files using specific keywords for quick searching without the need of extensive image analysis. The application's Entity Relationship Diagram (ERD) is shown in Figure

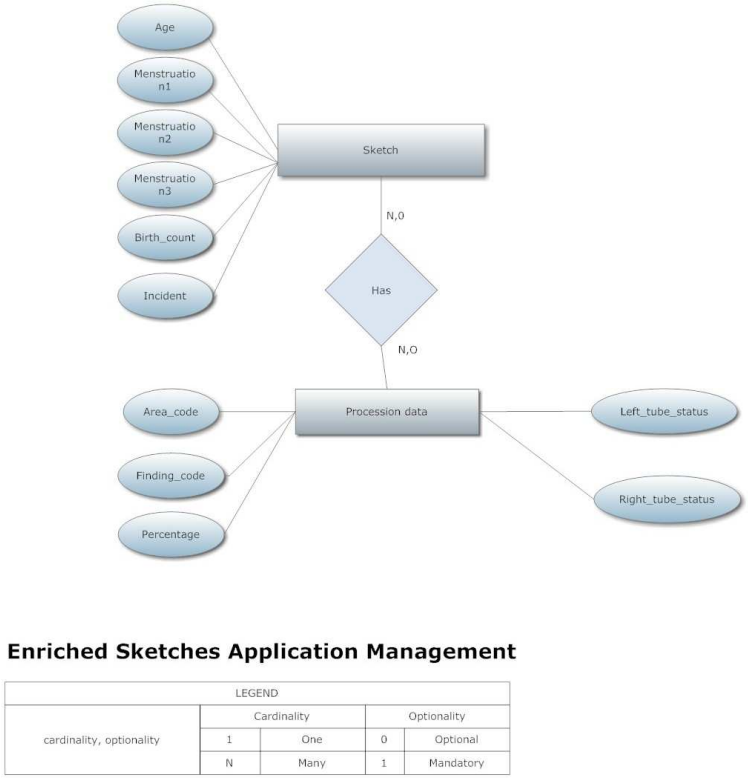


Fig. 6. The ERD application

2.2.5 Searching

A convenient searching user interface is created for quick search results in the library of sketches, based on objective such as key codes SNOMED.(Figure 7)

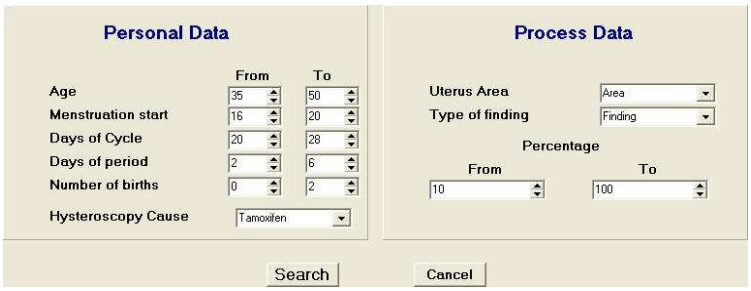


Fig. 7 Searching dialog

2.2.6 CDA Generation

The possibility of producing a CDA document from the metadata incorporated in each sketch via the SNOMED was added to the application. This option offers the potential output data in a format that supports the interoperability of computer systems. In more details, we create a mapping between the CDA data and the specific terminology and coding systems of SNOMED (Yuwen, S. & Yang, X, 2010) which fits the Hysteroscopy practice. Based on this idea, the system framework in accordance with gynecologists of the medical Scholl of the University of Ioannina is depicted in Figure 8.

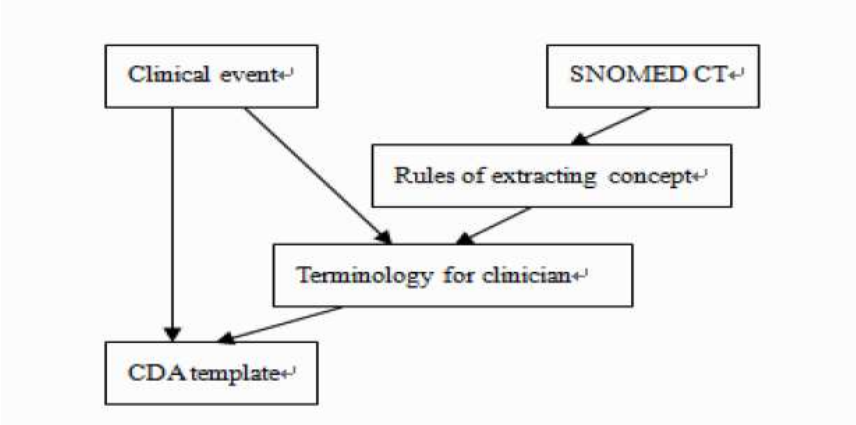


Fig 8. System Framework

3. Case Study

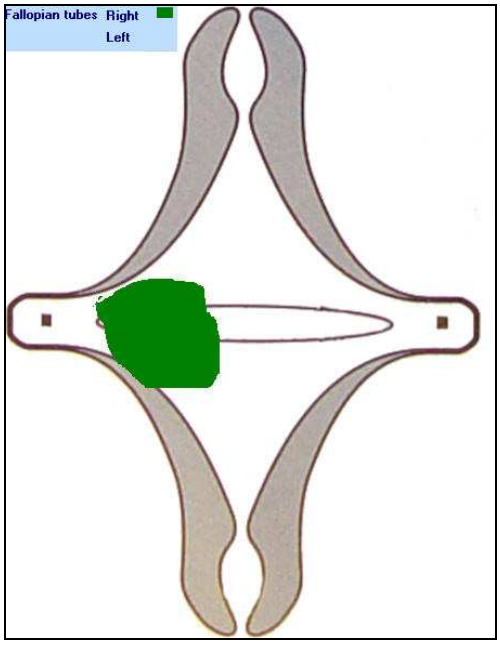
In this case, we implement the proposed method using a medical incident which is analyzed as follows:

3.2 Pre-processing stage

The patient who will become the hysteroscopy give the following data to the doctor and he entered them in the parent application

a) Age: 44 years
b) Menstruation data
    1. Cycle 27 days
    2. Period 4 days
    3. Start 13 years old.
c) Number of term births: 2
d) Indication of procedure : Use of tamoxifen for many years

After gynecological examination of the patient, the patient underwent in hysteroscopy procedure where produced the above sketch:



3.3 Image Analysis

The linear analysis of the image file gives the following results

a) Big number of pixels in the area of the left side of the image (corresponding to the place of the bottom and left horn of uterus) have green color
b) The pixels corresponding to the value of the right fallopian tube stage are all green
c) The pixels corresponding to the value of the left fallopian tube stage has no color
d) Small amount of pixels corresponding to the left wall of uterus have green color.

3.4 Data grouping and clearing

The automatic grouping and measuring of the data give the following results.

a) 87.2 % of uterus left horn have green color

b) 44 % of bottom of uterus have green color
c) 7 % of left uterus wall have green color
d) The third part of the results corresponding to the uterus wall was too small so was considered to be in error drawing by the doctor so cleared by the application

3.5 Metadata editing

The SNOMED Concept data for the field names is depicted in Table 1:

Table 1. SNOMED Concept data

Personal Data				
Description Data	**SNOMED CONCEPT CODE**		**Value**	**SNOMED VALUE CODE**
Age	397669002		44	9362000 9362000
Menstruation cycle	248960000		27	19338005 65607009
Menstruation Period	248959005		4	93602000
Years of menstruation life	310467003		31	79605009 38112003
Number of births at term	440425002		2	19338005
Indication	230165009		Tamoxifen adverse reaction	41395001
Procedure Data				
Finding site	363698007		Bottom of uterus	263354004
Percentage of total	258755000		87.20%	278497006 (75-90%)
Finding with explicit context	413350009		Uterine fibroid polypus	254880000
Finding site	363698007		Left horn of uterus	245482004
Percentage of total	258755000		44%	278495003 (25-50 %)

After Finding with explicit context	413350009		Uterine fibroid polypus	254880000
After Left fallopian Tube	417621010		Open	93326018
After Right fallopian tube	417620011		Closed	48829017

After the description data of table 1 are mapped in the metadata keys according to SNOMED. More details are presented in Table 2.

Table 2. Case study Metadata

Personal Data	
AGE	397669002;9362000 9362000
MENSTRUATION1	248960000;19338005 65607009
MENSTRUATION2	248959005;93602000
MENSTRUATION3	310467003;79605009 38112003
BIRTH_COUNT	440425002;19338005
INDICATION	230165009; 41395001
Procedure Data	
IMAGE_FINDING1	413350009;254880000
PERCENTAGE1	258755000;278497006
IMAGE_AREA1	363698007;263354004
IMAGE_FINDING2	413350009;254880000
PERCENTAGE2	258755000;278495003
IMAGE_AREA2	363698007; 245482004
LEFT_FALL_TUBE	417621010; 93326018
RIGHT_FALL_TUBE	417620011; 48829017

3.6 CDA Generation

In this stage, the production of the CDA document is produced via the aforementioned case study of incident. In more details, we described a session of this case according to architecture which depicted in figure 8. Also, this example of the CDA coding is presented as follows:

```
<targetSiteCode code="413350009"
           codeSystem="2.16.840.1.113883.6.96"
         codeSystemName="SNOMED CT"
```

```
          displayName=" finding with explicit context>
          <description>
                <name code="246090004"
                codeSystem="2.16.840.1.113883.6.96"
                displayName=" associated finding "/>
                  <value code="254880000"
                    codeSystem="2.16.840.1.113883.6.96"
                displayName=" uterine fibroid polyp "/>
                 <qualifier>
                      <name code="418775008"
                       codeSys-
tem="2.16.840.1.113883.6.96"
                       displayName=" finding method "/>
                      <value code="88355005"
                      codeSystem="2.16.840.1.113883.6.96"
                      displayName=" diagnostic hysteros-
copy "/>
                </qualifier>
<qualifier>
                      <name code="246075003"
                       codeSys-
tem="2.16.840.1.113883.6.96"
                       displayName=" causative agent "/>
                      <value code="41395001"
                      codeSystem="2.16.840.1.113883.6.96"
                      displayName=" tamoxifen citrate "/>
                </qualifier>
                 <qualifier>
                      <name code="363698007"
                       codeSys-
tem="2.16.840.1.113883.6.96"
                       displayName=" finding site"/>
                      <value code="245482004"
                      codeSystem="2.16.840.1.113883.6.96"
                      displayName=" entire left horn of
uterus "/>
                </qualifier>
             </description>
 </targetSiteCode>
```

4 Conclusions

- The application of proposed method in the Hysteroscopy's examination shows that this procedure is simplified for the following reasons:
- The hysteroscopy procedure produces easier, more correct and efficient registration of medical encoded information. This is achieved by automated production of the codes of medical model that eliminates the possibility of random error generation and standardize the terms used, bypassing the problems of overlapping terms that exist within the standard.
- The doctor is exempted from the need for knowledge management, since applications use specialized medical standards. This feature makes easier to harmonize a gynecological clinic with the modern requirements of medical information production since it does not require extra effort by the doctor.
- An easily manageable library of medical information is created for use by appropriate researchers. The big advantage of it is that simplifies and normalizes the data of each incident using only a sketch. In this way the researcher has the possibility of not only the statistical research of incident's data, but the immediate and very rapid visual presentation.
- An easy interconnection of different applications is enabled using the data export possibility in form of CDA document
- In the future, this method could be used in order to full utilization of the latest computing devices (tablets). Furthermore, this method could be applied to encode other simple medical examinations as well. Finally the creation of «network» image libraries enriched sketches, editable via WEB could be made possible.

