Classifying historical documents

Maria Theodoridou, Martin Doerr
Institute of Computer Science, Foundation for Research and Technology Hellas
Heraklion-Crete, Greece

Paper Proposal for "Applications of IT to Biblical Studies", Summer School, Sofia
25 July - 8 August 1998

Introduction

In this paper we propose a methodology for the classification of large volumes of historical documents. These ideas have been developed in the context of the ARCHON project - A Multimedia System for Archival, Annotation and Retrieval of Historical Documents - which is managed by ICS-FORTH and will be applied at the archive of the Vikelaia Municipal Library of Heraklion. The library posseses documents of remarkable cultural and historical value, dated from the late 1600s to early 1900s, which have undergone varying degrees of deterioration. The project will develop a state-of-the-art archive management system, which supports the main functions of recording digital images of all documents, subject indexing and retrieval and creation of descriptions, transcriptions and translations of documents. The multimedia system will allow efficient use and study of the documents by scholars without involvement of the prototypes.

The classification problem

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 state-of-the-art OCR software is inaccurate on all but the most uniformly printed documents (let alone manuscripts) requiring proof-reading 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. However, 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, since keywords are not necessarily unique, it is important to identify concepts and not keywords. Scholars will classify the documents according to their interests of study and we can never assume that the classification of a document or a whole archive is complete. We have to deal with an open world paradigm. However, if the system has knowledge about which parts of the archive have been classified and according to which facets, we may view parts of the archive as closed worlds and improve the retrieval of documents. Taking into account the above, it is crucial to implement a user interface that will be faster than usual keyword assignment and will provide a more precise and fast classification of the documents. The time needed by the scholar researcher to classify a document while studying it should not exceed 1 minute. The process of classification itself should be intuitive and the system should be able to fill automatically information already given which is derived from the system's built-in domain knowledge. Additionally, the system should follow the context and the user profile and present automatically to the user the most probable options.

Our approach

We have based the implementation of the ARCHON digital library on the Semantic Index System [1], a general purpose object-oriented semantic network database, developed by the Information Systems and Software Technology Division of ICS-FORTH. Document input, querying and retrieval will be done through a Java/HTML User Interface which communicates with the SIS through a Java Database Connectivity Driver (JDBC) [2].

Following the requirements of the users (Vikelea Municipal Library of Heraklion), we have modelled the collections of documents. 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 integral view of material that is copied or scattered in different physical locations. For example, the Venetian archives are in Italy while the Vikelea Municipal Library maintains microfilms of the same material. Additionally, we have modelled the organization of the archives into Fonds, Subfonds, Units of description (series, books, files, pages, sheets etc.) 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 modelled actions affecting the material such as editing or scanning and states during material processing that will facilitate the automation of some parts of the document input, classification and retrieval user interface (Figure 3).

