Historical documents as monuments and as sources

Panos Constantopoulos, Martin Doerr, Maria Theodoridou, Manolis Tzobanakis

Institute of Computer Science
Foundation for Research and Technology – Hellas
{panos, martin, maria, tzoban}@ics.forth.gr

Abstract. The main functions of a digital library of historical documents are the digital recording of documents, transcriptions and translations of documents, subject indexing, annotation and retrieval. Using such a system, scholars can efficiently study the documents without involvement of the originals, thus ensuring the preservation of documents and the protection of researchers from exposure to potential health hazards. An important part of the study of historical documents consists of classifying the material and annotating it in such a way that retrieval is facilitated in the future.
We present the development of a historical document management system that supports both digital library functionality and archival management of the original documents. This system includes semantic indexing and multifaceted classification of historical documents with the use of a built-in thesaurus aimed at attaining satisfactory levels of efficiency of the classification process, completeness and precision of the retrieved information, and user-friendliness.

Key words: digital libraries, historical documents, subject classification

1 Introduction

Along with the expansion of computer systems, networks and of the Internet in particular, has come a rapid expansion of digital libraries. With the term digital libraries we refer to collections of documents which have been stored and are accessible electronically together with a set of relevant services. The management of historical documents which have significant cultural and historical value and may have undergone varying degrees of deterioration gains increasing interest in the area of digital libraries [1].

Historical documents can be viewed in two perspectives, that of monuments and that of source material. On one hand, viewing the documents as monuments, the objective is to support the creation of a digital archive that will ensure preservation of the originals and facilitate the management of the physical archive using consistent, appropriate archival catalogue entries. Such a digital library should provide the entire spectrum of functions for the various stages in the treatment of historical documents: document acquisition, archival and management, cataloguing and annotation, processing and transmission, like in museum information systems. These functions serve a variety of purposes, such as supporting the preservation, documentation, study and management of historical documents. They also protect people from exposure to potential health hazards and assist in the production and dissemination of electronic versions of publications and exhibits, which promote cultural education.

On the other hand, we can view documents as source material and as such we need to manage content descriptions, transcriptions, translations and provide thematic indexing. An important part of the study of historical manuscripts consists of classifying the material and annotating it in such a way that retrieval is facilitated in the future.

The documentation and management of source material and monuments are commensurate and of the same importance. In an information management approach the two views are both notionally and functionally interconnected. In this spirit, we developed a historical document management system that comprises a digital library of historical documents and supports the main functions of recording digital historical documents, transcriptions and translations of documents, subject indexing, annotation and retrieval. Using such a system, scholars can efficiently use and study the documents without involvement of the originals, thus ensuring the preservation of documents and the protection of researchers from exposure to potential health hazards. The system supports semantic indexing and multifaceted classification of historical documents with the use of a built-in thesaurus. A main feature of the system is its rich and extensible document model. Our aim during the design and development of the system was to attain satisfactory levels of efficiency of the classification process, completeness and precision of the retrieved information, and user-friendliness. The system supports remote access to the digital library through the Web, allowing users to work in a familiar way, using their own preferred environment independently of their platform. The implementation of the system is based on Java for the client side and on the Semantic Index System [2] for the server side.

Two important collections with significant cultural and historical value have provided material for the development of the historical document and archive

management system. In the context of project "ARCHON - A Multimedia System for Archival, Annotation and Retrieval of Historical Documents" we investigated the classification and archival of the Turkish Archive of Heraklion, the Municipal Archive of Heraklion and the Venetian Archive which comprise the historical archives of the Vikelea Municipal Library of Heraklion, dated from the late 1600s to early 1900s. The other project concerned the Turkish Archive of Chania.

2 Subject classification

An important part of the study of historical documents consists of classifying the material and annotating it in such a way that retrieval is facilitated. The state-of-the-art OCR software is inaccurate on all but the most uniformly printed documents (let alone manuscripts) requiring proofreading and error correction. Thus, the automatic transcription of historical manuscripts is not possible and we can only rely on manual techniques. An obvious approach to the document classification problem would be the implementation of a keyword system. Manual keyword assignment is a time consuming process and historians, scholars and other researchers that study the documents are not willing to spend too much time to input and classify the documents. Moreover, keywords are not necessarily unique and we may easily end up using keywords from at least three different vocabularies: the vocabulary of the author(s) of the documents, the vocabulary of the cataloguer, indexer, or classifier of the document and the vocabulary of the searcher. As these vocabularies have evolved over different time periods, it is generally not easy to create satisfactory mappings between them. In addition the risk of mismatch in transitions from one vocabulary to another is quite high. An approach to this problem is to identify concepts, rather than words, in a given knowledge domain, which, organized in term thesauri, can provide consistent classification, better retrieval precision, preservation of the identity of concepts and a domain knowledge base.