References

1. Belmartino, S. The role of the state in health systems. Social Science & Medicine 39, 1315–1321 (1994).
2. Gelijns, A. C. & Rosenberg, N. From the scalpel to the scope: Endoscopic innovations in gastroenterology, gynecology, and surgery. Medical Innovation at the Crossroads 5, 67–96 (1995).
3. Neis, K. J., Brandner, P. & Hepp, H. Hysteroscopy: textbook and atlas. (G. Thieme Verlag: 1994).
4. Paschopoulos M, Paraskevaidis E, Stefanidis K, Kofinas G, Lolis D. Vaginoscopic approach to outpatient hysteroscopy. J Am Assoc Gyn Lap 1997; 4: 465–467.
5. Baggish, M. S. Colposcopy of the cervix, vagina, and vulva: a comprehensive textbook. (Mosby Inc: 2003).
6. Rezayat, M. Knowledge-based product development using XML and KCs. Computer-Aided Design 32, 299–309 (2000).
7. Benson, T. Principles of health interoperability HL7 and SNOMED. (Springer Verlag: 2010)
8. Yuwen, S. & Yang, X. Conversion between SNOMED and CDA. Biomedical Engineering and Computer Science (ICBECS), 2010 International Conference on 1–4 (2010).

9. Lagueux Jr, R. A. et al. Graphical user interface for configuration of a storage system. (2003).
10. Kollias, A., Paschopoulos, M., Evangelou, A. and Poulos M., Digital Management of a Hysterosccopy Surgery Using Parts of the SNOMED Medical Model. Open Medical Informatics Journal (6) 15-25, (2012)

Query Expansion and Context: Thoughts on Language, Meaning and Knowledge Organisation*

Anna Mastora[1] and Sarantos Kapidakis[1]

[1]Laboratory on Digital Libraries and Electronic Publishing, Department of Archives and Library Science, Ionian University, Corfu, Greece
72, Ioannou Theotoki str., GR-49100
{mastora, sarantos}@ionio.gr

Abstract. This study revisits the query – document terms mismatch problem and discusses identified perceptions on language, meaning and knowledge organisation, mainly of Ludwig Wittgenstein, in view of query expansion. We conducted literature review on query expansion techniques, knowledge organisation in general, as well as systems and Wittgenstein's theories on respective issues. The analysis commences from the discussion of the query – document terms mismatch problem and proceeds to the allocation of theories on the aforementioned issues. Additionally, we mention query expansion techniques in view of the definition of what constitutes the context of the query. We identify a theoretical background concerning the context which current query expansion techniques try to define and further exploit to the benefit of the searcher.

Keywords: Query expansion, Knowledge organisation, Meaning, Language, Context

1 Introduction

Information retrieval is about retrieving relevant results as response to information needs expressed through queries or by using any other available technique offered by current information retrieval systems, for example browsing. The great challenge for the community being involved in the creation and development of information retrieval systems is to define and decide what is relevant for whom at each occasion.

To accomplish that, a number of conditions have to be fulfilled depending on numerous factors, like the users, the interface, the available collections of documents and their organisation to name a few. This is why the research on Information Retrieval has so many different approaches and engages researchers from various disciplines. This is also why it is an intensive and complicated process although simple to comprehend.

Why does it end up being such a great challenge when it can be broken down to simple words? The user submits the query terms (mostly using natural

* 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.

language representations) to a system; the system has to locate these terms in the stored documents and deliver them to the user. The problem is that it is not only about *locating*, which can, too, be tricky, but also about *matching*, since query terms and document terms are not always expressed by the same linguistic representation. This situation is what has been established as the *query – document terms mismatch problem*.

Our study revisits the query – document terms mismatch problem and discusses current efforts for dealing with it, namely query expansion techniques. These techniques are juxtaposed here vis-à-vis with identified perceptions on language, meaning and knowledge organisation mainly of Ludwig Wittgenstein and others as well. His relation to Information Systems in general and to issues related to Artificial Intelligence and the Semantic web in particular has been demonstrated in several studies (Seidel: 1991; Blair: 2006; Halpin: 2009; Milonas: 2011).

Wittgenstein dived into the problem of assigning or extracting the meaning of words concluding that this can be achieved only through the uses of words as realised in human activities and practices -the "language games"- and as perceived within a certain context -the "forms of life". We indicate that Wittgenstein's postulates on these matters serve as the theoretical ground for implementations on current, robust query expansion techniques. Query expansion efforts strive to define, manipulate and deliver this context within which the words extract their meaning aiming at "how to obtain the right information for the right user at the right time" (Chu: 2003).

2 The Research Hypothesis

Deriving information about the context within which words are used will help us better clarify the identified indeterminacies caused by the use of natural language in the process of Information Retrieval. We believe that by studying the actual use of language in every-day activities, which in our research involves the users' search behaviour patterns within certain context, we can clarify the intended meaning in a more comprehensive way.

3 The Query – Document Terms Mismatch Problem

Within current Information Retrieval practices, the need is to find documents containing the information we are looking for. The user having an information need creates a conceptualisation of it; then, she expresses it using a linguistic representation. If a system treats queries as bags-of-words and, thus, performs only string matching, i.e. matching the words used in the query with the words contained in the documents only at a lexical level, then the user might come across with one of the following scenarios.

First, a document may contain the exact same representation and bear the meaning which the user intended. In that case, the document is retrieved and the user stands more chances for being satisfied with the outcome. Second, a document may contain the same representation but mean something different than the user intended, which means that the document is retrieved but the user stands more chances of being unsatisfied with the result, or rather puzzled. Finally, a document may contain a different linguistic representation but mean what the user intended. This means that the document is not retrieved and the user is left with the impression that no relative-to-her-request documents are contained within the collection. Figure 1 below shows a graphical representation of the aforementioned situation.

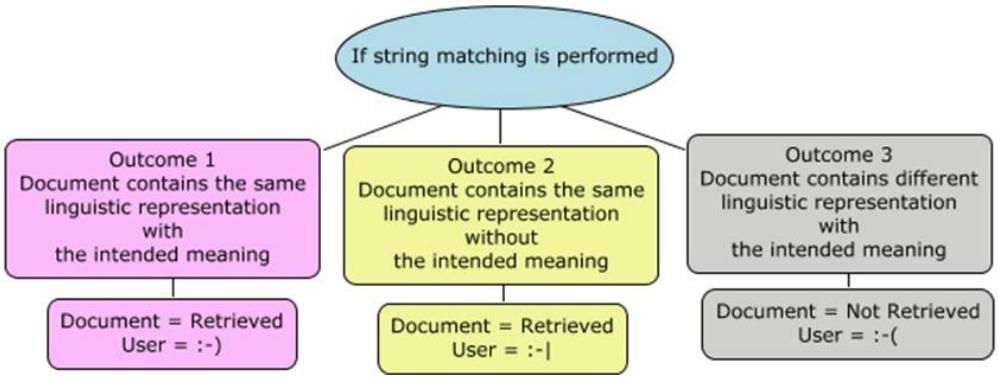


Fig. 1. A visualisation of the query – document terms mismatch problem and its consequences for the searcher.

To make things even more complicated, rather than more simple as we might expected, at the aforementioned situation, we have to also consider another significant factor which plays, or at least it was meant to play, the role of the mediator: the document's metadata, and more specifically either the linguistic or the systemic classification of the indexer. The indexer was meant to be some kind of a problem solver; a mediator between the inquirer and the content which is stored in the documents.

However, it is established that each person identifies reality in different ways and the indexer -let us accept him being any creator of any sort of index- is not the one holding the one and true meaning of the word. Blair (2006) claims that the expert does not hold a "more complete" definition of the words than we do. He simply knows more about certain words than we do, and by providing this additional information about them, it may be useful for identifying each word in different circumstances. It is not a prerequisite, though, that the definitions the indexer holds would match any of the definitions the searcher has in mind during the information searching process.

4 About Language and Meaning

As quoted in (Blair: 2006) Wittgenstein stated in "Zettel", §173, "Only in the stream of thought and life do words have meaning". Words, if considered individually, may not carry unique and distinctive meanings, but, if related to certain uses, they could form distinguishing entities. In addition to that, according to Wittgenstein it is the examples of the uses that are more likely to deliver the meaning; this is why we should ask about the use and not the meaning of a word.

Wittgenstein, in "Tractatus Logico-Philosophicus" (1922) states that "Colloquial language is a part of the human organism and is not less complicated than it" while in §23 of his Philosophical Investigations (1967) he questions:

"But how many kinds of sentence are there? Say assertion, question, and command?—There are countless kinds: countless different kinds of use of what we call "symbols", "words", "sentences". And this multiplicity is not something fixed, given once for all; but new types of language, new language-games, as we may say, come into existence, and others become obsolete and get forgotten."

And in the same work, being more articulate, he uses a metaphor: "Language is a labyrinth of paths. You approach from one side and know your way about; you approach the same place from another side and no longer know your way about" (§203).

Foris (2010), states that linguistic studies of word frequency (by Giraud: 1954) have pointed out that word frequency is connected to the semantic properties of the word: the rarer the use of the word is, the smaller its frequency and the probability of its use is, the more defined its meaning is, and the higher its information value is. So, frequently used words tend to have more meanings than rarely used words which are more defined and with high informational value because they stand for distinctive concepts. But the problem is not with rare words; instead it is the more common words that have to be submitted to the disambiguation process.

We conclude this section with a quote from Szostak (2010) saying that "language is clearly ambiguous", so not all linguistic representations are distinguishably informative of the user's query intent unless certain context is identified.

5 Query Expansion

In order for Information Systems to deal with the problems caused by the diversity of linguistic representations, Query Expansion is implemented. Query expansion is a popular technique in information retrieval for modifying users' queries in order to perform better in terms of precision and recall. It is defined as the stage of the information retrieval process during which a user's initial query

statement is enhanced by adding search terms to improve retrieval performance (Shiri and Revie: 2006). Basically, it is about supplementing the original query with additional, meaningful words or phrases (manually, automatically, semi-automatically). The key-term in the previous sentence is *meaningful*. In order to provide more meaningful terms related to the initial query, we first have to identify and define the context within which each term derives a certain meaning.

In the *Blue and Brown Books* (1958) Wittgenstein is alleged to state that "We are unable clearly to circumscribe the concepts we use, not because we don't know their real definition, but because there is no real "definition" to them". And he famously argued that the best way to define a concept was to provide examples of it. This position is stated in "Philosophical Investigations" as: "We must do away with all explanation, and description alone must take its place. […] The problems are solved, not by giving new information, but by arranging what we have always known" (§ 109).

6 About Context

Consequently, the next question is about what constitutes the *context*. Briefly, it is about deriving more information about the involved parties within the information retrieval process. Figure 2 below presents the multi-faceted concept of context as discussed in Bhatia & Kumar (2010).

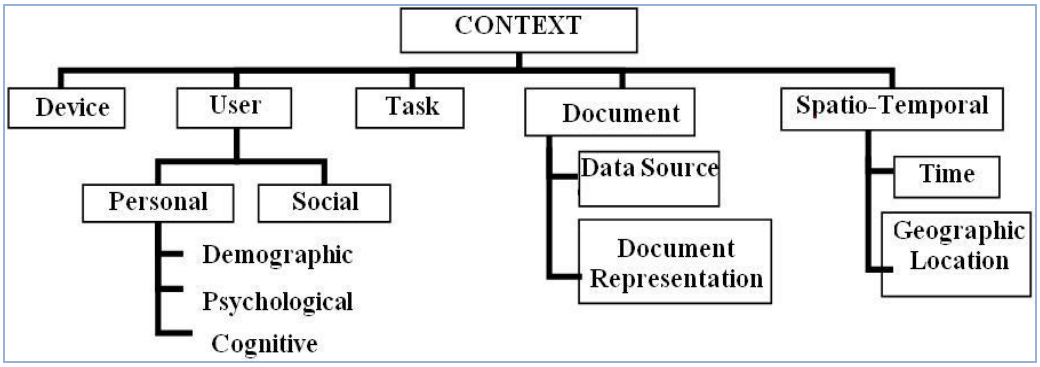


Fig. 2. The Multi-faceted concept of Context

A good example of how important is the context in forming our perception of things is presented below (Fig.3). In Porpodas (1991) it is mentioned that each sequence of images was shown to different participants during an experiment. It was observed that when people followed the upper series of images, perceived the last image as representing the face of a senior man wearing glasses. On the contrary, the participants who were shown the second sequence of images perceived the last one as the image of a mouse, although both images are exactly the same in design. This means that the sequence of images, i.e. the context within which they were represented, was determinative of the perception the

participants formed of the last image. This is a simple proof of the importance of the context within which we classify things.