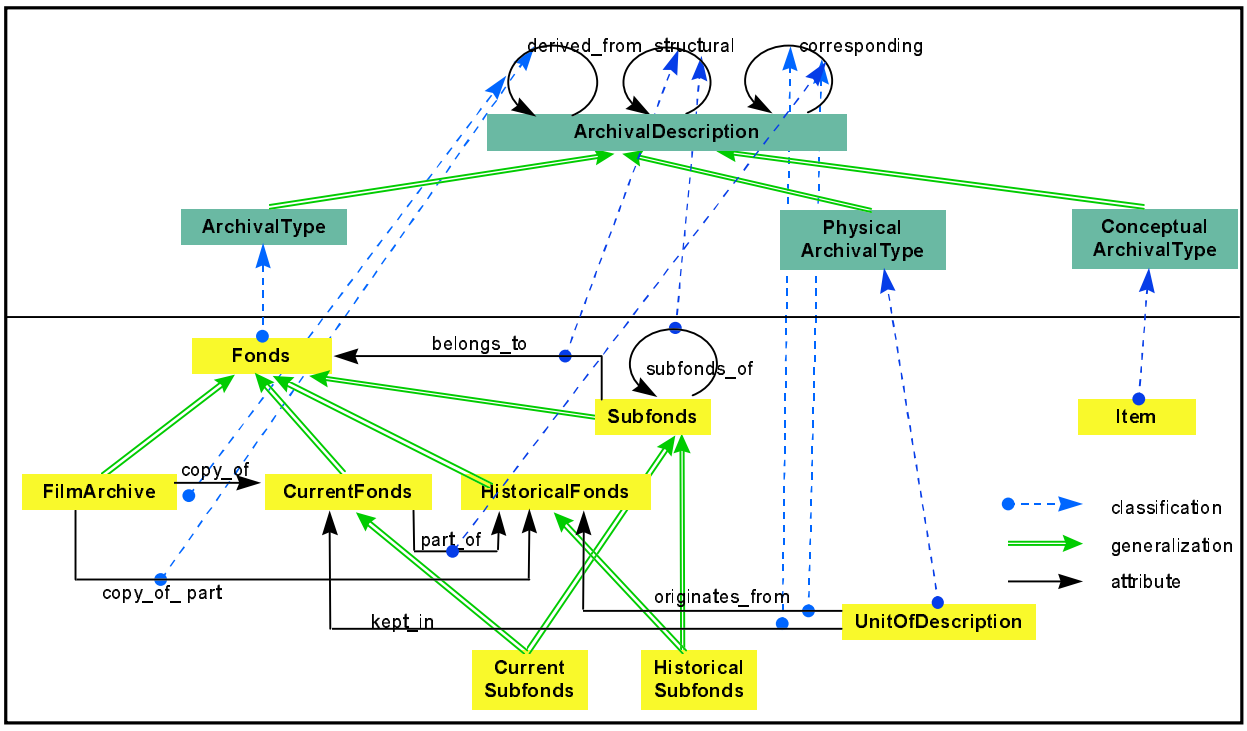

Figure 1: Model for collections of historical documents

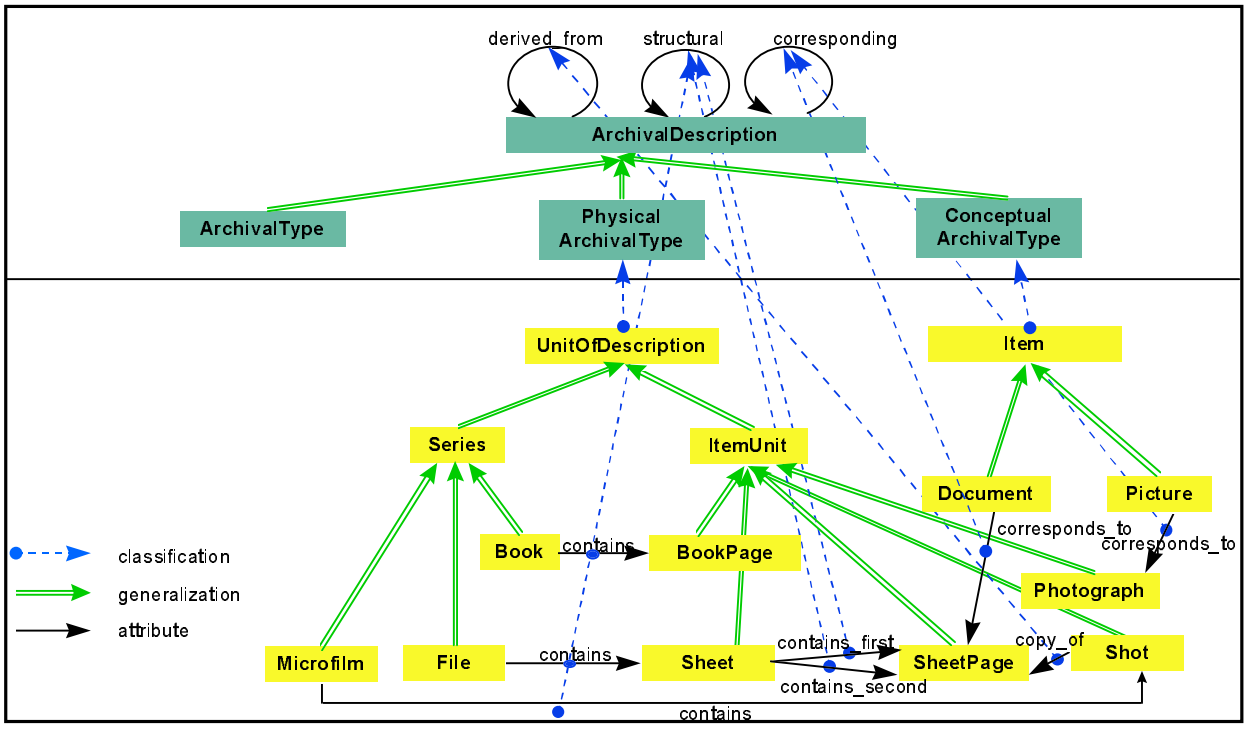

Figure 2: Model for collections of historical documents

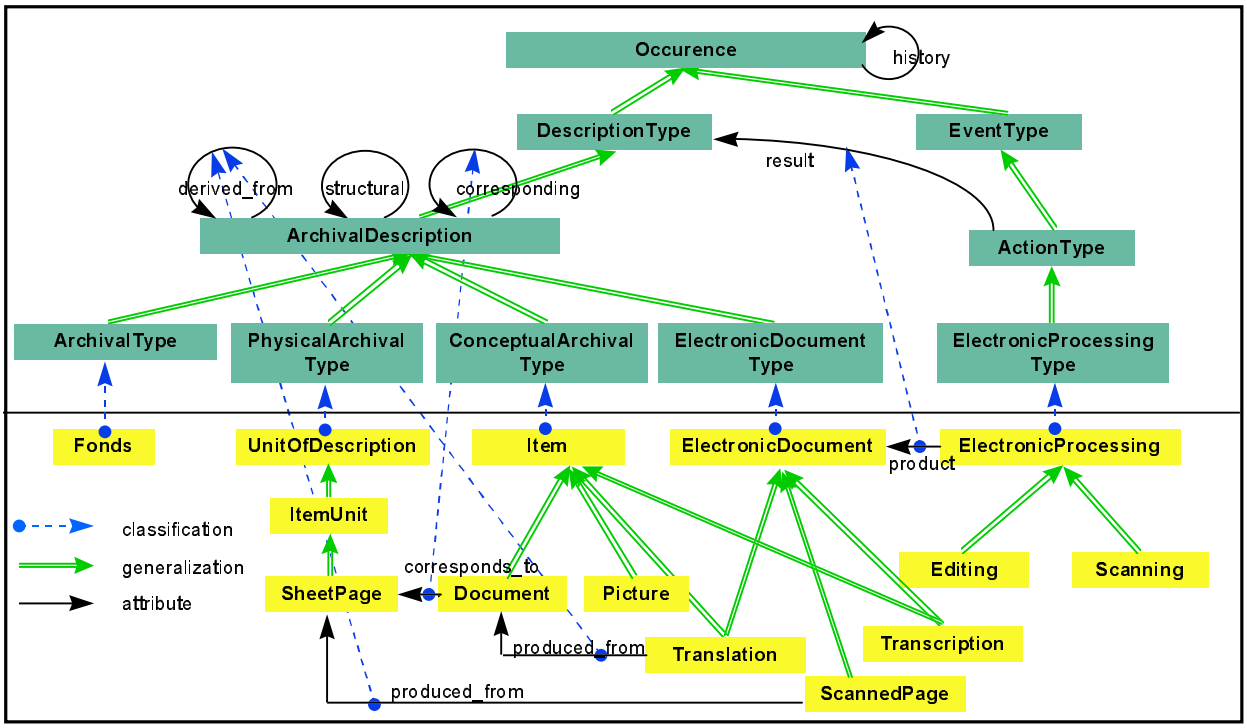

Figure 3: Model for collections of historical documents

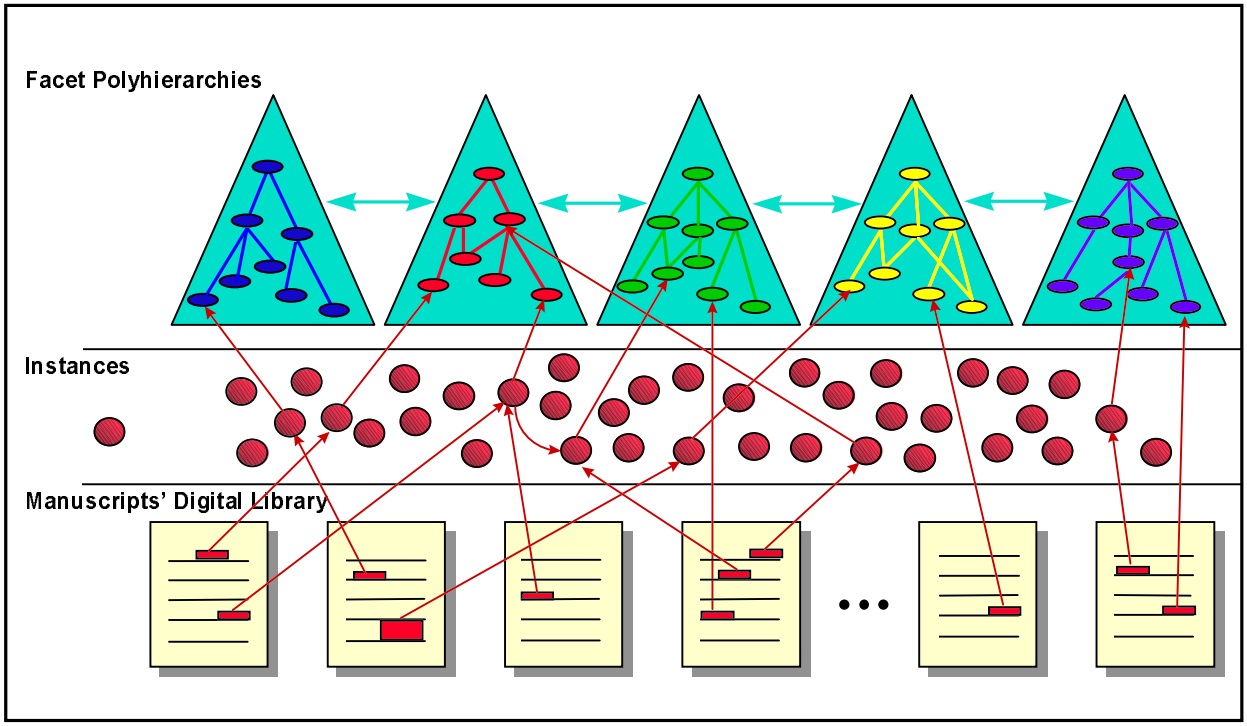

Figure 4: Faceted classification of historical documents

We base the classification of the historical manuscripts on five facets which represent distinct concepts used to group the actual things that appear in a manuscript. Important elements in a document are actors, dates, places, purposes, objects and their names. The questions of interest while studying a document are who wrote the document, to whom it was addressed, when and where it was written, for what purpose it was written, what it addresses. In other words the questions that have to be answered are "Who?", "Where?", "When?", "What?" and "How?". It is common practice by scholars to assign a title to documents while they study them - usually when they translate documents. We provide a formal syntactic structure to assign a combination of facets for each document which will result in an information similar to that in the document title but more detailed and precise. The facets