Historical research in archives: user methodology and supporting tools

Elena Torou, PhD, Researcher, Dept. of Informatics and Telecommunications, University of Athens, Greece

Akrivi Katifori, PhD, Researcher, Dept. of Informatics and Telecommunications, University of Athens, Greece

Costas Vassilakis, PhD, Assistant Professor, Dept. of Computer Science and Technology, University of Peloponnese, Greece

George Lepouras, PhD, Assistant Professor, Dept. of Computer Science and Technology, University of Peloponnese, Greece

Constantin Halatsis, PhD, Emeritus Professor, Dept. of Informatics and Telecommunications, University of Athens, Greece

ABSTRACT

Historic research involves finding, using and correlating information within primary and secondary sources, in order to communicate an understanding of past events. In this process, historians employ their scientific knowledge, experience and intuition to formulate queries (who was involved in an event, when did an event occur etc), and subsequently try to locate the pertinent information from their sources. In this paper, we investigate how historians formulate queries, which query terms are chosen, and how historians proceed in searching for related information in sources. The insight gained from this investigation can be subsequently used for organizing documents within historical source repositories and building tools that will enable historians to access the needed information more rapidly and fully.

1. INTRODUCTION

Libraries and historical archives (HAs) are regarded as the main repositories for preserving and maintaining historical documents. Their documents may have originated from either primary or secondary sources, and be maintained in the form of books (pages bound together), manuscripts, single pages, photos, paintings, video etc. A source is characterized as *primary* if it has been created

during the period of interest, whereas *secondary* sources are those created later on and are based on the analysis of primary sources [2].

Historians conducting research systematically examine past events to give an account; historic research may involve interpretation to recapture the nuances, personalities, and ideas that influenced these events, and the expected research outcome is to communicate an understanding of past events [10]. In this process, historians employ their scientific knowledge, experience and intuition to decide which information they will need to find and study during each next step, and subsequently attempt to locate sources that contain this information. In the context of a library or historical archive, the source location task may proceed in either of the following directions:

1. historians request from the library/archive personnel to retrieve for them the documents, by describing to them the information that the documents should contain and/or specific characteristics of the documents (author, period that the document was written, subject etc). The archive personnel will then exploit the conceptual model on the archive that they have developed and the tacit knowledge they have amassed from past searches to retrieve the pertinent documents and present them to the historian. The personnel's search may be aided by the archive's *content categorization scheme*, which is typically developed and maintained by the archive staff (mainly by augmenting it when the need arises, by adding new categories and/or sub categories) to assist them in organizing and managing documents.
2. if content digitization activities have been taken by the library/archive [4], typically contents are at least tagged with some keywords and/or structured metadata (creation date, author, etc), which are regarded to be useful towards locating documents of interest [1],[3]. Historians are provided with an interface that allows them to specify the characteristics of the documents they want to retrieve and matching documents can be then viewed on-screen (if their content has been digitized) or fetched by the archive personnel. In this context however, the digital archive categorization scheme into which documents are fitted has proven to provide little or no help at all for information location purposes [11], since categorization scheme is commonly compiled by archivists to suit their own needs for document organization and management, rather than for serving the needs of archive

users and researchers. This is a serious limitation, since the available information cannot be adequately exploited for retrieving the documents of interest.

According to the above, searches are very difficult without the help of the experienced archive personnel, which mainly relies on its tacit knowledge and experience, rather than on some explicit representation of knowledge about the archive content and tools that would offer guidance and automation for search tasks. Nevertheless, in order to build effective information retrieval tools for historians, their information requirements, search strategies and work patterns must be first analyzed and, insofar, very few data are available on these aspects. In this work we attempt to investigate the historians' search methods in the context of printed and digitized libraries and historical archives, to approach their strategy on their search on primary and secondary sources and to record their current practices and needs. We combined two different approaches to this end, the study of queries made by historians to an Historical Archive and the use of semi-structured interviews with historians. Our survey was conducted in the Historical Archive of the University of Athens.

The rest of this paper is structured as follows: section 2 briefly overviews related work, and section 3 describes the current status of metadata organization and digital aids in the HA of the University of Athens. Sections 4 and 5 present the methodology and the findings of the study of queries and the semi-structured interviews, respectively. Section 6 discusses the provisions that can be made by digitized archives and libraries to assist historical research and, finally, section 7 concludes the paper and outlines future work.

2. RELATED WORK

An important factor in our study was to understand what kind of data or information historians are looking for in a library/historical archive, either printed or digitized. Historians and researchers collect and process historical data in order to produce information connecting historical facts. Their main objective is to recreate the past through existing records and their interconnections. The collection of historical data is accomplished through methodical and comprehensive research in primary and secondary sources. Primary materials, which include the remaining records of archives, mail, books, etc of the time-

period of interest, are of special importance to historians, as they constitute the basis for original historical research.

Dworman in [20] focuses on pattern-directed queries to historical archives and collections ("What other animals are depicted with dogs in 19th vases of the collection?") versus item/record – oriented ones ("How many 19th century vases in the collection depict dogs?" and proposes an automated system for supporting the former. The focus of our work is more general and attempts to record and investigate the general historical research process.