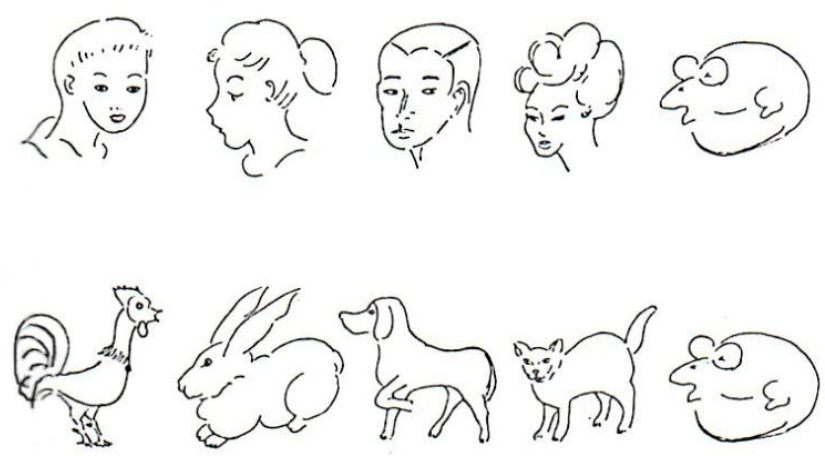


Fig. 3. Contextualised perception of reality [figure from: Porpodas (1991)]

In terms of query expansion implementations we can derive additional information about the context of the query through numerous ways. Some of them are: i) asking the users explicitly, ii) the statistical processing of log files (e.g. taking advantage of formerly created and stored query chains), iii) the qualitative evaluation of log files, data or systems, iv) pre-processing of document corpora, v) using machine learning techniques (un-, semi-, supervised), vi) the implementation of user behaviour models, vii) personalisation/ profiling of users, viii) knowledge organisation structures and many more.

The context within which things are perceived is determinative since most perceptions are conventions we make within certain frameworks. If we can communicate these conventions to many others and they embrace these conventions, then these conventions might be considered successful due to their wide acceptance. Such conventions are the Knowledge Organisation Systems (KOS).

7 About Knowledge Organisation

Alexiev and Marksbury (2010), state that the epistemological basis of any theory of knowledge organisation is an accepted postulate. In other words, how knowledge is organized and represented depends largely on the understanding of how knowledge is organized and represented. Knowledge organisation systems constitute ways for formalising knowledge and, consequently, communication through linguistic expressions, or any other kind of definitional or descriptive sign, like visual. They reflect how people (or at least some of them) perceive knowledge or how (some) people would find convenient if all significant others

perceived knowledge in the way the former have a priori defined and categorised it, with respect to facilitating communication and interoperability.

According to Hodge (2000) knowledge organisation systems are schemes for organising information and promoting knowledge management. Such schemes include 1) the term lists (authority files, glossaries, dictionaries, gazetteers), 2) classification and categories (classification scheme, taxonomy, subject headings) and 3) relationship lists (thesaurus, semantic network, ontology).

Knowledge Organisation Systems, though, bear themselves the inherent characteristics of the use of language namely ambiguity, homonymy, polysemy and synonymy, as well as they undergo the same process with which meanings are assigned to words or to any other symbol used for the purpose of communication. This leads to the understanding that whichever difficulties are present between the query and the document terms, they are also present when Knowledge Organisation Systems are involved. Additionally, even if KOSs try to capture and deliver the absolute meaning (or a more targeted one) of what they describe, they still are considered "collection independent knowledge structures" (Efthimiadis: 1996). This means that there are still missing parts for the communication of the intended meaning.

Nevertheless, Szostak (2010) states that placing a concept within a hierarchical definition establishes what sort of thing this is and what sort of thing it is not, and often sorts the subsidiary elements of which it may be comprised. In the same work is also mentioned that the degree of ambiguity lessens within groups that regularly interact (though it does not disappear). And continues that ambiguity differs only by degree between universal and domain-specific classifications, though that difference of degree is likely quite significant.

8 Query Expansion and Context in Practice

In this section we attempt a connection between techniques of query expansion and identified perceptions of what constitutes the *context* along with how it can be derived. First we encounter the simplest of the approaches, meaning the ones which treat query terms as bags-of-words, as well as flat lists of words or dictionaries or sets of synonyms (synsets), like Wordnet. These approaches rely much on part-of-speech tagging and morpho-syntactic analysis to match query terms with terms having the same or similar meaning. Usually these efforts use natural language processing techniques in order to calculate the degree of similarity between the terms.

At the same direction is the concurrent use of multiple Knowledge Organisation Systems for query expansion in order to cover more uses of the concepts involved. Approaches which consider the user's feedback (implicit or explicit) could also be perceived as trying to define a context or a network within which words obtain more discriminative meanings. Also prominent is the use of Wikipedia for expanding the users' queries. Wikipedia contains context developed and organised by the users themselves, meaning the ones that do use

it for retrieving information. Hence, it is believed that such an approach, meaning the use of terms made from users and addressed back to them, creates a more pragmatic context.

The use of ontologies, general or domain-specific, could be considered as a Wittgensteinean approach in query expansion techniques. The use of ontologies may be the best paradigm of implementing Wittgenstein's "language games" and "forms of life". Words, query terms, obtain their meaning within a specific context having defined and explicitly expressed the conditions for their use.

Even when statistical processing is involved -like in the case of exploiting users' query chains in order to detect the users' query intent- it is, once more, an expression of trying to define the context within which the user used specific words and exploit this context for delivering better results to the next user. It is the creation of a dynamic ontology which is dictated, even implicitly, by the users. User queries consist of few words in each query and the use of co-occurrence metrics does not provide significant information for the context since users also tend to use varying terms in their queries. Under this effort we can also categorise the personalisation techniques which create the user's profile, either with explicit or implicit information, and provide information by "building" on knowledge acquired for the specific user from her recorded behaviour.

In accepting that the description along with the examples of the use of a word discriminate its meaning we find the implementation of a query expansion technique which exploits the scope notes within a thesaurus (Tudhope et al.: 2006). This technique could provide more information on the use of each word, and by extent to its meaning.

Another query expansion implementation of Wittgenstein's perception is users' annotations, like social tagging, which can be used for query expansion (Biancalana & Micarelli: 2009). According to Wittgenstein the meanings of words derive from the forms of lives which are driven from the various activities of everyday life. Consequently, users could offer to others the meanings that would be understandable and useful as long as the users who are engaged in this process share common characteristics, like the family resemblances Wittgenstein defended. The users, thus, play the role of another indexer, being able to bring additional knowledge to the meanings and use of words.

9 As a Conclusion

Jaworski (2011) states that symbols in a language must be assigned meanings. The reason is that the relationship between a symbol and its meaning is a contingent one. Symbols need not have the meanings they have in fact. This is why it would not be possible to assign symbols as a very private and subjective process. We have to use these symbols to interact with others, which makes it mandatory that others know the established relations, too.

As Blair (2006) acknowledges, instead of building a "logically perfect language" that would be more precise than our day-to-day language, or using logical methods to analyze linguistic mistakes, we must reorient our investigation: Instead of looking for an *underlying logic* of language, we need to look at how language is *actually used*, for it's not an underlying logic that clarifies what we mean, it's the context, activities and practices in which we use language that provide the fundamental clarification of meaning we are looking for. And we may add here that we need user models which would be derived by behaviour studies on actual conditions showing the "real" use of words and the meanings they bear through "real" activities.

Related to the previous, Wittgenstein states in "Tractatus" (§5.6 & 5.62):

"The limits of my language mean the limits of my world"

As far as the reason and worth for investigating such a demanding and controversial domain, we embrace the statement of John Locke (1632-1704) appearing in his work "*An essay concerning human understanding"* (here quoted from Locke: 1948):

"Since it is the understanding that sets man above the rest of sensible beings, and gives him the advantage and dominion which he has over them; it is certainly a subject, even for its nobleness, worth our labour to inquire into. The understanding, like the eye, whilst it makes us see and perceive all other things, takes no notice of itself; and it requires arts and pains to set it at a distance and make it its own subject" (John Locke, 1690)

References

Alexiev, B., & Marksbury, N. (2010). Terminology as organized knowledge. In C. Gnoli & F.Mazzochi (Eds.), *Paradigms and conceptual systems in knowledge organization* (pp. 363-370). Wurzburg: Ergon.

Bhatia, M. P. S., & Kumar, A. (2010). Paradigm shifts: from pre-web information systems to recent web-based contextual information retrieval. *Webology*, *7*(1).

Biancalana, C., & Micarelli, A. (2009). Social Tagging in Query Expansion: A New Way for Personalized Web Search. *International Conference on Computational Science and Engineering, 2009. CSE '09* (Vol. 4, pp. 1060-1065). IEEE.

Blair, D. (2006). *Wittgenstein, language and information "Back to the rough ground!"* Dordrecht: Springer.

Chu, H. (2003). *Information representation and retrieval in the digital age*. Information Today, Inc.

Efthimiadis, E.N. (1996). "Query expansion", *ARIST*, vol. 31, pp. 121-187.

Foris, A. (2010). Change of paradigm in terminology: new model in knowledge organisation. In C. Gnoli & F.Mazzochi (Eds.), *Paradigms and conceptual systems in knowledge organization* (pp. 57-63). Wurzburg: Ergon.

Halpin, H., & Thompson, H. S. (2009). Social Meaning on the Web: From Wittgenstein to Search Engines. *IEEE Intelligent Systems*, *24*(6), 27-31.

Hodge, G. (2000). *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files.* Digital Library Federation.

Jaworski, W. (2011). *Philosophy of mind: a comprehensive introduction.* Oxford: Wiley-Blackwell, 2011.

Locke, J. (1948). *An essay concerning human understanding: abridged and edited by Raymond Wilburn.* London; NewYork: J.M Dent & Sons LTD; E.P. Dutton & Co Inc.

Milonas, E. (2011). Wittgenstein and web facets. In Smiraglia, R. P. (Ed.) *Proceedings from North American Symposium on Knowledge Organization* (pp. 33-40). Vol. 3 (1). Toronto: NASKO.

Porpodas, D.K. (1991). Cognitive psychology. Vol. 1: Learning process. Vol. 2: Θέματα Issues on language psychology. Problem solving. Athens: [s.n.], 1991. [in Greek]

Seidel, A. (1991). Plato, Wittgenstein and Artificial Intelligence. *Metaphilosophy*, *22*(4), 292-306.

Shiri, A., & Revie, C. (2006). Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation. *Journal of the American Society for Information Science and Technology*, *57*(4), 462–478.

Szostak, R. (2010). Universal and domain-specific classifications from an interdisciplinary perspective. In C. Gnoli & F.Mazzochi (Eds.), *Paradigms and conceptual systems in knowledge organization* (pp. 71-77). Wurzburg: Ergon.

Tudhope, D., Binding, C., Blocks, D., & Cunliffe, D. (2006). Query expansion via conceptual distance in thesaurus indexed collections. *Journal of Documentation*, *62*(4), 509-533.