constitute an extensible structure of hierarchical catalogues. In contrast to keyword systems, the index terms and names are typed by their facet and interrelated, forming a semantic net with unique identity and associated information so, ambiguities are not present. Hierarchical structures and specially polyhierarchies, as in our case, serve identity and facilitate fast search and recognition. The maintenance of such a semantic net can be assisted by a full text indexing facility. Classification may take place at type/class level or at "instance" level where we have references to correlated real things (Figure 4).

Taking into account the above we distinguish the following five facets:

[1] 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 any way by the document. The facet is further analyzed according to types of agents i.e. professions, social/religious casts, administrative roles, origin etc. For efficient, fast retrieval based on names we will 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 will make the initial assumption that two persons are different unless proven differently.

[2] 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 [3] for this purpose.

[3] Places

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

An interesting research issue concerning this facet is how to provide associative search, for example how to identify "Manolis' house".

[4] Time

This facet groups the chronologies referenced by the document.

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

[5] Objects

This facet groups the objects referenced by the document. Types include movable/unmovable objects, monatery systems and payment transactions. In several cases an object may be identified through an activity or action.

The classification will be done through a user interface (ImageTagger [2]) that will allow the user to mark-up parts of the representation of the document (areas on the document image) and associate them with the information in the semantic net.

Conclusions

We have designed a system for the classification of historical archives. The advantages of such a classification system are:

- It is faster than manual keyword assignment
- It provides high precision
- It maintains identity of concepts and not keywords

- It creates a base of domain knowledge
- It preserves the syntactic structure of the expression used for the classification

References

[1] http://www.ics.forth.gr/proj/isst/Systems/SIS

[2] 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

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