Tibbo in [7] presents the preliminary results of a user study concerning the way historians locate primary resource materials in the digital age. Preliminary results suggest that electronic finding aids, well-designed websites and digitized documents, are helpful and should be available in archives, but cannot still be considered as replacements for more traditional methods of making collections available; the role of the archive personnel also remains important in aiding the historian. These results suggest that existing finding aids for digital archives do not cover all the needs of historical research. Although historical research is greatly based on intuition during the information seeking stage, it is not a casual discovery of facts. History is a science and although researchers use a wide range of practices and methods, common points may be noted and provide insight on a general methodology of historical research. As suggested in [7], there is a lack of user studies on this issue.

3. CURRENT STATE OF MATERIAL ORGANIZATION AND DIGITAL AIDS IN THE HA OF THE UNIVERSITY OF ATHENS

3.1 The Goal of the Historical Archive of the University of Athens and its Contents

The main goal of the Historical Archive of the University of Athens (HAUA) is the collection, classification and processing of historic material related to the University of Athens. The historic material currently owned by HAUA pertains to the period starting from 1837 (i.e. the era of its establishment) and ending in the decade of 1970; the ending period slides along with the "30-year rule", according

to which a document can be incorporated into an historical archive if 30 years have passed from its initial publication.

Insofar, the complete files of the following organizational units have been classified and catalogued:

- Senate secretariat (transcripts of assemblies and file of "unbound documents")
- Protocol (more than 500.000 unbound pages)
- Directorate of Faculties
- Directorate of Public Relations
- Secretariat of the School of Theology
- Secretariat of the School of Law
- Secretariat of the Medical School
- Secretariat of the School of Philosophy
- Secretariat of the School of Science

Each of these files is organized using (a) bound volumes or (b) envelopes containing unbound documents (or a combination of the two). The organization of the material was designed to fulfil the following two criteria:

1. *historical accuracy:* to maintain the structure that the material is believed to have had at the time of its creation, i.e. reflect the organization of the University during different periods of its existence and
2. *practical usefulness:* to facilitate processing of the material and extraction of useful information as answers to queries regarding administrative and research issues of the University of Athens. Note that the administrative issues include the way that files were organized during different periods.

3.2 Organization of printed material

Printed material was classified and organized using two methods, grossly corresponding to the periods 1837-1900 and 1900-now. During the first period, the administrative structures were still shaping -and thus were in an immature state-, and documents were classified according to the academic year they had been created in. After 1900, the document classification scheme changed and documents were classified according to their topic, as it was felt that the former classification scheme was impractical. To this end, a *topic taxonomy* was crafted.

Initially the taxonomy was shallow, but was later refined to accommodate more focused branches, while it also followed the major organizational changes of the University.

During the period 1900-1950, the archive personnel initiated a process of re-organizing the material of the first period, according to the categorization scheme of the second period. In order to save time, however, classification was not as elaborate, i.e. documents were categorized only on the top taxonomy branches, and additionally the type of the document (such as meeting transcripts, official letter etc) was noted. The classification scheme that resulted from this process is known as *Old Encoding Scheme* and applies to documents of the period 1837-1900, whereas the classification scheme that applies to documents of the period 1900-now is known as *Middle Encoding Scheme*.

In order to retrieve documents from the printed archive, a researcher submits a relevant query to the archive personnel, e.g. "I want the speech given by Missail Apostolides at the ceremony held for his appointment as a Rector". Since the organization of the archive is based either on years or thematic categories and contains no indication regarding the rectors' names, the HA personnel should use its tacit knowledge regarding the history of the university to conclude that Missail Apostolides served as a rector during the academic years 1842-1843 and 1850-1851 (this piece of information could be also contributed by the researcher); therefore, the set of documents pertaining to these academic years can be examined to locate the ones that the user had requested for. If the rector in question had served during the academic year 1930-1931, the archive personnel could exploit the classification according to the topic taxonomy, and examine only documents in the category "Elections of Administrative Bodies".

TABLE I. METADATA ATTRIBUTES ACCORDING TO THE OLD ENCODING SCHEME

Attribute	*Description*
Organizational Unit	The unit of the University of Athens that has produced the document, such as the Senate, a department's secretariat, the Public relations and history directorate and so forth.
Document Type	The type of the document, such as meeting transcripts, official letter, appointment decree etc.
Academic Year	The academic year within which the document was produced; e.g. "1931" refers to the period "September 1931-August 1932"
Creation date	The exact date on which the document was created in ISO format, e.g. 19310922 corresponds to "September 22, 1931".

TABLE II. METADATA ATTRIBUTES ACCORDING TO THE MIDDLE ENCODING SCHEME

Attribute	*Description*
Organizational Unit	Same as for the Old Encoding Scheme
Document Type	Same as for the Old Encoding Scheme
M	An indicator that metadata assignment is in accordance to the Middle Encoding Scheme
Thematic category	The topic category to which the document belongs to, e.g. "Financial", "Educational" etc.
Academic Year	Same as for the Old Encoding Scheme
General Category	The top-level category of the document, e.g. "Administrative"
Subcategory	A second-level category, e.g. "Healthcare" under "Administrative"
Sub-subcategory	A third-level category further specializing the second-level one, e.g. "First aid station"

3.3 Metadata schema for digitized material

The metadata schema for the digitized material closely follows the taxonomic scheme used for printed material. Therefore, two different attribute sets are used for digitized documents, the first one being employed for documents classified

according to the Old Encoding Scheme and the second one being assigned to documents classified according to the Middle Encoding Scheme. These attribute sets are illustrated in Table I**Error! Reference source not found.** and Table II**Error! Reference source not found.**, respectively.