Wittgenstein, L. (1922). *Tractatus logico-philosophicus*. New York; London: Harcourt Brace & company inc.; K. Paul Trench Trubner & co. ltd.

Wittgenstein, L. (1958). *Preliminary studies for the "Philosophical investigations": generally known as the Blue and Brown books*. London: Blackwell.

Wittgenstein, L. (1967). *Philosophical investigations, Transl, by G. E. M. Anscombe, (Repr.).* Oxford: Basil Blackwell.

Policies for Geospatial Collections: a Research in US and Canadian Academic Libraries

Ifigenia Vardakosta and Sarantos Kapidakis

Laboratory on Digital Libraries and Electronic Publishing
Department of Archive and Library Sciences,
Ionian University, Corfu, Greece,
{ifigenia, sarantos}@ionio.gr

Abstract. Geographic information is essential to economical and political implications. The technological development of recent years and the simplicity of many applications have made it part of everyday life of citizens.
It is common to libraries, especially those who serve departments interested in geographic content to organize such collections and provide services to their users.
The rapid diffusion of free data in the internet and the growth of open access software have begun to affect libraries in adopting policies related to the collections and services of GIS.
The work presented in this paper seeks, through the investigation of geospatial collection development policies' of 21 academic libraries in United States and Canada to identify those characteristics reflecting their adaptation to the new era of open data.

Key Words: Geospatial collections, collection development policies, academic libraries, open data, surveys

1 Introduction

The geographic information constitutes an important type of information connected to the daily activities of all citizens and to issues related to the broader environment in which they live and develop as well. The potential offered now by the digital publishing and communication environment have contributed to the rapid diffusion of digital geographic information through the internet via different channels and output in different forms. The term "geographic information" may until recently have been connected to the meaning and use of printed maps, the spatial diffusion datasets, demographic data, remote sensing images, orthophotos, etc., ie information relating to a site, has led to the use of the term "spatial data" while the original term "geographic information" performs better issues related to features that describe the earth's surface. Both terms are used in the international bibliography and are closely connected to the technology of GIS.
It is commonly accepted that libraries are typical agencies that organize, manage and disseminate knowledge, therefore their involvement with the printed geographical information is not a new discovery. However in recent years, the various economic

and social conditions and adaptation to the growing needs of their users in conjunction with geographic information continuous flow on the Internet, led to the development of new collections and services to their patrons. Consequence of technological evolution constitutes the rapid development of GIS applications and their users, while increasing the demand for geospatial data and the transfer over the Internet [7]. Computing technologies, such as sensor computing, cloud computing, mobile computing, visual computing, business intelligence, spatial database server, and high-performance computing, play key roles in geospatial technologies and applications[1].

Geospatial Collection Development Policies refer to all the necessary procedures adopted and recorded by a library in order to develop print and digital geospatial collections capable of meeting their user's information needs.

Although several studies have been performed for GIS in libraries, though there is a gap on researches about policies related to the development of geographical collections and this paper attempts to contribute towards this direction.

2 Policies and Libraries

According to IFLA Guidelines for a collection development policy "the main reasons for having a written collection development policy can be put under four broad headings: 1.selection 2.planning 3.public relations 4.the wider context".

Collection development policy statement is a necessary tool for a librarian used to "describe an individual library's objectives in developing its collections" [4], and as [28] argues "is no longer just about creating the physical collection but more important is the notion of providing access to information regardless of format or location".

In the world of digital libraries, a policy is typically described as a condition, term or regulation governing the operation of a digital library or some aspect thereof. People (such as digital library staff members, managers, and stakeholders) make policies for digital libraries. Sometimes, these policies can be expressed as rules. Rules provide mechanisms to express complex policies in ways that computer systems can interpret and apply them. At a user's level, digital library access policies must be enforced, and users often need to "be informed of the policies and educated as to what constitutes a

[1] The importance of the above are clearly underlined by US Government which recently announced a 'Big Data' research and development initiative in response to processing the large amount of data collected by geospatial and other systems. Under this initiative, several federal government agencies, NSF, USGS, DARPA, DOD, NIH, and DOE, commitment for the programs total $200 million. Big data refers to the rising flood of digital data from many sources, including the sensors, digitizers, scanners, software-based modelling, mobile phones, internet, videos, e-mails, and social network communications. The data type could be texts, geometries, images, videos, sounds, or their combination. Many of such data are directly or indirectly related to geospatial information. The emerging opportunity arises from combining these diverse data sources with greatly improving computing tools and techniques needed to access, organize, analyze, visualize, and extract useful information from huge diverse data sets [http://www.com-geo.org/conferences/2012/topics.htm]

reasonable behaviour" normally through usage policies. At a repository or at a collection level, formalized policies can be followed through trusted systems or through secure combiner (encryption, digital signatures, and public-key encryption). [19].
In the "Framework of Guidance for Building Good Digital collections" that Institute of Museum and Library Services (IMLS) Digital Library Forum [2] created the *Framework Collection Principle 1* recommends that digital collections be "*created according to an explicit collection development policy that has been agreed upon and documented before digitization begins.*" As in the same paper appears "there is confusion between collection development policy and digitization selection guidelines, which though closely related are not synonymous" and authors based this argue in "the lack of substantial pre-existing collection development policies hint at problems engendered by the opportunistic way that digitization and digital collection creation is undertaken".
As geospatial data by nature are unique and complicated and dependent upon software and hardware for access and analysis an essential step in creating and integrating GIS services and collections in an academic library is in creating a sound collection development policy. A number of factors as user needs, available budget, technological infrastructure and staff development programs, are important factors in constructing a policy [1]. A policy can be understood as political, management, financial, and administrative mechanisms structured to ensure the delivery of certain consistent outcomes or behaviours.

3 Researching US and Canadian academic libraries for geospatial policies

3.1 Research Questions

Academic libraries that already own or want to develop geographical collections, have to deal with the phenomenon of the rapid diffusion of open geographical data on the web while must offer their users additional services in order to cover the continuously increasing demands in times of low budget. Our research question relates to whether or not reflected in the policies adopted by each library to develop its geographical collections, the move towards open data. Furthermore, a research in the content of written policies will clarify what are finally the issues that a library consider as important to include in its regulations and communicate to patrons through its website.
The research questions formed in this context are:

1) What are the main features of geospatial collection development policies?
2) Do geospatial collection development policies include features that reflect the adjustment of libraries to the rapid growth of open geospatial data?

[2] In the spring of 2001, the Institute of Museum and Library Services (IMLS) convened a Digital Library Forum to discuss the implementation and management of networked digital libraries (DLs), including issues surrounding DL infrastructure, metadata, the use of thesauri and other forms of authorities for controlled terminologies, and the use of automated processes for content enrichment, e.g., to better support inclusion of digital resources in curriculum materials and teacher guides.

3) Do the existence geospatial collection policies reflect the adjustment of libraries to limited financial means the last few years?

The aforementioned research question comes also to explore the conclusions of an [5] where participants in their comments *"indicate that they expect growth in the demand for digital spatial data and are revising collection development policies to address this need"*. Our research involves libraries that are members of ARL and initial members of ARL GIS Literacy Project (8 from US, 7 from Canada), however, it does not focus in libraries that meet this requirement.

3.2 The Followed Methodology

The specific work expands our previous researches which were mainly quantitative and aimed to identify the existence of geospatial collections in academic libraries and geospatial collection development policies as well [29-30]. More specifically relied on the survey we undertook from May to August 2011 and in which we searched websites of a stratified sample of academic libraries in US and Canadian Universities which inter alia operate those departments whose curricula are based on the use of geospatial information and GIS e.g. Geography, Geology, Topography, Earth sciences, Environmental sciences etc. in other words to serve departments where GIS systems are necessary for education and research. To identify those academic libraries we used Libwebcats[3], a directory of libraries throughout the world, and in addition the Libweb[4] a directory of library home page. Among the examined policies are those of major universities that were pioneers in developing geographical collections and establishing GIS services (e.g. University of California – Santa Barbara).
This method of analysis of web content and data collection using the combination of browsing and searching is similar to the coding technique described by [16], who proclaim that *"a library's web site can provide a powerful forum for communicating with users"*. The aforementioned researchers used this technique to analyze the use of library web pages to communicate specific information to faculty, while [8] used it to search the use of library web pages to promote data resources to all researchers. [33] used the same methodology, in part, to "*understand to what extent academic libraries are participating in GIS Day events on their campuses, as well as to what extent those events are being promoted and described on the library's web pages and through the dedicated web site gisday.com"*.
We considered content analysis method as the proper one for the specific research since our purpose was to investigate the policy texts, therefore we had to deal with specific words which they represent specific activities (e.g. acquisition) or issues (e.g. purpose) of libraries which in the bottom line harmonize their operations in a specific way easily understandable by its users.
What differentiates our research from those mentioned above is the focus on policies regarding geographical collections nowadays that libraries facing financial problems and the technological potential can be used as a means of continuous provision of services and collections.

[3] http://www.librarytechnology.org/libwebcats/
[4] http://www.indiana.edu/~librcsd/internet/libweb-mirror/

3.3 Data Collection

We focused our research on policies provided by 21 libraries (13 in US and 8 in Canada).
More specifically our research was held through the following steps:

1) At first we recorded the number of libraries that have policies. This was the result of browsing the initial sample of 331 libraries' websites for: a) the library's "GIS collection/data", "GIS services", or "geospatial collection/data", or "geographical collections/data" webpage if existed, b) for relevant terms "geospatial policies", "collection development policies". If there was no result when using those terms, we searched the library's home page and tried to locate the link "policies" or "collection policies".
2) Our next step was to download all the policies we found.
3) Step three was to make an Excel spread sheet where the basic elements were the libraries' names.
4) Each policy was read carefully and it appears that each of those texts comprises of certain specific categories of information (Fig.1) and each category includes several topics. The lists of words based on the top words of topics indicative of policies were used as *features*.
5) In step five, those specific categories (which also form the features we seek to measure) mentioned by each of the policies, were grouped together so as to isolate what was common to all and highlight those referred to specific cases. Thus, we grouped results in 5 basic broader categories.
6) As each policy text is a mixture proportion of categories and topics and each topic corresponds to a specific category of information, it was in step 6 that we assigned topics to categories, depending on the library's will to give detail or brief information related to one category or topic than the others. Similarly, the layout of position of each category of information varies, depending on library's decision.
7) After this, we recorded the number of topics related to the broader ones and ranked them by frequency of appearance in policies.
8) Finally, we constructed Tables which contain each Feature and related Topics.

Fig.1: Categories and Topics in policies' text

3.4 Results

Our study focused on 21 academic libraries (13 in U.S. and 8 in Canada as shown in Fig.2) which had geographical collections development policies as shown by prior research [29-30].