Historical documents may be classified through different procedures. One scenario is that an indexer is assigned to classify a specific corpus of material. A second scenario is that a researcher studying a specific topic assigns classification information to the documents that he encounters during his study. Scholars classify documents according to their interests of study and we can never assume that the classification of a document or of an entire archive is complete. We essentially have to deal with an open world. Thus, it is important to provide an easy to use, straightforward and efficient user interface that will enable precise and fast classification of the documents. The process of classification itself should be intuitive and the system should be able to fill in automatically information that is stored already in or can be derived from the knowledge base of the system. Additionally, the system should follow the context and the user profile and offer automatically the most probable options.

3 Building a historical document classification and management system

The archival and management of documents in the digital library are driven by a model of documents, archive organization and processes, which take into account the ISAD (G) General International Standard Archival Description [6], the EAD Encoded Archival Description Document Type Definition [7] and the Dublin Core Metadata Elements Set [8].

The model distinguishes between the historical (original) and the current organization of the material and keeps the correlations between the two organizations (Figure 1). This distinction allows an integrated view of material that is copied or scattered in different physical locations. For example, the Venetian archives are physically located in Italy, while the Vikelea Municipal Library maintains microfilm copies of the same material.

The organization of the archives comprises Fonds, Subfonds, Units of description (series, books, files, pages, sheets etc.), according to the ISAD (G) terminology, and we distinguish between the physical, the conceptual and the electronic material providing the correlations among them (Figure 2). For example, a page is a physical unit that may contain more than one documents (conceptual units). Finally, we have modeled actions affecting the material such as editing or scanning and states during material processing that will facilitate the streamlining of some parts of the document input, classification and retrieval user interface (Figure 3).

Subject classification of historical documents is based on five facets which represent distinct concept categories related to information *about*, as well as information contained *in* the documents.. Important elements about or in a document are actors, dates, places, purposes, objects and their names. There exists significant information regarding document creation, such as who wrote the document, to whom it was addressed, when and where it was written, for what purpose it was written, what it quotes. Moreover, scholars are interested in information quoted in the document. This includes significant actions or activities mentioned in the document such as what activity is described, who is involved, where and when it took place, what objects were involved. In other words the questions that have to be answered are "Who?", "Where?", "When?", "What?" and "How?".