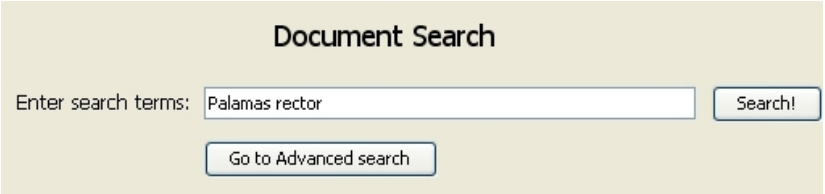

Figure 1. Simple search interface

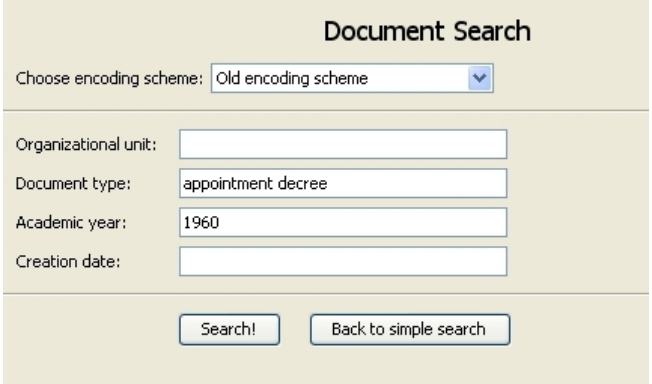

Figure 2. Advanced search interface.

When searching in the digital archive, an application is employed where the user can designate the criteria that the documents should fulfill in order to be retrieved. Criteria can be specified using either the *simple search* or the *advanced search.* In the simple search mode, users simply type in keywords in a search box (much like Google - Figure 1), and these values are matched against all metadata slots of documents; a document is retrieved if all entered keywords are matched. In the advanced search mode, the user initially selects whether documents from the Old or the Middle encoding scheme should be returned (in this mode, a query may only return documents from a single encoding scheme). Afterwards, the user is presented with a form having one input area for each metadata slot of the selected scheme (Figure 2), and the user may type in any of these areas the desired value for the particular metadata slot. Values are entered as free text (as opposed to selecting from a list of values, which is not supported), while for fields on the values of which no restriction is to be placed, the default value of "*" (asterisk) can be used as a wildcard.

Note that while the researcher may directly search the digital archive without the intervention of the archive personnel, still the archive personnel's tacit knowledge

is invaluable for correctly formulating queries that will return the desired documents, since the metadata attributes effectively remain the same as in the case of the printed archive.

4. STUDY OF USER QUERIES

The first approach we employed for surveying historian's research methods was to analyze the queries users have made to the Historical Archive (HA) of the University of Athens, requesting documents; since each query aims to retrieve documents relevant to a specific subject, query analysis could provide useful indications regarding the historians' interests in relation with the HA contents.
We performed an analysis of approximately 100 user queries made to the Historical Archive of the University of Athens. The queries were posed to the Historical Archive using natural language, and each query was modeled as a request for finding information regarding a number of *concepts* (persons, departments, locations etc) or interactions between these concepts (e.g. person X becomes president of department Y). Such a modeling of natural language queries is always possible through typical languages such as NKRL ([15], [16]). These languages also has the potential of modeling events, by introducing a "taxonomy of events" in parallel to the "taxonomy of concepts", and practically an event description is an relation between an arbitrary number of concepts, with the relation being drawn from the taxonomy of events.
In order to find the concepts within the queries and classify the queries into topic categories, we used Kaon's semi-automated concept extraction tool [6] to perform term extraction on the query texts, and subsequently identify frequently requested concepts. The results are grouped by query topic and are presented in the Table I. Figures in Table III correspond to the information that researchers wanted to obtain from the archive – in database terms, this corresponds to the *select list* of a query.

TABLE III. TOPICS OF QUERIES MADE TO THE ARCHIVE

Topic	*Percentage*
Person Biographies	24%
Historical Evolution of Institution/Organization	18%
Ceremonies	14%
General Socio-political issues	12%
Economic issues	10%
Administration of Institution/Organization	7%
Request for artistic or photographic material (photographs, of persons, portraits, monuments, etc)	6%
Books	4%
Time	4%

As seen from Table III, evolution-related queries, either person biographies or institution histories are predominant among the queries; this indicates that time and entity evolution is of great importance for historic research in the context of an archive (evolution-related queries constitute the 42% of the query bulk). Besides the implicit reference to the temporal dimension in these two query topics, approximately 32% of the queries involved an explicitly specified time period or time point to restrict the search scope (in database terms, this corresponds to the *where* clause of the query), while an additional 4% of the queries targeted to retrieving time points when certain events occurred (e.g. when the Hippocratic Oath was first taken in a graduation ceremony of the medical school) or periods (e.g. during which period did professor Kastorchis serve as a rector). Periods and points were provided in varying granularities, ranging from a whole century ("Names and biographies of professors who taught philosophy in the University of Athens after its creation in the 19th century") to specific dates ("Speech given in the Great Hall of the National University at the 21st of April, 1896"). Nevertheless, the vast majority of queries were restricted by periods or time points specified using year-level granularity ("Information for the Chair of Physiology of the Medical School from 1931 to 1939"). As a result, it seems that providing support for entity evolution and time-restricted queries would be important for historic research in the context of an HA.