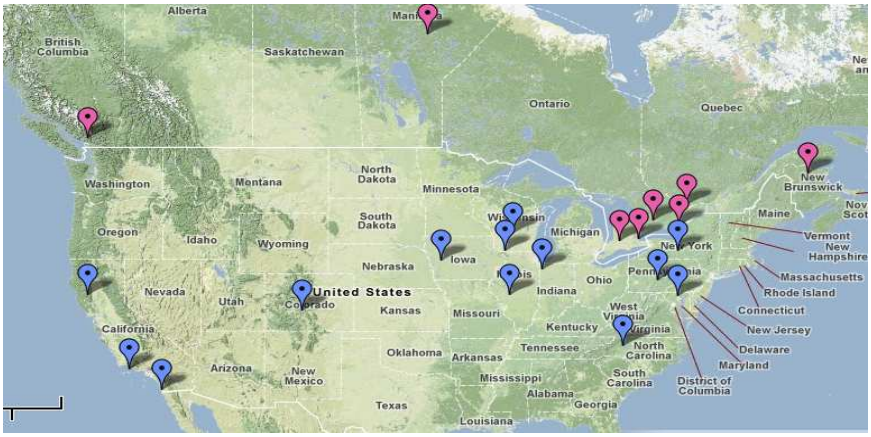


Fig.2: Examined academic Libraries of US and Canada

As we mentioned above, results were grouped and organized in Tables. Thus, answering our **first** research question which concerns the *main features of geospatial collection development policies*, our survey revealed that these are:

1) "General information"(Table 1)
2) Information regarding "Collection" (Table 2)
3) Information regarding "Data" (Table 3)
4) Information regarding "Open Access" (Table 4)
5) Information regarding "Cooperation"(Table 5)

Analyzing each one of the characteristics above that the majority of academic libraries provided on their website for informing their users, we can identify that in the first feature *"General Information"* we can distinguish the following subset of topics presented below,[5] classified according to the extent of their appearance in the policies' text: 1) Date created/revised/updated[6] (14), 2) Person related to/responsible for the collection development policy (11), 3) Department Description/Academic Program Support (5) 4) Special considerations for collection development (1) 4) History (1) 5) Location of GIS Collection (1). As Table 1 illustrates the occurrence of *Date created*/revised/updated/ and *Person related to /responsible for collection development policy* are in high percentages while *Special considerations for Collection Development*, *History* and *Location of GIS* remain slightly less prevalent than *Department Description*.

[5] Results are presenting in numerical order as ranking shown in Tables. The number in parentheses indicates the number of policies on which this term detected.
[6] We consider appropriate to point out all the names under which the particular category was recorded.

Table 1.
General Information

Rank	Topic	No of policies	Percent (n=21)
1	Date created/revised/updated/	14	66.6%
2	Person related to/responsible for collection development policy	11	52.4%
3	Department Description/Academic Program Support	5	24%
4	Special considerations for collection development	1	4.8%
4	History	1	4.8%
4	Location of GIS Collection	1	4.8%

In Information regarding "*Collection*" (Table 2) are several topics relating to the collections. These are: 1) *Collection Purpose/Purpose of the collection/General Collection principles* (10), 2) *Collection Guidelines* (10) 3) *Selection/Evaluation & Prioritization* (4) 4) *Audience/Description of users/*Distribution (4) 5) Collection Profile/Description/Level/Brief Overview (4) 6) Acquisition/s (2) 6) *Price* (1).
It is worth analyze the subset "Collection Guideline" since in most academic libraries consists a main part of policies which is divided in other topics as : *Subject boundaries/priorities* (11), *Publication dates collected* (9), *Languages* (9), *Geographical range* (8), *File formats and types* (8), *Type of materials included and excluded* (5), *Chronological span/limits* (4).

Table 2.
Collection

Rank	Topic	No of policies	Percent (n=21)
1	Collection Purpose/Purpose of the collection/General Collection principles	10	47.6%
1	Collection Guidelines:	10	47.6%
	Subject boundaries/priorities	11	52.4%
	Publication dates collected	9	42.9%
	Languages	9	42.9%
	Geographical range	8	38.1%
	File formats and types	8	38.1%
	Type of materials included and excluded	5	23.8%
	Chronological span/limits	4	19.04%
2	Selection/Evaluation & Prioritazation	4	19.04%
2	Audience/Description of users/Distribution	4	19.04%
2	Collection Profile/Description/Level/Brief Overview	4	19.04%
3	Acquisition/s	2	9.5%
4	Price	1	4.5%

Information regarding the feature "Data" as shown in Table 3 gathers the following topics: 1) *Use/Licensing/Restrictions/Copyright* (4), 2) *Data* (3) 3) *Weeding* (3) 4) *Metadata* (2) 5) *Documentation* (2) 6) *Software support* (2) 7) *Citation* (1)

Table 3.
Data

Rank	Topic	No of policies	Percent (n=21)
1	Use/Licensing/Restrictions/Copyright	4	19.04%
2	Data	3	14.3%
2	Weeding	3	14.3%
3	Metadata	2	9.5%
3	Documentation	2	9.5%
3	Software support	2	9.5%
4	Citation	1	4.8%

Availability of "Open Access" is shown in Table 4 and is expressed through 1) *Governmental sources* (e.g. US Sensus Bureau, municipal agencies) (10) 2) *Depository programs* (e.g. FDLP, USGS, Canadian Topographic maps & data) (8) 3) *Commercial firms* (8) 4) *Free data* (3) 5) *Gifts* (3) 5) *Consortia arrangements* (2) 6) *Non-profit entities* (e.g. professional organizations or environmentally focused non profits) 7) *Products issued by people* (1)

Table 4.
OPEN ACCESS
(availability of data)

Rank	Topic	No of policies	Percent (n=21)
1	Governmental sources (e.g. US Sensus Bureau, municipal agencies)	10	47.6%
2	Depository programs (e.g. FDLP, USGS, Canadian Topographic maps & data)	8	38.1%
2	Commercial firms	8	38.1%
3	Free data	3	14.3%
3	Gifts	3	14.3%
4	Consortia arrangements	2	9.5%
5	Non-profit entities (e.g. professional organizations or environmentally focused non profits)	1	4.8%
5	Products issued by people	1	4.8%

"Cooperation" details in policies are addressing according Table 5 with 1) *Cooperative arrangements and related collections* (7) and 2) *Interdisciplinary Relationships* (2).

Table 5.
COOPERATION

Rank	Topic	No of policies	Percent (n=21)
1	Cooperative arrangements and related collections	7	33.3%
2	Interdisciplinary Relationships	2	9.5%

According to the findings that answer the second research question of the investigation, the features *that reflect the adjustment to the rapid growth of open geospatial data* could be considered as Governmental sources (e.g. US Sensus Bureau, municipal agencies) and Depository programs (e.g. FDLP, USGS, Canadian Topographic maps & data) as shown in Table 4, since they appear 10 and 8 times accordingly in library's policies. On the contrary, Non Profit Organizations or Products issued by people are not familiar in academic libraries, since only 1 library mentions them as a source of data while Free Data and Gifts is used by 3 libraries, and Consortia Arrangements by 2 libraries.

"Open Access" and *"Cooperation"* could be considered as features that reflect the adjustment to limited financial means, and which answer to our third research question, as shown in Table 4 and Table 5 of our results.

4 Related Work and Discussion

In the international bibliography, literature relating to policies concerning geographical/geospatial information can be classified in 2 types: 1) those related to researches on the implementation of GIS in libraries referring policies aspects in their content and 2) those articles that were written specifically for policies.
In the above context and in the *first type* of articles the majority of them appeared after 1992, which was the year that ARL GIS Literacy project implemented in libraries. [2] points out the CDPs because "the management of and efficient access to it [spatial data] is one of the key challenges that librarians face as GIS service providers" while [24] states that "in an academic library in institution with active GIS initiatives need to identify and establish contact with faculty to determine teaching and research needs". [3] argues that "developing collections of GIS-related materials and spatial information in support of teaching, research, and public access is an important first step in initiating a GIS service policy and in assisting library staff to become GIS literate". [5] conducted a survey for member libraries (123) and as revealed of it the demand for geospatial data seem to be growing and participants' comments that are revising collection development policies so to address this need.

[13] in his paper for GIS collection development in Harvard University argues that "GIS collection development does not always coincide with the organization's traditional collection policy" and names how should a librarian act so to formulate a successful CDP. [25] after examined 69 academic libraries' websites concluded that "regularly assessing and revising policies helps academic library adapt GIS services to strike a balance between ever-changing needs of users and finite library staff, equipment and budgetary resources". For the best accomplishment of National Geospatial Digital Archive project[7], University of Santa Barbara at California and Stanford University which are partners in it, created three CDP because "in the long term this strategy will support more breadth to the archive as well as leverage the strengths of each institutions" [12].
In the *second type* of articles, [31] outlines CDP that can be applied to many types of information agencies especially as a step toward the identification and standardisation of effective practices in which were based on in Mann Library at Cornell University. [9-10] in his several works regarding policies for geospatial collections emphasizes their necessity through the focus on the key areas of interest to the geospatial community which indicates as pricing, copyright, security, privacy, licensing, and access and use. In the same philosophy lies [27] when describes the development of a data management policy for the Cornell University Geospatial Information Repository (CUGIR) while she illustrates that "in developing a policy, data distributors are advised to consider such issues as intellectual property rights, liability issues, distribution methods and services, data and metadata management practices, security risks posed by geospatial data, and user limitations".
The bibliography related to the involvement of libraries in developing geographical collections over the last years is significantly increasing [34] as the aforementioned articles indicate. Nevertheless, the content of policies related to geographical collections needs further exploration.

The present study aimed to highlight the policies' characteristics of geospatial collections as they are displayed on the web pages of academic libraries.
This examination of academic libraries web pages has shown that many libraries have chosen to make available collection development information through the internet. Of the web sites examined 21 had some type of collection management statement that ranged from a thoughtful detailed policy to a single sentence mission statement. This left a large number of libraries with no collection information that could be found in their web pages.
The approach that libraries chooses to develop geospatial data determines in a way how policies will be communicated since according to the analysis of our findings 6/21 policies were only for GIS collections, 5/21 along with map collection and 8/21 along with geographical collection.
It is worth mentioning the heterogeneity of policies texts we studied for completing this research. They did not follow a specific formula since in some libraries are analyzed and recorded in detail and are multi paged while in some other documents provided, contain epigrammatic information regarding important issues like acquisition or data distribution (e.g. Emory University Library).
The difference in used terminology is one of the policy's attributes highlighted through this work. This difference in terminology can easily be explained since each

[7] The project funded by the Library of Congress and the goal of the collaboration was to collect, preserve, and provide long-term access to at-risk geospatial data http://www.ngda.org/

library formulates its own policies in accordance with its own priorities and potentials and there is not any guideline text from e.g. an Association that libraries could rely on for developing their own documented policies. As we noticed through this study, there are libraries which could be considered as pioneers in geospatial data collections and have developed well formed texts which could easily be used as a guideline.
The collection development statements we studied relies their usefulness mainly on information about the data and their format as well as the way that a user can have access to it. Other information similar to those given for the non geospatial material is also provided e.g. who has the right to access the information.
Despite the fact that some libraries have developed portals of freely available data that librarians detected on the internet, in the policies' text open geospatial data are mentioned by the minority of libraries only in order to costs and budgets in conjunction with the types of data and their scales. Another point worth to be highlighted is the fact that a number of libraries are developing geospatial collections taken into account are the collections of other libraries nearby (e.g. University of Pennsylvania, University of Chicago). An absolute increase in emphasis on collaborative approaches to collection development can be detected through these movements and all these trends derive from a need to reduce the financial costs.
Although Free Data and Gifts are considered to be for library professionals common practice for data supply, however in geospatial data as we can identify this does not happen regularly since only 3 libraries refer those two ways of having data
The lack of GCDPs from countries outside America may mislead the potential researcher since US and Canada has lots in common in applying library science. Without any input from participants in library environment (users and librarians) this research is limited to what can be seen and inferred from the written policies. While the sample includes libraries that have published their policies in the World Wide Web, we cannot ignore those ones that although they have written policies nevertheless for some reasons have chosen not to upload them on their website. Therefore the focus on internet published policies will not allow any comparison with internal written documents that may or not exist at the rest libraries with GIS. Given these limitations, any general statement will be clearly limited.