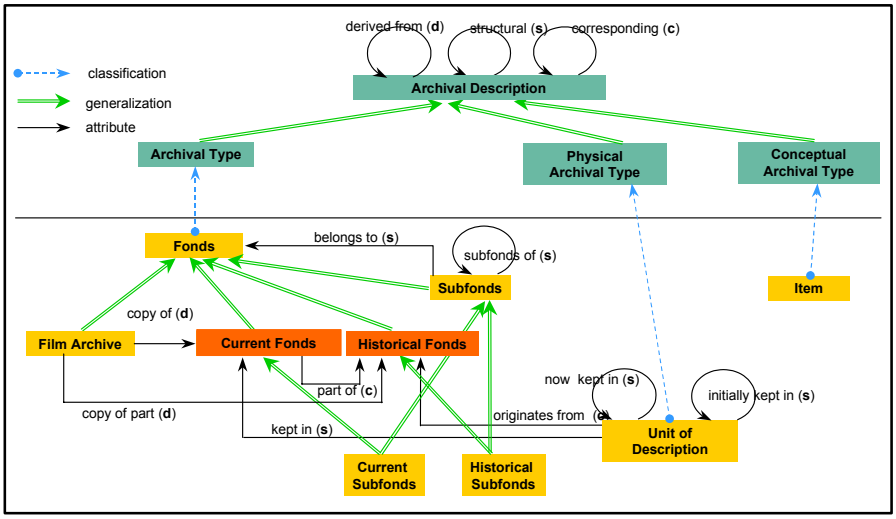


Figure 1: Modeling collections of historical documents

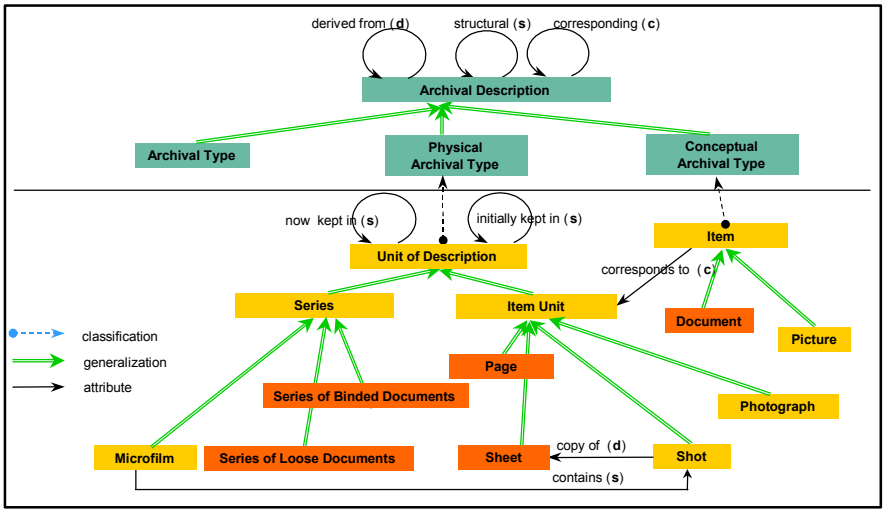


Figure 2: Modeling objects versus contents

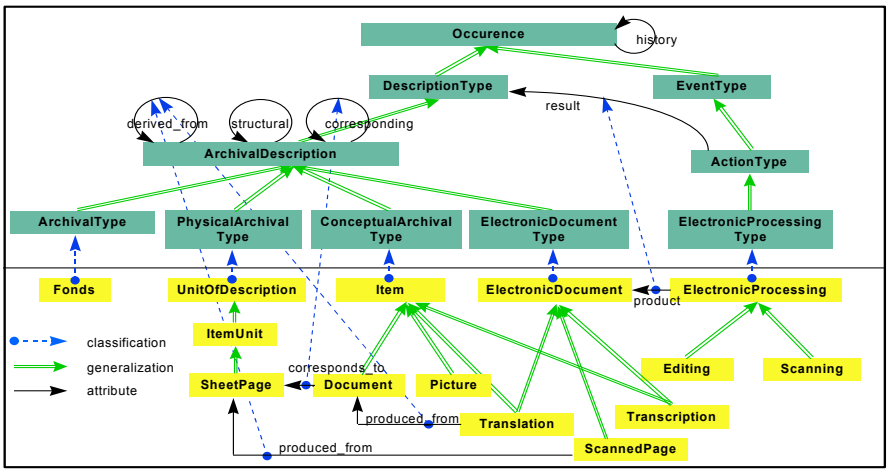


Figure 3: Modeling the electronic documents

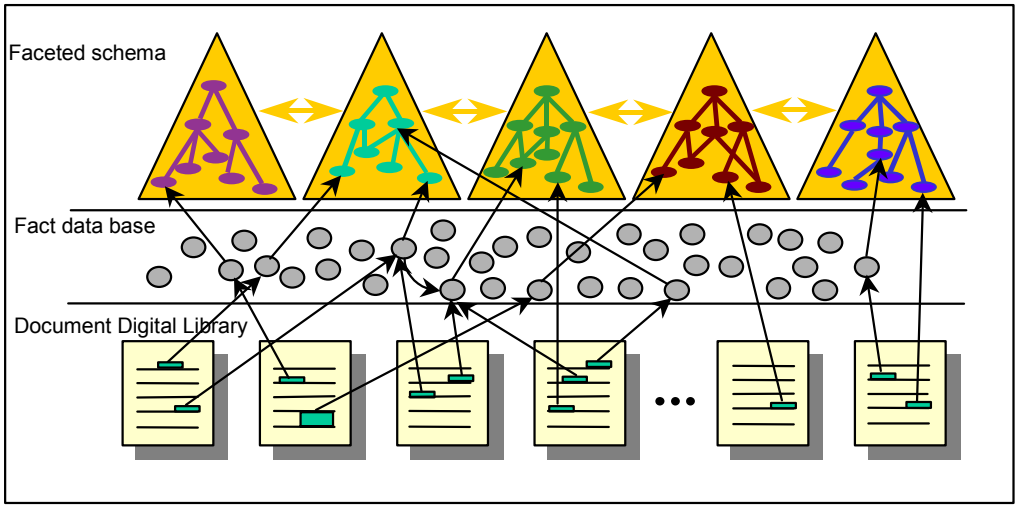


Figure 4: Concept -based faceted classification of historical

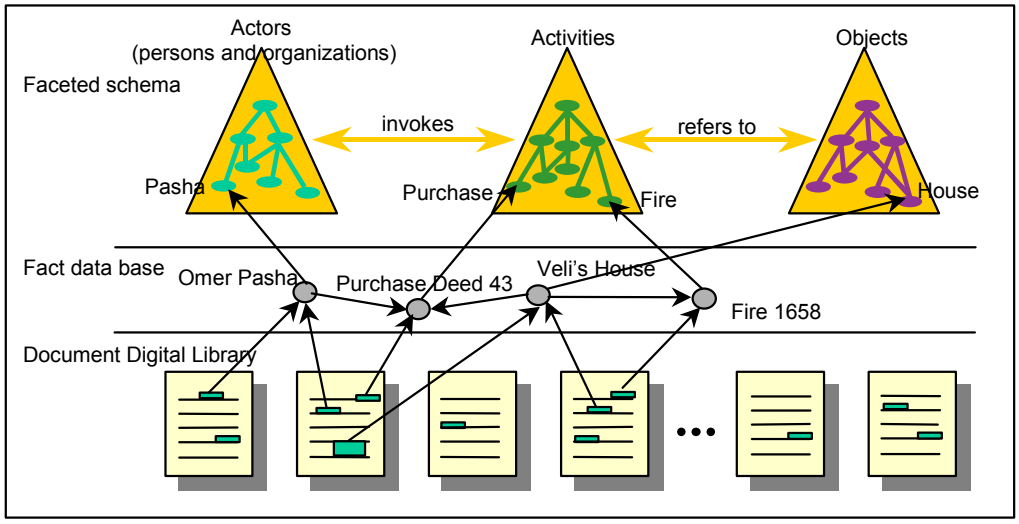


Figure 5: Concept-based faceted classification of historical documents: An

It is common practice by scholars to assign an annotation to documents while they study or translate documents. In our approach, we base classification on the analysis of the original (prototype) text or of its annotation, description or translation. We provide a formal syntactic structure to assign a combination of terms from different facets for each document, which result in a precise characterization of the document according to the set of criteria represented by the facets. The facets constitute an extensible structure of hierarchical catalogues. In contrast with keyword systems, the index terms are typed by their facet and interrelated, forming a semantic network with unique identity and associated meaning thus avoiding ambiguities (Figure 4). Hierarchical structures and especially polyhierarchies, as in our case, facilitate fast search and recognition. Classification takes place at type/class level or at "instance" level where we have references to correlated real things. Simultaneous classification from independent aspects is also possible. The system builds up knowledge about the domain, which may be helpful during retrieval. The maintenance of such a semantic net can be assisted by a full text indexing facility. In Figure 5, we present an example of document classification and the knowledge building process: A document refers to the purchase of Veli's house by Omer Pasha. Another document refers to a fire that took place in 1658 during which Veli's house was burned. A scholar interested in documents concerning Omer Pasha might consider the document about the fire as relevant. Our system has built a link between the two

documents and thus a query on Omer Pasha would retrieve both, although there is no direct reference of Omer Pasha in the second document neither has it been classified having Omer Pasha as actor.

We now briefly present the five facets defined in the system:

Persons and Organizations

This facet groups the agents - either individuals, groups of people or organizations - that are involved in the creation of the document or referred in it. The facet is organized according to types of agents i.e. professions, social/religious casts, administrative roles, origin etc. For efficient retrieval based on names we keep a separate index of names with references to documents and additional information such as name of father, mother, origin, date.