Another useful conclusion that can be gained from Table III, is that tools and aids provided to historical researchers should include means for locating documents

falling under the listed topics. This can be achieved in various ways, including listing the topics within the keywords (unstructured information) or specifying the topics in specific metadata fields (more structured information). Providing a *topic taxonomy* which allows topics to be further classified into more specific concepts (e.g. *Ceremonies* can be broken down to Inaugurations, Commencements, Medal and award presentations, etc) can improve the search effectiveness, since queries can be better targeted.

5. User Study

While the study of the queries provided useful insight as to which are the topics historians are interested in, it offered no information whatsoever regarding the methods and strategies employed by historians for query formulation. Furthermore, the queries recorded in the Historical Archive's logs included only the queries that *could be answered* – e.g. if a query requested for documents referring to events that occurred in a certain place and this query could not be answered (because answering would involve an exhaustive examination of all documents which is clearly infeasible), this query was not recorded in the log, while researchers would refrain in the future from posing similar questions. Thus, in order to gather the missing information, as well as investigate whether any differences exist in search strategies and habits in printed and digital sources, we formulated a questionnaire, and asked historians participating in the user study to fill it in. Rather than giving the questionnaire away to the participants and collecting it afterwards, the approach of the semi-structured interview was chosen, in order to avoid misunderstandings and probe participants for explanations or more information where needed. The interviewed historians were researchers with knowledge and experience in information retrieval from various historical sources. The participants were chosen to be familiar with digital technologies related to information retrieval, in order to provide a more complete view on how they search for information both on digital and printed sources. The user group was composed of 5 men and 10 women. 4 of them are employees of the Historical archive, and 11 are historical researchers who have visited the Historical Archive of the University of Athens more than 3 times.

5.1 Questionnaire structure and interview procedure

For the needs of our study, the questionnaire was separated into two parts: The first part contains general questions recording the historian's profile, which primary sources of material -digital and/or printed- s/he employs, general types of queries s/he poses and closed questions [5] for generic concepts that s/he researches. In these closed questions the respondent should designate which of generic concepts presented to him/her s/he employs while searching for documents. The respondent should also rank the chosen concepts according to the frequency s/he employs them.

The generic concepts appearing as options in the closed questions were identified by analyzing the queries that historians had posed to the historical archive. Each concept appearing in these queries was extracted, and then mapped to top-level concepts in the domain of discourse. For example, for the query "what was the name of the professor that served as Dean in the University of Athens in 1912" the concepts are "Professor", "Dean", University of Athens", "University", " Athens" and "1912", and their mappings to generic (top-level) concepts in the domain of discourse are as shown in Table II. Using generic concepts rather than more specific ones was opted for, because it provides a more manageable and concise view of what users search for in the available material.

TABLE IV – MAPPING CONCEPTS TO GENERIC CONCEPTS

Concept	*Generic Concept*
Professor	Occupation
Dean	Occupation
University of Athens	Place
University	Institution
Athens	Place Name
1912	Time

TABLE V – GENERIC CONCEPT FREQUENCY RANKING

Generic Concept	*Order of Preference*			*Percentage of respondents that use this concept*
	Mean	StdDev	Median	
Name	1,4	0,507093	1	90%
Place	2,066667	1,032796	2	80%
Place-name	4,533333	0,99043	5	20%
Occupation	3,666667	1,112697	3	34%
Time	3,533333	1,245946	4	40%
Institution	5,666667	0,816497	6	40%

Table V illustrates the respondent percentages that stated to use the listed generic concepts in their queries, together with the frequency ranking they specified. Generic concepts Name and Place are the most extensively used ones, being employed by 90% and 80% of the users, respectively. Users were also asked to rank the concepts, starting from the first one they use in their queries and proceeding towards the last. The responses to this question are also summarized in table III, under the *Order of Preference* column; for this metric three figures are given, *mean* (i.e. the average of responses), *standard deviation – StdDev* (which shows how close to the mean value the individual responses are – the smaller the StdDev value, the less the distance from the mean) and *median* which corresponds to the answer most frequently given by respondents. *Name* appears to be the most preferred criterion, being ranked as first by the majority of respondents whereas the ones that did not rank it as first, assigned an order of preference equal to 2. *Place* is ranked second, having though been assigned a considerable number of rankings with order of preference equal to 1. Interestingly, the time criterion appears to be used by a moderate number of respondents (40%) and to have a low order of preference (fourth). We have to note, however, that two important category topics, namely *Person Biographies* and *Historical Evolution of Institution/Organization* contain an indirect reference to the temporal dimension, which may not have been taken into consideration by respondents.

An issue worth noting here is that most respondents declared to prefer to use Place and not Place-Name, even though they stated that they do not clearly understand the difference between the two concepts e.g. the term “University of Athens”

would be regarded as *Place* and not as *Name*. On the other hand, the concept "Athens" would be regarded as *Place* and as *Place Name*. The concepts Occupation and Time were used less frequently. It was believed that, if a specific date is not given to you, it is difficult to search for something, among several different periods of time, especially as far as historical research information is concerned.

Overall, it seems that historical researchers tended to choose concepts whose meanings were clear to them, disregarding concepts whose meaning was unclear or ambiguous. This suggests that the vocabulary to be used in any tools and aids that will be made available to historians should be carefully chosen, so as to be clear to the tools' users, since otherwise the related features might not be used at all.

The second part of the interview was composed of seven information retrieval tasks, and respondents were asked to describe in detail how they would proceed in retrieving information, both in digital and printed sources, in order to complete each task. Four of these tasks were typical queries to the Historical Archive of the University of Athens, whereas the remaining three were based on queries made to the H.A, but transformed to facilitate recording information for different types of searches. Through this procedure we aimed to investigate the different ways a historian may face a specific question with different sources available and what are his/her expectations and preferences.