9. Conclusions and Future Work

It is clear that separate collection management pages are the preferred vehicles for presenting information about the collections [16]. The present research reported that the main features of GCDPs are information regarding: General Information, Collection, Data, Open Access, and Cooperation. The topics that in the majority of collection management policies for geospatial collections appears: Person related to/responsible for collection development policy (52.4%), Collection Purpose (47.6%), Collection Guidelines (47.6%), Subject boundaries/priorities (52.4%), Governmental sources (e.g. US Sensus Bureau, municipal agencies) (47.6%), Use/Licensing/ Restrictions/Copyright (19.04%), Cooperative arrangements and related collections (33.3%).
These five features aforementioned are those we revealed from our research in US and Canadian academic libraries. It would be interesting to further explore the written policies of academic libraries in other countries of the world e.g. Europe, where also have developed geographical collections. Recent development in managing geospatial data (e.g. linked data) along with the adoption of new strategic actions (e.g. co

operations) are potentials that libraries should exploit. Therefore, we consider that policies related to geospatial data have not been adequately examined.

References

1. Abresch, J.e.a. Integrating GIS into Library Services: a guide for academic libraries. Hershey: Information Publishing Company (2008)
2. Abbott, L.T. and Argentati, C.D. GIS: a new component of public services. The Journal of Academic Librarianship (July), p.251-256 (1995)
3. Adler, P.S. & Larsgaard, M.L. Applying GIS in libraries. In P. Longley, ed. Geographic Information Systems. Chichester,New York: Wiley, pp. 901-908 (2002)
4. ARL. The ARL GIS Literacy Project. Spec Kit 238 (1999) http://www.eric.ed.gov/PDFS/ED429609.pdf.
5. ARL. Spatial Data Collections and Services.Spec Kit 291 (2005). Available at: http://www.arl.org/bm~doc/spec291web.pdf
6. Arms, W.Y.Digital Libraries. Cambridge, MA: MIT press (2001).
7. Beard, K.. Digital spatial libraries: A context for engineering and library collaboration. Information Technology and Libraries, 14(2), pp. 79- 86 (1995).
8. Bennett, T.B. and Nicholson, S.W. Research Libraries: Connecting Users to Numeric and Spatial Resources. Social Science Computer Review, 25(3), p.302-318 (2007). http://ssc.sagepub.com/cgi/doi/10.1177/0894439306294466
9. Boxall, J. and Anderson, C. Geospatial Information Management: spatial is still special. Dalhousie Journal of interdisciplinary Management, Spring (2005).
10. Boxall, J. Advances and Trends in Geospatial Information Accessibility – Part II□: Policy Dimensions, Journal of Map & Geography Libraries, vol. 3, no. 1, pp. 57-78, (2006).
11. Cole, T.W. e.a. Findings Pertaining to the Framework of Guidance for Building Good (2006). Digital Collections https://www.ideals.illinois.edu/bitstream/handle/2142/722/UIUC-IMLSDCC-FrameworkAssessmentfinal.pdf?sequence=2
12. Erwin, T. and Sweetkind-Singer, J. The NGDA: a collaborative project to archive geospatial data. Journal of Map & Geography Libraries, 6(1), pp.6-25 (2010).
13. Florance, P. GIS collection development within an academic library. Library Trends, 55(2), pp.222-235 (2006).
14. Gabaldon, C. and Repplinger, J. GIS and the academic library: a survey of libraries offering GIS services in two consortia in Issues in Science & Technology Librarianship, 48, Fall, (2006). http://www.istl.org/06-fall/refereed.html
15. Guptil, H. Spatial data standards and information policy. Government information Quarterly 11(4), pp 387-401 (1994).
16. Hahn, K.L. and Schmidt, K. Web communications and collections outreach to faculty. College and Research Libraries 66(1), pp.28-40 (2005).
17. Herold, P.(e.a). Optimizing web access to geospatial data: the Cornell University Geospatial Repository (CUGIR) (1999) http://www.istl.org/99-winter/article2.html
18. IFLA: About the Acquisition and Collection Development Section. Available at: http://www.ifla.org/en/about-the-acquisition-collection-development-section
19. Innocenti, P., Vullo, G., Ross, S. Towards a digital library policy and quality interoperability framework: the DL.org project. New Review of Information Networking Vol. 15, pp. 29-53 (2010)

20. Janee, G. Preserving geospatial data: The National Geospatial Digital Archive's approach. (2009).
http://www.alexandria.ucsb.edu/~gjanee/archive/2009/archiving-2009-paper.pdf
21. Kinikin, J.N. and Hench, K. Survey of GIS implementation and use within smaller academic libraries, Issues in science and Technology Librarianship, vol. spring, 8 p., (2005).
22. Kinikin, J.N. and Hench, K.). Follow-up survey of GIS at smaller academic libraries in Issues in Science and Technology Librarianship, Summer (2005a).
23. Kim, S. and DeCoster, E. Organizational schemes of information resources in top 50 academic Business Library websites in The Journal of Academic Librarianship 37 (2) pp.137-144 (2011).
24. Longstreth, K. GIS collection development, staffing and training in The Journal of Academic Librarianship, July, 267-274 (1995).
25. Sorice, M. An analysis of GIS services websites in academic libraries, Master Thesis, (2006). [http://etd.ils.unc.edu/dspace/handle/1901/303]
26. Salem Jr., J.A. Spatial Data Collections and Services. ARL Spec Kit 291 (2005). [http://www.arl.org/bm~doc/spec291web.pdf]
27. Steinhardt, G. Libraries as distributors of geospatial data: Data management policies as tools for managing partnerships. Library Trends 55(2), pp. 264-284 (2006).
28. Tucker, J.C. and Torrence, M. Collection development for new librarians: advice from the trenches. Library Collections, Acquisitions & Technical Services, 28, pp. 397-409 (2004)
29. Vardakosta, I. and Kapidakis, S. Geographic collections development policies and GIS services: a research in US academic libraries' websites, In First Workshop on Digital Information Management, Corfu 30-31 March, pp.89-98 (2011).
http://eprints.rclis.org/bitstream/10760/15851/1/08.Vardakosta.pdf
30. Vardakosta, I. and Kapidakis, S. Geospatial collection development policies in academic libraries: a worldwide research. In 17th European Colloquium on Quantitative and Theoretical Geography (ECQTG2011), September 2-5, (2011a). http://eprints.rclis.org/handle/10760/16096#.T0-PsHncC2o
31. Walters, W. Building and maintaining a numeric data collection. Journal of Documentation 55 (3), pp.271-287 (1999)
32. Weber, R.P. Basic Content Analysis. London: Sage publ. (1990)
33. Weimer, K.H (e.a) GIS Day and Web Promotion: Retrospective Analysis of U.S. ARL Libraries' Involvement. Journal of Map & Geography Libraries, 8 (1), p.39-57 (2012), http://www.tandfonline.com/doi/abs/10.1080/15420353.2011.629402
34. Wolf, J.H. Are we doing enough to prepare future librarians to meet the challenges and demand? (2011),
http://jenwolf.net/wordpress/wpcontent/uploads/2011/10/Wolf_MLIS_capstone_Dec20111.pdf

Appendix 1: Examined Academic Libraries List

LIBRARY NAME	GEOGRAPHICAL/GEOSPATIAL COLLECTION DEVELOPMENT POLICIES URLs
Carleton University Library	http://www.library.carleton.ca/about/policies/collection-development-gis-resources
Cornell University	http://cugir.mannlib.cornell.edu/CUGIRCollectionDevtPolicy_20060825.pdf
Duke University	http://library.duke.edu/research/subject/guides/maps/map_policy.html
George Washington University	http://www.gelman.gwu.edu/collections/policies/maps-and-gis.pdf/view
Iowa State University Library	http://www.lib.iastate.edu/cfora/pdf/3000057.pdf
McMaster University	http://library.mcmaster.ca/collections-services/policies/lloyd-reeds-map-collection
Queen's University	http://library.queensu.ca/research/collections/maps-geospatial-data-and-air-photos#geograph
Ryerson University	http://www.ryerson.ca/library/info/collections/colldev/material.html
Simon Fraser University	http://www.lib.sfu.ca/collections/collections-policies/geography
Stanford University / GIS at Branner	http://lib.stanford.edu/gis/
University of California-San Diego	http://libraries.ucsd.edu/_files/sshl/pdf/Geospatial-Data-Collection-Plan.pdf
University of California-Santa Barbara	http://www.library.ucsb.edu/services/policies/collections/geogcdp1.html
University of Chicago	http://guides.lib.uchicago.edu/content.php?pid=115216&sid=1220061
University of Colorado at Boulder	http://ucblibraries.colorado.edu/collectiondevelopment/geography.htm
University of Illinois Urbana-Champaign	http://www.library.illinois.edu/gex/classes/collectiondevelopmentgeosciences.html
University of Manitoba	http://www.umanitoba.ca/libraries/units/datalib/gis/gis.html
University of New Brunswick	http://www.lib.unb.ca/about/policies/colldev-UNBF.php#II
University of Pennsylvania	http://www.library.upenn.edu/collections/policies/maps.html
University of Waterloo	http://www.lib.uwaterloo.ca/staff/irmc/collectionsmanagement.html
University of Wisconsin-Madison	http://www.geography.wisc.edu/maplib/Docs/GISData_Dist_Policy.pdf
University of Wisconsin-Milwaukee	http://www4.uwm.edu/libraries/CollPolicy/u-agsl.cfm

Path-based MXML Storage and Querying

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,
Artemidos 6 & Epidavroy 15125 Maroussi, Greece.
yannis@imis.athena-innovation.gr

Abstract. 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 path-based techniques for storing MXML documents in relational databases, based on similar techniques previously proposed for conventional XML documents. Also, we present the basic ideas behind an algorithm for converting context-aware MXML queries to SQL queries for execution over the MXML data which are stored in the relational database.

1 Introduction

The previous years, the problem of storing XML data in relational databases has been intensively investigated [4, 12, 13, 15]. 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 equivalent SQL for evaluation. After the execution of SQL queries, the results are translated back to XML and returned to the application.

Multidimensional XML (MXML) [9] 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.

In this paper, we present a path-based approach for storing MXML in relational databases. For querying MXML data, we use *Multidimensional XPath* (MXPath) [8], which is an extension of conventional XPath containing all additional characteristics of MXML. Finally, the basic ideas behind an algorithm for converting MXPath queries to SQL queries is described.

* Supported by State Scholarships Foundation of Greece(IKY), Makri 1 and Dionisiou Areopagitou 117 42 Athens - Greece

2 Preliminaries

2.1 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 [17, 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 [11, 15, 5, 16, 12, 3, 13].
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 [15, 5, 16, 18, 12, 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.

2.2 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 [9, 10]. 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 [9] where a related process called *reduction* is presented.

3 Properties of MXML documents

3.1 Graphical model of MXML and Node Indexing

In this section we present a graphical model for MXML called *MXML-tree*. The proposed model is node-based and each node is characterized by a unique "id".

For indexing the nodes of a MXML tree, we use the Dewey labelling schema. In general, this schema assigns to each node a dotted format identification number, according to the hierarchical position (level, sibling number) of that node in the MXML tree. So, assuming that $\{N_i\}$ with i=1,2,...,d is the set of nodes contained in a path of the MXML tree, such that N_1 is the root node and N_{i-1} is the parent node of node N_i, we define the Dewey labelled index of node N_d denoted as the dotted format identification number $s_{N_1}.s_{N_2}.s_{N_3}\ldots s_{N_d}$, where s_{N_i} is the position number of node N_i among its siblings.