The identification of a person may be difficult in several cases. We make the assumption that two persons are different unless proven differently.

Activities and actions

This facet groups the purposes and kinds of documents. For activities and actions it is better to do the classification according to their types and not the instances themselves. We investigate the possibility of using an existing thesaurus, such as SHIC [5], for this purpose.

Places

This facet groups the places referred in a document or the place it was created. Types under this facet include natural division (e.g. mountain, lake, river, valley), administrative division (e.g. prefecture, city, village) or buildings (e.g. monastery, church).

An interesting issue concerning this facet is how to support associative search, for example how to identify "Martin's house".

Time

This facet groups the chronological references made in or attributed to documents.

In the case of dates recall is often considered more important than precision in document retrieval, thus an interesting issue is how to provide an "intelligent" browser based on a dynamic temporal index for events and a "historical clock".

Objects

This facet groups the objects referenced in documents. Types include movable/fixed objects, monetary systems and payment transactions. In several cases an object may be identified through an activity or action.

4 Example of use

In this section we will give a brief presentation of the historical document management system that we developed. The system supports semantic indexing and multifaceted classification of historical documents with the use of a built-in thesaurus. The subject classification functions are built using the Semantic Index System [2], a general purpose semantic network information management system, developed by the Information Systems Laboratory of ICS-FORTH. Document querying and retrieval are done through a Java/HTML User Interface, which communicates with the SIS through a Java Database Connectivity Driver (JDBC) [3].

Two important collections with significant cultural and historical value have provided test material for the development of the historical document and archive management system. The first consists of the Turkish Archive of Heraklion, the Municipal Archive of Heraklion and the Venetian Archive which comprise the historical archives of the Vikelea Municipal Library of Heraklion, dated from the late 1600s to early 1900s. The second consists of the Turkish Archive of Chania.

The three above-mentioned historical archives of the Vikelea Municipal Library comprise approximately 1,500,000 pages of manuscripts. As a test for the digital archive of our historical document management system, we proceeded with the digitization of about 80,000 pages of the archives. These documents were scanned to an electronic form, processed for image correction [4] and archived in our digital library. The users access the digital library through client programs that have been developed using Java and WWW technology. Java allows the same code to be executed on different platforms. To ease installation and maintenance we have built client programs to execute as Java applets inside standard Internet browsers (Netscape, Internet Explorer). To represent the information on the user's monitor a graphical user interface similar to the well-known Microsoft Windows Explorer GUI was created.

The user interface caters for both the monument and the source nature of documents during retrieval. The first is through the use of the archival catalogue. The archival catalogue entries used by the library to identify the documents in their physical location are preserved in the digital library and can be used for fast retrieval of documents by those users who are familiar with the physical organization of the archives.

The second and most interesting way of retrieving documents is by formulating document queries regarding the document content. Query formulation is achieved by selecting and combining terms from the five facets that have been defined. The terms of each facet are displayed

to the user as expandable tree structures similar to the trees of Microsoft Windows Explorer (Figure 6).

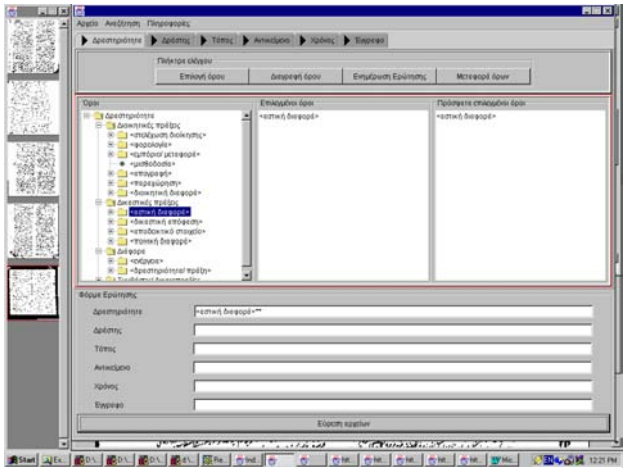
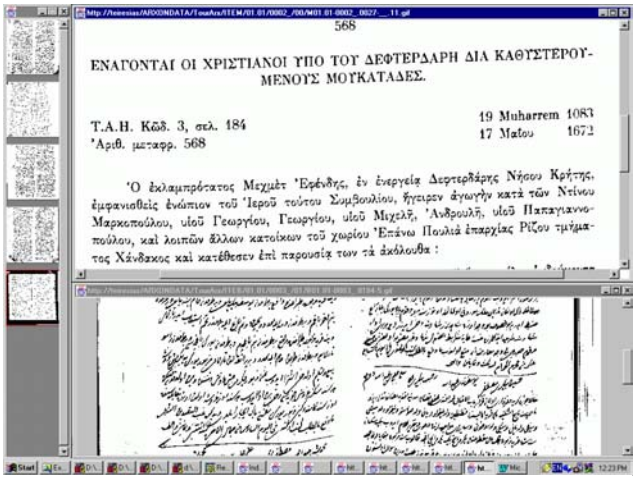