In this part of the questionnaire we used open questions [5] since the respondent would describe how s/he would proceed in locating documents related to a specific historic question. Closed questions could not be used for gathering this information, since each historian employs a personal strategy for information foraging, thus the number of options is practically unlimited. Moreover, if certain strategies appeared on the questionnaire, respondents might be influenced and include them in their answers, even though they do not usually employ them.

The seven information retrieval tasks are:

1. Describe how would you search for information regarding «Kostis Palamas as Secretary General of the University of Athens»

2. Describe how would search for information on the historical evolution of the Chemistry Department of the University of Athens
3. Describe how would search for information regarding the historical evolution of an organization or city.
4. Describe the way that you would seek for the PhD thesis of X who lived between 1850 and 1920.
5. Describe how would you search for the Curriculum Vitae of a teacher (e.g. P. Papageorgiou), who taught the "Greek Studies" course in the Department of Philosophy during the academic year 1909-1910.
6. Describe how you would look for information in case of synonymy. How would you verify that two pieces of information actually refer to the same entity or synonymous entities?
7. Describe how would you search for information regarding «female graduate students of University of Athens coming from Smyrna».

5.2 Study Results

A view shared by almost all interviewed historians was that digital sources were less reliable than printed, traditional ones; however, they stated that they did use digital sources like archive web pages and general web search in order to locate additional material.

The maintenance of a personal archive of notes and copies of documents, when possible, either printed or digital, is a common practice among all researchers. The researchers explained that they organize this archive and keep various kinds of metadata like notes, dates, document descriptions, interesting citations copied from documents etc. The form of this archive varies, as in some cases they organize their notes per document copied and in others per research subject they are working on.

The analysis of the second part of the interview, which contained specific information retrieval tasks, produced several observations related to the historian search method.

As interviewees explained, when faced with a particular topic, they break it down into several questions that define their information retrieval tasks from primary and secondary sources. For example, for the question: "What is the work of Kostis Palamas during the years of his tenure as a Secretary General in the University of

Athens", users provided several responses as to which sub-topics they would investigate. An example is "During which period was Kostis Palamas a Secretary General?", "Did he visit foreign universities during this period?", "Are there any documents in the University Archive with his signature or mentioning his name?" etc.

The researcher then proceeds to explore these questions to the appropriate sources. The way that each question is investigated in the primary sources has been presented in [21] and summarized in the following steps:

1. They identify and isolate entities, like persons, places or organizations, related to the topic of their research. These entities are represented by one or more keywords, as in some case an organization, for example, may be referenced by its full name or its acronym. In many cases the search is restricted by a time point (date, year, etc) or period. More details on this issue may be found in [21].
2. They focus on one keyword at a time and look for material in the primary and secondary sources available. As the users explained, they firstly focus on the keyword which they believe are more closely related to their topic and then investigate the rest of them one by one in order to have a clear view of the material produced by each different keyword and not to miss useful material.
3. They attempt to perform searches combining more than one of the identified keywords, for example name – date, or place – name – date. These combinational searches produce in some cases more focused and relevant results, so in this case precision is the main objective of the historian.
4. They use synonyms and derivatives of the keywords. For example, for the topic "history of the department of Chemistry", apart from the word "Chemistry" they would use the word "chemical". This approach is used mostly if the previous steps did not produce many useful results.
5. They introduce new concepts that they consider related; e.g., for the "Department of Chemistry", they would introduce "study programme", "professor" or "book". These concepts are most often a result of the study of the material retrieved in previous steps. As the researchers gradually get more familiar with their topic, they are able to identify more and more related concepts.
6. They create and investigate various combinations of the initial terms, their

synonyms and related terms. Related terms may be derived from generic or more specialized concepts relevant to the initial ones (e.g. for "Postgraduate Student" they may use "Student" [generalization] or "PhD Student" [specialization]): the specialization/generalization relationships between the terms stem from the mental model that researchers have formulated for the domain of interest, and organize terms in a *hierarchical taxonomy*; Figure 3 shows an excerpt of such a taxonomy for the domain of a University.

Related terms may also be connected to the initial ones with relations like "belongs to" or "works at" (For the "Department of Chemistry", "Faculty" or "University" could be possible related terms). These relationships originate from a mental model of the domain which is more semantically rich and expressive, as compared to the taxonomy, since this model does not only record specialization/generalization hierarchies but other (domain-specific) relationships between concepts as well. An example of such a mental model in the form of a *semantic net*, is illustrated in Figure 4. This semantic net is derived from a researcher's answer on how she would proceed for answering the query "Find information on the history of the Department of Chemistry". Note the generalization relation for "Student" and "Professor" (denoted through directed arcs), as she explained that after looking for "Students" or "Professors" she would search for other "Persons" related to the specific department.

Within our experiment, eight respondents used a hierarchical taxonomy for selecting new terms, while the remaining seven employed the semantic network structure for enriching their search. It is worth noting here that only four stated expressly that they start from the most specialized keyword of their query and then proceed to generalizations or related terms, while the rest of the respondents performed this task intuitively.