In MXML-tree, 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 tree. 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,

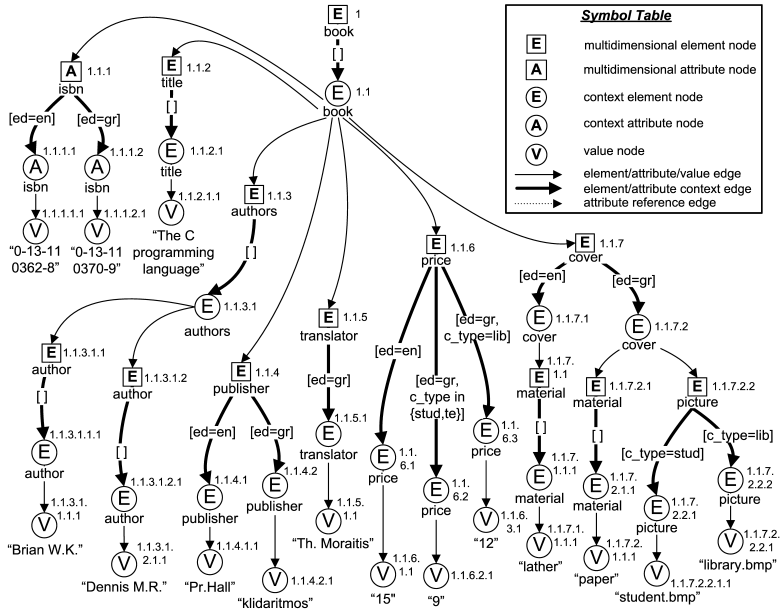


Fig. 1. Graphical representation of MXML (MXML tree)

we see the representation of the MXML document of Example 1 as a MXML-tree. Note the dotted format of dewey indexing, with the number of dots representing the hierarchical level of each node and the number after a dot, the

position number of a node among its siblings. For example, node `1.1` is a level 1 node (starting from level 0 of the root node) and the first child of node 1. Also, note that some additional multidimensional nodes (e.g. nodes 1.1.2 and 1.1.3) have been added to ensure that the types of the edges alternate consistently in every path of the tree. This does not affect the information contained in the document, but facilitates the navigation in the tree 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-tree 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 [14] 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-tree. By definition, the inherited context of the root of the tree is the universal context [ ]. Note that contexts are not inherited through attribute reference edges.

As in conventional XML, the leaf nodes of MXML-trees 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 [9]. 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 [14] 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-tree.

4 MXML Querying

4.1 Multidimensional XPath

Multidimensional XPath (MXPath) [8] is an extension of XPath used to navigate through MXML-trees. 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.

Bellow, we present a MXPath example with the explanation of its evaluation on the MXML-tree of Figure 1.

Example 3. Retrieving multidimensional nodes according to their explicit context.

Query: *What is the material of the cover of the english edition of the books?*

MXPath:

`/child::book/child::cover[ec()="edition=en"]/child− >material`

This query returns multidimensional nodes (because of the notation $-$ >) labeled "`material`". Also, the step `child::cover[ec()="edition=en"]`, because of the contained ec qualifier, returns the context node with ID 1.1.7.1. So, the final result is node 1.1.7.1.1.

4.2 MXPath and Tree Patterns

Given MXPath queries containing branches, context (/) or multidimensional ($-$ >) child edges, descendent edges (//), context qualifiers, "*" wildcards and value nodes, we are able to construct *Multidimensional Tree Patterns*. A Multidimensional Tree Pattern is a graph that depicts the constrains posed by a MXML query. The sub-path of a Multidimensional Tree Pattern which starts from the root and leads to the final result of the query, without containing any branches, is called *Selection Path*.

Example 4. In Fig. 2, we see the multidimensional tree pattern expressing the query of Example 3. Notice the nodes marked as (M) denoting that these nodes should be multidimensional and the thick edges, which should lead in context nodes. Also, we can see the ec context qualifier "`[ec()="edition=en"]`" expressing the constraints, that should be posed as a predicate of the query, over the context node "`cover`". Finally, we can see that output nodes (at the end of the selection path) are denoted as circles with node labels inside them.

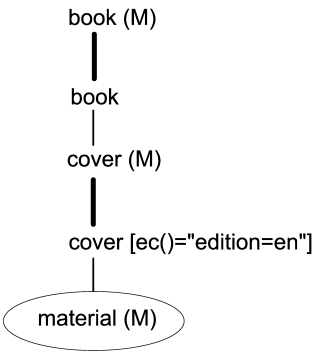


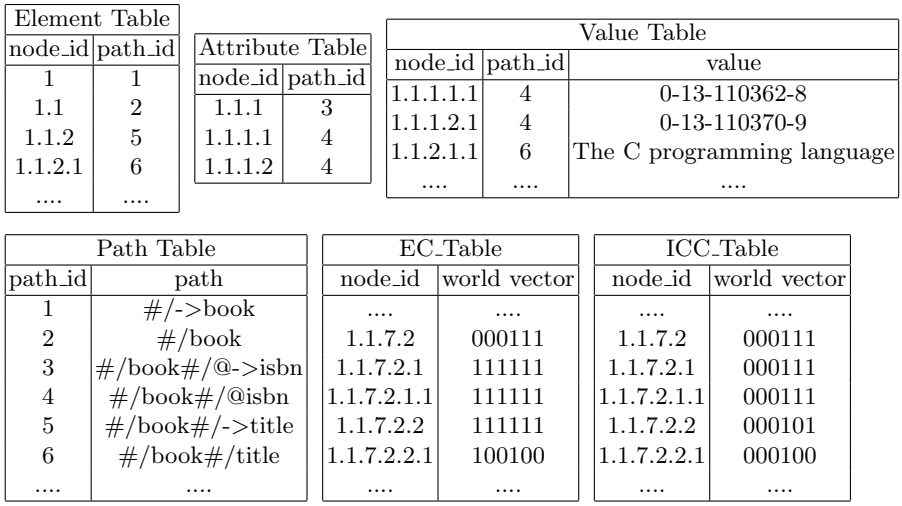
Fig. 2. Multidimensional Tree Pattern

5 Path-based Relational Schema

In a Path-based relational schema [18], the nodes of a XML document are able to be stored in different tables in the relational database, according to their type (elements, attributes and value nodes). Furthermore, there is a *Path Table* which stores all possible paths of the XML tree. Each path is identified by a path identification number through which the elements of the XML document are assigned to this path.

For MXML, we use similarly three tables in the relational schema to store different types of MXML nodes. So, we have an Element Table for storing element nodes, an Attribute Table for storing attribute nodes and a Value Table for storing value nodes. Each one of these tables contains the id of each node (node id) based on Dewey-labelling schema and the id (path id) representing the path in the Path Table, which leads to that node. For the Value Table, the value of the node is additionally stored. On the other hand, the Path Table contains all possible paths of the MXML document starting from the root node, assigning to each path a unique id (path id). As we show in the following example, the paths stored in the Path Table are path expressions with two more additional characteristics and we will call them *simple path expressions*. The first one is the "$->$" notation used in MXPath for indicating multidimensional nodes and the second one is the character "`#`", which is added before every "`/`" separator for helping through the MXPath to SQL query conversion. Also, there are two more tables in the relational schema, the explicit context (EC) table and the inherited context coverage (ICC) table, which are used for storing MXML nodes' context (explicit context or inherited context coverage respectively) in a binary-based format (world vectors).

Example 5. Fig. 3, depicts (parts of) the tables contained in the Path-based relational schema, for storing the MXML document represented in the MXML-tree of Fig. 1. Notice the representation of context for each node in EC Table and ICC Table, using the world vector representation [7]. This representation uses



Element Table

node_id	path_id
1	1
1.1	2
1.1.2	5
1.1.2.1	6
....	....

Attribute Table

node_id	path_id
1.1.1	3
1.1.1.1	4
1.1.1.2	4

Value Table

node_id	path_id	value
1.1.1.1.1	4	0-13-110362-8
1.1.1.2.1	4	0-13-110370-9
1.1.2.1.1	6	The C programming language
....	....	....

Path Table

path_id	path
1	#/->book
2	#/book
3	#/book#/@->isbn
4	#/book#/@isbn
5	#/book#/->title
6	#/book#/title
....	....

EC_Table

node_id	world vector
....	....
1.1.7.2	000111
1.1.7.2.1	111111
1.1.7.2.1.1	111111
1.1.7.2.2	111111
1.1.7.2.2.1	100100
....	....

ICC_Table

node_id	world vector
....	....
1.1.7.2	000111
1.1.7.2.1	000111
1.1.7.2.1.1	000111
1.1.7.2.2	000101
1.1.7.2.2.1	000100
....	....

Fig. 3. The Path-based Relational Schema.

binary numbers to declare the absence (value 0) or presence (value 1) of specific worlds, forming the environment under the associated node exists. Moreover, we can see that the paths to any context or multidimensional element or attribute are stored in the Path Table, with the addition of the spacial character "`#`" before every "`/`" separator. This is a technique, also used by XRel [18], in order to solve the problem of converting XML queries containing "`//`" separators to SQL queries, using appropriate SQL conditions (for example the query "`/book//price`" could be replaced by the string "`#/book#%/price`" in a LIKE expression of the SQL query, in order to return the "`price`" nodes which are descendants of node "`book`", according to their assigned paths).

6 Converting MXML queries to SQL queries

For the conversion of MXML queries to SQL queries, an appropriate algorithm is used. According to this algorithm, an MXPath expression is divided into subpaths (segments), based on the *segmentation rules* mentioned bellow. Based on the derived segments of the initial query, the appropriate table instances are chosen in order to be used in the FROM part of the final SQL query. Also, predicates containing context qualifiers, are processed based on the ordered-based representation schema of context, so as to be included as conditions in the WHERE part of the SQL query. Finally, we should notice that this algorithm is recursive in order to cover cases including branches in the MXPath query.

6.1 MXPath segmentation

Considering MXPath query Q as a tree pattern G, we can divide G in subpaths (segments) according to the segmentation rules defined by the following

definitions. Each rule contains a segmentation point, according to which G is divided into segments:

Definition 1. *Given a tree pattern G and assuming as a segmentation point S_b the root node of a branch, we define the Branch Segmentation Rule, which divides G in three pieces (sub paths) at every S_b point. Traversing G from the top to the bottom, the first piece contains the path up to the root node of the branch (S_b), the second one contains the rest of the path in the selection path and the third one contains the rest of the branch G_b.*

Definition 2. *Given a tree pattern G and assuming as a segmentation point S_c a node indexed by a context specifier, we define the Context Segmentation Rule, which divides G in two pieces (sub paths) at every S_c point. Traversing G from the top to the bottom, the first piece contains the path up to the indexed node (S_c) and the second one contains the rest of the path.*

Definition 3. *Given a tree pattern G and assuming as a segmentation point S_w the parent node of a "*" wildcard, we define the * Segmentation Rule, which divides G in two pieces (sub paths) at every S_w point. Traversing G from the top to the bottom, the first piece contains the path up to the parent node of the "*" node (S_w) and the second one contains the rest of the path.*

Definition 4. *Given a tree pattern G and assuming as a segmentation point S_v the parent node of a value node (linked with a dotted edge), we define the Value Segmentation Rule, which divides G in two pieces at every S_v point. Traversing G from the top to the bottom, the first piece contains the path up to the parent node of the value node (S_v) and the second one contains the value node itself.*

Note that during the query conversion procedure, all the above segmentation rules are applied again for every branch part G_b.

6.2 Convertion algorithm

At this point, we describe the algorithm, which is used for the conversion of a MXPath query over a MXML document to a SQL query over the corresponding Relational Schema. In order to apply this algorithm we should first make the following assumptions.

Assumptions:
- The MXPath queries can contain branches, context (/) or multidimensional (/− >) child edges, descendent edges (//), context qualifiers, "*" wildcards and value nodes.
- Tables ICC_Table and EC_Table contain binary world vectors as decimal numbers.

For better understanding, we provide the following example, where we show how a MXPath expression can be converted to a SQL query over the relational schema of Figure 3, according to the conversion algorithm.

Example 6. Lets consider the following MXPath query.
MXPath:

```
/book[authors[author="Brian W.K."]]/cover[ec()="ed=gr"]
/->material
```

In Fig. 4, we can see the multidimensional tree pattern expressing the query of Example 6.