Figure 6: Document Management System – Query Formulation

For each facet the user may select one or more terms which are automatically combined with the AND operator. The query is thus formulated in steps and the user has an overall view of his selections at any point of the formulation process. The result of the query execution appears as a set of thumbnails that represent the digitized documents that matched the query. By clicking on the thumbnails the user can retrieve the image of the original document as well as its translation, if one is available (Figure 7). Additionally, the user may ask to view the classification information of a retrieved document, which might be useful in selecting new appropriate terms for query reformulation.


568

ΕΝΑΓΟΝΤΑΙ ΟΙ ΧΡΙΣΤΙΑΝΟΙ ΥΠΟ ΤΟΥ ΔΕΦΤΕΡΔΑΡΗ ΔΙΑ ΚΑΘΥΣΤΕΡΟΥΜΕΝΟΥΣ ΜΟΥΚΑΤΑΔΕΣ.

Τ.Α.Η. Κώδ. 3, σελ. 184
Ἀριθ. μεταφρ. 568

19 Muharrem 1083
17 Μαΐου 1672

Ὁ ἐκλαμπρότατος Μεχμὲτ Ἐφένδης, ἐν ἐνεργείᾳ Δεφτερδάρης Νήσου Κρήτης, ἐμφανισθεὶς ἐνώπιον τοῦ Ἱεροῦ τούτου Συμβουλίου, ἤγειρεν ἀγωγὴν κατὰ τῶν Ντίνου Μαρκοπούλου, υἱοῦ Γεωργίου, Γεωργίου, υἱοῦ Μιχελῆ, Ἀνδρουλῆ, υἱοῦ Παπαγιαννοπούλου, καὶ λοιπῶν ἄλλων κατοίκων τοῦ χωρίου Ἐπάνω Πουλιὰ ἐπαρχίας Ρίζου τιμάριτος Χάνδακος καὶ κατέθεσεν ἐπὶ παρουσίᾳ των τὰ ἀκόλουθα :

Figure 7: Document Management System - Retrieval

5 Conclusions

We have presented a management system for historical documents that supports semantic indexing and multifaceted classification of historical archives with the use of a built-in thesaurus. Documents are treated both as monuments, to be preserved and managed, and as sources of information content, collected, inter-linked and managed in a digital library. This system has been tested using archives of historical documents of the Vikelea Municipal Library of Heraklion and of the Turkish Archive of Chania.

Future work includes the development of an annotation system, the support for the development of specialized vocabularies per application, and the provision of Web accessibility.

References

[1] Digital Libraries: Future Research Directions for a European Research Programme, June 13-15, 2001, San Cassiano (Dolomites), Italy, http://delos-noe.iei.pi.cnr.it/activities/researchforum/Brainstorming/brainstorming-report.pdf

[2] http://www.ics.forth.gr/isl/r-d-activities/sis.html

[3] http://www.ics.forth.gr/isl/manuals/api.doc

[4] A Visual Tagging Technique for Annotating Large-Volume Multimedia Databases. K.V. Chandrinos, J. Immerkaer, Martin Doerr, P.E. Trahanias, 5th DELOS Workshop on Filtering and Collaborative Filtering

[5] http://www.holm.demon.co.uk/shic.htm

[6] ISAD (G) General International Standard Archival Description: http://www.ica.org/

[7] EAD Encoded Archival Description http://lcweb.loc.gov/ead/

[8] Dublin Core Metadata Elements Set http://dublincore.org