Finally, the enrichment of the initial terms with new ones is performed incrementally, introducing to the search firstly those that seem more relevant and then the less relevant ones. It should be also noted that the process of enrichment of the initial terms with related ones, as derived from the study of the experiment subjects' responses, is in accordance with the model of the human mental lexicon as described in ([8] pp. 289-294). It is suggested that concepts in our brain are represented in a semantic network of words, as in

Figure 4. The strength of the connection and the distance between the nodes are determined by the semantic relations or associative relations between the conceptual nodes. This model assumes that activation spreads from one conceptual node to those around it, with greater emphasis to the closer ones. A hierarchical structure is also present in this network, classifying concepts in more generic and more specific ones.

Administrative body
- Department administration body
  - Department general assembly
  - Department president
  - Department vice-president
- University administration body
  - Rector
  - Senate

Person
- Professor
- Student

Figure 3. Sample hierarchical term taxonomy for the domain of the University. More specific terms appear below the respective generic terms and indented to the right.

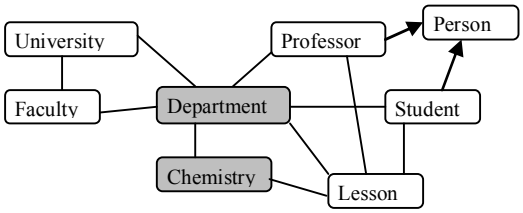

Figure 4. Example of a Semantic Net

Regarding the way that the mental model for the domain is built (either taxonomic or semantic net-structured), respondents stated that they try to identify a minimum set of high-level concepts that are contained in or are relevant to the historical information they are searching for. Once the relevant concepts are identified, they are structured in hierarchies and linked through relationships. According to the historians' point of view, organizing information in this way helps them to better understand the concepts they find in the historical sources, the way they are joined, and integrate them to a comprehensive conceptual and chronological frame.

Another point to be noted here is the differences observed on the way researchers would investigate a query using traditional printed sources and digital ones:

1. In digital search, they used fewer keywords than in the traditional one.
2. In digital search, they used fewer *combinations* of keywords and confined themselves to using simple search only, neglecting the advanced one.
3. In digital search, they used less synonyms or related concepts (in some cases none at all) and limited themselves to the keywords present in the topic.
4. Most of them believed they would not get the desired results using digital search and reported that in this case they would then turn to traditional search methods.

These results suggest that the current state of digital HAs (cf. section 3, Current state of material organization and digital aids in the HA of the University of Athens) has not managed to win the trust of the history researchers. They seem to feel more confident that with using the traditional primary sources in printed format they will have better access to the historical data needed in their research. From the responses provided by the history researchers in the context of the interviews, the main reasons for this lack of trust towards digital HAs are the following:

1. high volumes of irrelevant documents are retrieved (low *precision* [12]). Researchers commented that while in the printed versions they considered "natural" to go through a number irrelevant documents and reject them, in the case of a digitized HA they expected the system to be able to filter-out most of the documents not related to their queries. Coarse-grained metadata (i.e. specification only of the organizational unit that has created a document, rather than a specific author) or lack of metadata are the most usual causes.
2. not all relevant material is retrieved (low *recall* [13]). Obscure classification schemes that are meaningful only to the personnel of the archive, missing or erroneous metadata, lack of consistency (e.g. use of abbreviations in some instances and expanded forms in others) and lack of linkage between entities (e.g. a Department is not explicitly linked to its professors or its presidents; linkage is only implicit through the contents of the documents) are the most typical causes for failing to retrieve relevant material. It has to be noted here that when searching in

printed archives, errors in the documents (e.g. misspelled names) or inconsistencies are tackled, since the researcher can understand the meaning of the document; in the case of digital archives though, search is performed by software and matching is performed at string level only. The fact that a researcher first reads a document in the printed archive and then decides if it is relevant or not, alleviates the need for considering all search criteria beforehand, as must be done in the case of searching in digital archives. The lack of metadata or errors in them could be tackled by broadening the scope of searches, but this results in retrieving even more irrelevant documents. Some researchers finally commented that they used fewer keywords, synonyms or keyword combinations, because they expected the system to be "smart enough" to retrieve documents that would match synonyms, keyword combinations, or terms semantically associated to the ones given.

3. the fact that query formulation needs the intervention of the HA personnel, who indicates which search terms should be used in which fields is an additional impediment, since it introduces delays (HA personnel may not be always available) and is an additional source of errors (the personnel's tacit knowledge may be incomplete or imperfect).
4. there may be issues with the users' skill levels in computer usage: approximately 25% of the users that participated in the survey have rated their own computer skills as "below average". Additionally, 35% of the users stated that they not feel confident enough to use advanced search, although the latter can limit the number of results irrelevant to the user's intention (e.g. if the user is searching for documents in the "Educational" thematic category, entering "Educational" in the simple search may match appearances of the term in *both* the "Organizational unit" and the "Thematic category" metadata slots, whilst if advanced search were employed, the user would clearly designate his/her intention by entering the search term in the appropriate area in the search form).

The above findings are inline with the observations in [13], where it is concluded that "the overwhelming majority of historians want to see and use historical sources in their original format" and that "electronic access and digital reproductions have great, untapped potential". However, the progress of

information retrieval and semantic web technologies in the past years should be put to the service of historical archive researchers, by incorporating techniques and practices into digital finding aids. Such techniques and practices are discussed in the following section.

6. Assisting Historical Research

Based on the results of both studies (user queries and questionnaires), a number of requirements for the organization of information and the functionality of the tools that will be available to researchers can be identified. These requirements are presented in the following paragraphs.