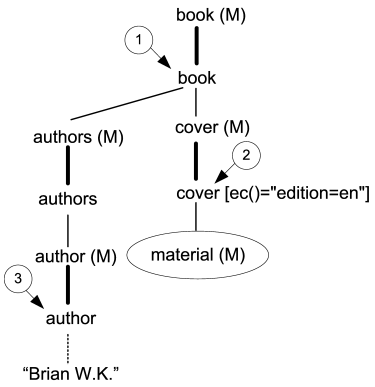


Fig. 4. Multidimensional Tree Pattern

Initially, the algorithm divides the processed MXPath query of Example 6 expressed by the corresponding multidimensional tree pattern into segments (*sub_path_1*, *sub_path_2*, ..., *sub_path_n*, *VALUE*), as it is shown in Example 7, according to the segmentation rules mentioned in Section 6.1.

Example 7. Executing the segmentation rules for the above query, we divide the tree pattern of Fig. 4 in three segmentation points indexed by labelled "1", "2" and "3" arrows (*Branch Segmentation*, *Context Segmentation* and *Value Segmentation* respectively). So, we produce the following parts:

sub_path_1: /book
sub_path_2: cover
sub_path_3: ->material
sub_path_4: authors/author
VALUE: "Brian W.K."

Note the *sub_path_4*: $authors/author$ and *VALUE*: "$BrianW.K.$", which represent the segments of a branch.

The ordered sub paths (*sub_path_1*, *sub_path_2* ...) mentioned in Example 7 are used from the algorithm to construct the corresponding simple path expressions. As we can see in Example 8, those expressions are constructed by concatenating

every next sub path with its previous one. Also, a prefix *Q_pref* representing the path from the root of the tree to the root of the processed branch and the appropriate "#" characters are added. When the algorithm is called for first time, *Q_pref* is initially set to NULL, as during that time the processed branch is the whole tree.

Example 8. Using the sub-paths of Example 7 we make the corresponding simple path expressions:

Path_expr_1': #/book
Path_expr_2': #/book#/cover
Path_expr_3': #/book#/cover#/->material
Path_expr_4': #/book#/authors#/author

When the simple path expressions are produced, the construction of the SQL query is started. During the construction of the FROM part, the algorithm uses a table instance of Element Table or Attribute Table and a table instance of Path Table for each segment (sub-path), except the case where a "*" wildcard exists. In that case, an extra table instance of Element Table or Attribute Table is required to represent any possible node may be represented by the "*". Also, according to the existence of context qualifiers or value nodes, table instances of ICC Table, EC Table or Value Table are used.

For our example query of Example 6, we show in Example 9 the section of the SQL query which contains the FROM part.

Example 9. Constructing the FROM part of the SQL query:

```
SELECT a_3.node_id FROM Element_Table a_1, Path_Table p_1,
Element_Table a_2, Path_Table p_2, EC_Table ec_2, Element_Table
a_3, Path_Table p_3 ...
```

In the WHERE part, the appropriate conditions are added. For each path expression, there is a condition checking the existence of a node in the Path Table, using the LIKE operator. If there is a ec predicate accompanying a path expression, the algorithm adds a condition comparing the set of worlds expressed by the predicate's context specifier, against the world vectors stored in the EC Table, based on the predicate's condition operator. If there is one or more branches, the condition which the algorithm adds to the WHERE part, contains the execution of the algorithm for each branch.

In Example 10, we show the SQL query which is produced by the execution of the conversion algorithm, containing the WHERE part.

Example 10. Constructing the WHERE part of the SQL query:

```
SELECT a_3.node_id FROM Element_Table a_1, Path_Table p_1,
Element_Table a_2, Path_Table p_2, EC_Table ec_2, Element_Table
a_3, Path_Table p_3 WHERE p_1.path Like '#/book' and
    a_1.path_id = p_1.path_id and
    a_1.node_id = ANY (
```

Evaluate_branch()

```
) and
    p_2.path Like '#/book#/cover' and
    a_2.path_id = p_2.path_id and
    a_2.node_id = ec_2.node_id and
    ec_2.world_vector = 7 and
    a_2.node_id Like CONCAT(a_1.node_id, '%')
    p_3.path Like '#/book#/cover#/->material' and
    a_3.path_id = p_3.path_id and
    a_3.node_id Like CONCAT(a_2.node_id, '%')
```

Note that the function call *Evaluate_branch()* represents the execution of the algorithm for the branch of Figure 4. After that execution, the produced SQL code will be placed where the function call occured.

At this point, we should note that for each two consecutive Element Table instances, algorithm adds a SQL condition based on dewey indexing in order to guaranty the ancestor-descendant relationship among these instances. This condition is expressed with the "%" operator in a SQL LIKE condition, involving the the identification numbers of the two instances.

In the case where there is a "*" wildcard in a sub path, the "%" operator is used in the SQL LIKE condition in order to express the "*". As we said before, in that case, we use an intermediate table instance and we check the difference between the levels of that instance and the current instance, using the dewey based identification numbers of the two instances.

Whether a path expression contains a value, a condition involving the instance of Value Table is included in the SQL query. On the other hand, if a icc predicate exists for the processed query, the algorithm works in the same manner with ec predicates, but this time using the ICC Table through the relevant table instance.

Going back to Example 10, the algorithm is executed through function *Evaluate_branch()* in order to process the branch part of the query. Generally, for each branch execution the depth number (*Level* variable) of the starting point of each branch is taken into account. This depth number is used with the SQL expression "SELECT SUBSTRING_INDEX", which returns a substring from a dewey index before *Level* occurrences of the dot delimiter. In that way, it is guaranteed through the dewey indexing that the processed branch is under the correct node.

After the branch execution of the conversion algorithm we have the SQL code shown in Example 11.

Example 11. Conversion algorithm's branch execution result:

```
SELECT SUBSTRING_INDEX (a_1.node_id, '.', 2) FROM Element_Table
    a_1, Path_Table p_1, Value_Table v
WHERE p_1.path Like
    '#/book#/authors#/author' and a_1.path_id = p_1.path_id and
    v.path_id = p_1.path_id and v.value = 'Brian W.K.' and
    v.node_id Like CONCAT(a_1.node_id, '%')
```

Inserting SQL code produced by the algorithm's branch execution (Example 11) into the code of Example 10, at the place where the function call *Evaluate_branch()* occurred, we have the final SQL query presented in Example 12.

Example 12. The whole SQL query:

```
SELECT a_3.node_id FROM Element_Table a_1, Path_Table p_1,
Element_Table a_2, Path_Table p_2, EC_Table ec_2, Element_Table
a_3, Path_Table p_3 WHERE p_1.path Like '#/book' and
    a_1.path_id = p_1.path_id and
    a_1.node_id = ANY (
    SELECT SUBSTRING_INDEX (a_1.node_id, '.', 2) FROM Element_Table
       a_1, Path_Table p_1, Value_Table v
    WHERE p_1.path Like
       '#/book#/authors#/author' and a_1.path_id = p_1.path_id and
        v.path_id = p_1.path_id and v.value = 'Brian W.K.' and
        v.node_id Like CONCAT(a_1.node_id, '%')
    ) and
    p_2.path Like '#/book#/cover' and
    a_2.path_id = p_2.path_id and
    a_2.node_id = ec_2.node_id and
    ec_2.world_vector = 7 and
    a_2.node_id Like CONCAT(a_1.node_id, '%')
    p_3.path Like '#/book#/cover#/->material' and
    a_3.path_id = p_3.path_id and
    a_3.node_id Like CONCAT(a_2.node_id, '%')
```

The result of the above SQL query, is the multidimensional node with identification number (node_id) "1.1.7.2.1".

7 Discussion and motivation for future work

In this work, we presented a path-based technique for storing MXML documents in relational databases. Also, we described how we can express MXPath queries using multidimensional tree patterns and how we can convert those queries into SQL queries through a conversion algorithm, which is based on a tree pattern segmentation method. Finally, we explained how this algorithm works in order to convert MXML queries to SQL queries, using an appropriate example. Future work will focus on experimental evaluation of the performance for the path-based storage technique and the conversion algorithm.

References

1. T. Amagasa, M. Yoshikawa, and S. Uemura. A Data Model for Temporal XML Documents. In *Database and Expert Systems Applications, 11th International Conference, DEXA 2000, London, UK, September 4-8, 2000, Proceedings*, volume 1873 of *Lecture Notes in Computer Science*, pages 334–344. Springer, 2000.

2. T. Amagasa, M. Yoshikawa, and S. Uemura. Realizing Temporal XML Repositories using Temporal Relational Databases. In *Proceedings of the Third International Symposium 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 *Proceedings of the 18th International Conference on Data Engineering, 26 February - 1 March 2002, San Jose, CA*, pages 64–75. IEEE Computer Society, 2002.
4. A. Deutsch, M. F. Fernandez, and D. Suciu. Storing Semistructured Data with STORED. In *SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1-3, 1999, Philadelphia, Pennsylvania, USA*, pages 431–442. ACM Press, 1999.
5. F. Du, S. Amer-Yahia, and J. Freire. ShreX: Managing XML Documents in Relational Databases. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases, Toronto, Canada, August 31 - September 3 2004*, pages 1297–1300. Morgan Kaufmann, 2004.
6. D. Florescu and D. Kossmann. Storing and Querying XML Data using an RDMBS. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 22(3):27–34, 1999.
7. Nikolaos Fousteris, Manolis Gergatsoulis, and Yannis Stavrakas. MXML Storage and the Problem of Manipulation of Context. In *First Workshop on Digital Information Management, 30-31 March, 2011, Corfu, Greece*, pages 45–60, 2011.
8. Nikolaos Fousteris, Yannis Stavrakas, and Manolis Gergatsoulis. Multidimensional XPath. In *iiWAS'2008 - The Tenth International Conference on Information Integration and Web-based Applications Services, 24-26 November 2008, Linz, Austria*, pages 162–169, 2008.
9. M. Gergatsoulis, Y. Stavrakas, and D. Karteris. Incorporating Dimensions in XML and DTD. In *Database and Expert Systems Applications, 12th International Conference, DEXA 2001 Munich, Germany, September 3-5, 2001, Proceedings*, volume 2113 of *Lecture Notes in Computer Science*, pages 646–656. Springer, 2001.
10. M. Gergatsoulis, Y. Stavrakas, D. Karteris, A. Mouzaki, and D. Sterpis. A Web-Based System for Handling Multidimensional Information through MXML. In *Advances in Databases and Information Systems, 5th East European Conference, ADBIS 2001, Vilnius, Lithuania, September 25-28, 2001, Proceedings*, volume 2151 of *Lecture Notes in Computer Science*, pages 352–365. Springer, 2001.
11. M. Ramanath, J. Freire, J. R. Haritsa, and P. Roy. Searching for Efficient XML-to-Relational Mappings. In *First International XML Database Symposium, XSym 2003, Berlin, Germany, September 8, 2003, Proceedings*, Lecture Notes in Computer Science, pages 19–36. Springer, 2003.
12. 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.
13. 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 *VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases, September 7-10, 1999, Edinburgh, Scotland, UK*, pages 302–314. Morgan Kaufmann, 1999.
14. Yannis Stavrakas and Manolis Gergatsoulis. Multidimensional Semistructured Data: Representing Context-Dependent Information on the Web. In *Advanced Information Systems Engineering, 14th International Conference, CAiSE 2002,*

Toronto, Canada, May 27-31, 2002, Proceedings, volume 2348, pages 183–199, 2002.

15. 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 *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, Madison, Wisconsin, June 3-6, 2002*, pages 204–215. ACM, 2002.
16. 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.
17. F. Wang, X. Zhou, and C. Zaniolo. Using XML to Build Efficient Transaction-Time Temporal Database Systems on Relational Databases. In *Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA*, page 131. IEEE Computer Society, 2006.
18. 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.