1. the digital repository contents should be tagged with *complete* and *structured* metadata. *Metadata completeness* refers to the need that information typically used in researchers' queries should be available as metadata. In particular, the topic of the item, its author, date of creation, period to which the content refers, and involved entities (entities referred to in the documents) should be listed within the item's metadata. *Metadata structuring* refers to the need that this information should be stored as separate fields in the item's metadata, not as mere keywords. Structuring allows the researcher to gain more control over the search procedure and get more relevant results. For example, if structured metadata are available, the researcher may request documents authored by "Palamas", whereas if no structure is imposed the query will return all documents that refer to "Palamas", most of which will be irrelevant to the researcher's search.
   It has to be noted here that in the context of even a medium-sized archive, it is not feasible, in terms of time and cost, to manually create complete and structured metadata for all its contents. Yet, two factors can alleviate this impediment:
   a. Historical archives are gradually incorporating documents in digital format; these documents can be automatically processed and information regarding their authors or other metadata with relatively high accuracy, precision and recall (e.g. [14]). Therefore, digitally available documents cannot be processed to have the metadata extracted, and be subsequently incorporated into the archive, together with the relevant metadata.

b. The paradigm of user-contributed tagging employed insofar successfully by museums can be adopted. Insofar, a number of museums, such as the Brooklyn Museum[1] and the Cornell's Lab of Ornithology NestCam project[2], allow their users to tag the contents, helping other users to find them or even themselves to re-find them. Similarly, historians performing research in an archive can contribute such information to the system through appropriate interfaces.

2. the topics available, the document categorization scheme, the entities of interest in the domain of discourse and the timeline covered by the repository should be expressly represented using an appropriate scheme (taxonomy or semantic network), and be made available to researchers, alleviating thus the need to rely on the personnel's tacit knowledge for conducting a successful search. The scheme should be populated with both generic and specific terms, suitably organized in hierarchies, to allow the researchers to tune the scope of their searches accordingly to the information they have available and the goal of their queries. (Naturally, a search for a generic term should fetch all documents that are tagged with any more specific term than the one searched for). Therefore, items should be tagged with the most specific term possible.

   Under the presence of exhaustive term hierarchies, this approach could be counter-productive, since taggers may need to spend considerable time browsing through the hierarchies to locate the specific tag. In such cases, tagging specificity can be relaxed in favor of productivity; yet, it could be possible to employ the techniques mentioned in (1) in order to either automatically create tags from documents available in electronic format or exploit researcher-provided tagging. Another technique that can be of use here is the use of OCR techniques, not for generic image-to-text recognition but for recognition of named entities in the text ([17] [18]), which can serve as specific tags.

3. the manner that users place queries against the archive should be kept as simple as possible, to allow users that are not highly competent with IT systems to work with the system. Advanced search features should be

[1] http://www.brooklynmuseum.org/opencollection/tag_game/start.php
[2] http://watch.birds.cornell.edu/nestcams/clicker/clicker/index

included, but access *to the same functionality* should be also provided to more naïve users. For instance, instead of requiring the user to explicitly specify the metadata field against which a keyword must be matched, the user could simply enter keywords as in simple search, and then the system could offer a menu through which the user could disambiguate his/her intentions (e.g. "Does *Educational* refer to the organizational unit *Educational Directorate* or the thematic category *Educational Affairs* (or *both*)?"). The latter approach can be also used for addressing the cases that using advanced search limits *recall*, since *all* results are retrieved and the user then limits the displayed results according to his/her desires.

4. the choice of the terms that are used for describing the domain of discourse should be careful, and –among candidate terms for this description– the ones deemed more clear and unambiguous should be preferred. Since clarity and ambiguity are subjective, terms should be appropriately clarified and/or disambiguated through accompanying descriptions.
5. since the mental model of each researcher for the domain of discourse may differ from the model adopted by the digital repository, the tools must provide means for researchers to align their mental model to the digital repository's model. Drilling down the concept hierarchy, searching, synonyms and thesauri, as well as descriptive texts for the adopted concepts could greatly assist the researchers in choosing the right concepts.
6. the tool should provide means to limit the scope of searches to points or periods in time, both regarding the time that a document was authored and the time to which the document refers to (the latter is particularly useful for secondary sources). Proposing of terms whose spelling most closely matches the terms given by the researcher could be also used to handle misspelling cases; this is particularly important in the context of historical archives, since names are often found to be written with different spellings in different time periods.
7. since the retrieval of person biographies and evolution of institution appears to be a frequent query, the tools should assist researchers in locating documents that refer to different periods of the same entity. [11] lists some heuristics that can be employed by such a tool.
8. given that researchers limit the number of query term combinations when

searching in digital repositories, as compared to when they search in printed repositories, the tools provided to researchers could compensate for this reluctance. A possible method could be to ask the researcher for all query terms to be employed and subsequently formulate automatically all possible term combinations. For example, if the user would enter the terms "chemistry", "laboratory" and "faculty", the tool could create the combinations "chemistry/laboratory", "chemistry/faculty", "laboratory/faculty" and "chemistry/laboratory/faculty".

9. finally, since researchers have been found to use less synonyms or related concepts when searching in digital archives, the tools could suggest related terms, extracted from semantic network connections, to assist researchers in the phase of enhancing their queries. The tools could also automatically extract query term synonyms from standard thesauri (e.g. Wordnet [19]) and suggest them to the user for search query enhancement.

7. Conclusions and Future Work

This work presents a user study aiming to record the historians' information retrieval methods in the context of an Historical Archive. The study was conducted both by studying typical queries that historians pose to the archive and by interviewing researchers. Through gaining insight to the practices employed by researchers, requirements for information organization and tool support so as to facilitate historic research within digitized repositories of primary and secondary sources can be formulated. Based on an initial set of these requirements, a prototype tool architecture has been drafted [11] and an initial ontology schema has been designed. The ontology schema has been populated by automatically processing the metadata present in the filenames of the digitized documents, however these metadata are coarse-grained and partial, necessitating thus their refinement and completion. Future work will include the completion of the prototype tool implementation, and the testing of this tool in the context of the Historical Archive. Extending the presented surveys to include subjects working in other archives and/or different historical subjects (e.g. national history) will also be considered.

We also plan to continue elaborating on the requirements for creating digital tools for specialized and demanding user groups like history researchers, exploiting the

extensive user studies undertaken within the EU funded project Papyrus[3], in which our group participates. Papyrus intends to provide the appropriate ontology-based technologies to bridge the domain of History with News Archives. This project, through access to several societies and groups of users (two participating educational and research institutions, both members of the "Tensions of Europe" network of experts for the History of Science and Technology) will contribute to the requirements elaboration phase through providing substantial amount of material that will be used to extract more specific user needs and requirements.

References

[1] K. Cruikshank, C. Daniels, D. Meissner, N. l. Nelson, M. Shelstad, How Do We Show You What We've Got? Access to Archival Collections in the Digital Age. Journal of the Association for History and Computing 3, no. 2 (2005).

[2] Identifying Primary and Secondary Sources - A Preliminary Guide , http://www.libraries.iub.edu/index.php?pageId=1483 [Accessed September 7, 2010]

[3] Pitti, D. V., Encoded Archival Description, An Introduction and Overview, D-Lib Magazine, Vol 5, No 11, November 1999

[4] Leiner, B. Bm: The Scope of the Digital Library, Draft, DLib Working Group on Digital Libraries Metrics, January 16, 1998. http://www.dlib.org/metrics/public/papers/dig-lib-scope.html [Accessed September 7, 2010]

[5] Open and Closed Questions: http://changingminds.org/techniques/questioning/open_closed_questions.htm [Accessed September 7, 2010]

[6] KAON, http://kaon.semanticweb.org/ [Accessed September 7, 2010]

[7] Tibbo, H. R., Primarily History: Historians and the Search for Primary Source Materials, in Proceedings of the 2nd ACM/IEE-CS joint conference on Digital Libraries, 1-10, 2002

[8] Gazzaniga, M. S., Ivry, R. B., Mangun, G. R., Cognitive Neuroscience, The Biology of the Mind, W. W. Norton & Company, New Ed edition (April 1998), ISBN: 978-0393972191

[9] Protégé project, Stanford University, http://protege.stanford.edu [Accessed September 7, 2010]

[10] Investigative Techniques Glossary, http://www.pbs.org/opb/historydetectives/techniques/glossary.html [Accessed September 7, 2010]

[11] Katifori, A., Torou, E., Vassilakis, C., Halatsis, C., Supporting Research in Historical Archives: Historical Information Visualization and Modeling Requirements, Proceedings of the 12th International Conference on Information Visualization IV 08, London 2008, pp. 32 - 37.

[3] http://www.ict-papyrus.eu/default.aspx?page=home

[12] Baeza-Yates, R., Ribeiro-Neto, B., *Modern Information Retrieval*, 2nd edition. Addison Wesley & ACM Press, 2008, ISBN-13: 978-0321416919.
[13] Duff, W., Craig, B., Cherry, J. Historians' Use of Archival Sources: Promises and Pitfalls of the Digital Age. *Public Historian*, Vol. 26, No. 2. (2004), pp. 7-22.
[14] Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z., Fox, E. Automatic Document Metadata Extraction using Support Vector Machines. Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries, Houston, Texas, 2003, pp. 37 – 48.
[15] Zarri, G. P. Semantic Web and Knowledge Representation. Proceedings of the 13th International Workshop on Database and Expert Systems Applications (DEXA'02), Aix-en-Provence, France.
[16] Zarri, G. P. Representation and Management of Narrative Information Theoretical Principles and Implementation. Springer-Verlag London Limited, 2009. ISBN: 978-1-84800-077-3, DOI: 10.1007/978-1-84800-078-0.
[17] Nadeau D., Sekine, S.. 2007. A survey of named entity recognition and classification. Linguisticae Investigationes, 30(1):3–26.
[18] Wang, W., Xiao, Ch., Lin, X., Zhang, Ch., 2009. Efficient approximate entity extraction with edit distance constraints. Proceedings of the 35th SIGMOD international conference on Management of data, pp. 759–770, Providence, Rhode Island, USA. ACM.
[19] Fellbaum, Ch. (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press. ISBN-13: 978-0-262-06197-1
[20] Dworman, G. O., Kimborough, S.O., and Patch, C. 2000. Pattern-Directed Search of Archives and Collections, Journal of the American Society for Information Science, Volume 51, Issue 1, Special topic issue: When museum informatics meets the World Wide Web, 2000, 14 - 23, ISSN:0002-8231
[21] Torou, E., Katifori, A., Vassilakis, C., Lepouras G., Halatsis, C., Capturing the historical research methodology :an experimental approach, International conference of education, research and innovation, Madrid, ICERI 2009