Research Methodology on Historical Archives

Torou Elena[*]

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
etorou@di.uoa.gr

Abstract. This dissertation attempts to investigate issues related to requirements that the archives' information systems should apply in order to effectively support the historical research. The requirements are based on the general rules of the historical research methodology, the practices used by the historians during their research and the studies related to the effectiveness of the several methods of information recovery and visualization in the historical archives.

Based on historical research methodology, the general rules and the practices that are used by the historians during their research, a methodology for the ontologies' development has been developed. These ontologies are used in the management of the historical archives material. The developed methodology is general and takes into consideration not only the methodological historical research rules but also the special characteristics of each archive.

The developed methodology will be used at any historical archive environment, while the usage of a number of heuristics which will be used on the ontologically managed knowledge, will assist the historians on the monitoring of the ontologies evolution throughout time, by making easier the historical research conduct and the modulation of the historical conclusions.

Keywords: Research methodology, Historical archive, Ontology, User study, Digital Libraries

1. Introduction

Libraries and historical archives (HAs) are regarded as the main repositories for preserving and maintaining historical documents. Their documents may constitute 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].

Digitized historical archives (HAs) could be considered as a special case of digital libraries; they have however, characteristics that differentiate them. In particular, the digitization process in the context of HAs is inherently more demanding than the equivalent in common digital libraries, mainly due to the large volume of the original material and its poor preservation state, as well as to the convoluted and archaic handwriting often found in documents of HAs. At the best case, keywords or other metadata (creation date, author etc) will be available [14]. Commonly, documents in a HA are fitted into a categorization scheme, which has proven to provide little or no help at all for information retrieval purposes, as it is typically compiled by archivists to suit archiving purposes. As a result, even browsing

[*] Dissertation Advisor: Constantin Halatsis, Professor

becomes very difficult without the help of the experienced archive personnel, which mainly relies on their conceptual model of the archive, rather than on some explicit representation of knowledge about the archive content and tools offering guidance and automation for search tasks.

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 [1]. Their main objective is to recreate the past, through existing records and their interconnections. 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 this work, we attempt to investigate the historians' search methods in the context of printed and digitized libraries and historical archives. An important factor in our study was to understand what kind of data or information historians are looking for in an historical archive (or library in general), either printed or digitized, and which research methodologies or research models they use, while they investigate an historical archive. Since this issue has not been addressed insofar [3],[4], no methods for elucidating research methodologies or research models that historians employ have been reported in the literature.

2. Related Work

The recent great digitization effort has resulted in the creation of numerous Digital Libraries that may be accessible by historians.

The European Culture Heritage Online[†] (ECHO) is a collection of digital libraries of 50 scientific and cultural institutions worldwide, which contribute cultural heritage content as well as scholarly metadata. Access to digitized material, the majority of which concerns philosophy and science, is possible either by inserting key words or by browsing in thematic categories. The Perseus Digital Library[‡] of Tufts University contains primary material and secondary sources for research in the humanities, which are accessible by browsing or by inserting keywords in simple or advanced search. The digital library of Perseus includes various collections, such as the Classics Collection, the Renaissance Collection, the Bolles Collection, the California Collection, the Upper Midwest Collection, the Tufts History and the Boyle's Papers. The site also offers historical information on the related areas. A lot of the primary material is also available in text fromat. For the material in Greek, a transliterated version with comments was chosen, based on various bibliographic sources.

[9] presents a project on historical scene investigation, which encourage students in the process of «doing history». In that project, students are provided with a set of questions to guide their analysis and their step-by-step analysis on historical clues. There are other many academic and cultural institutions that have started to digitize primary material, making it available in the form of PDF images. Greek digital libraries such as Pergamos[§] and Hellinomnimon[**], both of the National and Kapodistrian

[†] http://echo.mpiwg-berlin.mpg.de/home

[‡] http://www.perseus.tufts.edu/

[§] http://pergamos.lib.uoa.gr/dl/index

University of Athens, offer simple and quick access to rich collections of digitized material of relevance to Modern Greek Studies scholars. Most of this material is old and rare. The researcher can browse the digital representations of old prints, manuscripts and visual material for historical, biographic and bibliographic information.

The National Library of Greece has developed a digital library[††] that is accessible through the Web. This digital library contains five Greek newspapers in digitized form, covering the period from the end of 19th to the middle of the 20th century. The material appears in PDF form. Each page of the newspapers corresponds to one PDF document. Newspaper issues are accessible either by browsing a calendar for each newspaper or by inserting the desirable date as a keyword. In addition to offering access to the digitized material, this digital library offers the researcher a useful real-time OCR tool for searching into the digitized material. Its major disadvantages are: a) that it does not return all the relevant results, and, b) that the interface of the display of the results is rather inconvenient for the user, since he/she has to open all the relevant pages in order to find what he/she is looking for.

All these efforts do not offer functionality beyond simple keyword search and browsing of the catalogs of the archive collections.

3. Studying historian research methods

An important factor in our study was to understand what kind of data and/or information historians are looking for in a library/historical archive, either printed or digitized, and which research methodologies or research models they use while they investigate a historical archive. Since this issue has not been addressed insofar, and therefore there are no methods for elucidating research methodologies or research models that historians employ / use, we formulated a questionnaire comprising of seven information retrieval tasks commonly addressed in the context of historic research. History researchers were asked to describe in detail how they would proceed in searching for the information they need for completing these tasks. Through this procedure we aimed to investigate the different ways a historian can use to tackle a specific question, examine whether there exists a common research methodology, and the historic researchers' expectations and preferences.

In order to identify the historical researchers' needs and requirements, we combined two different approaches: (a) the study of queries made by historians to the Historical Archive of the University of Athens and (b) the use of semi-structured interviews with historians, results in [15][16][17]. More specifically, in order to collect the information on the different ways a historian may address a specific historical question, as well his/her expectations and preferences, we collected and analyzed information regarding the number and type of terms that researchers employ for retrieving the information they require, while searching either in a printed or in a digitized historical material.

Most of the interviewees found the idea to participate in the survey, recording on a piece of paper their line of though, the data or information they are looking for in a historical archive, and the queries they formulate to achieve the retrieval of this data/information, very exciting. They told us that it could be

[**] http://www.lib.uoa.gr/hellinomnimon/

[††] http://www/nlg.gr

very helpful if a methodology for conducting research on historical information was developed, since this could provide valuable guidance for conducting research, especially for the less experienced historical researchers. They also commented that the availability of digital tools to guide researchers through the steps of the methodology and assist them in performing each step would be of great assistance.

4. Practices on research methodology

Regarding the historical researchers' methodology for formulating queries in order to retrieve documents relevant to their research, we observed the following practices, which are more or less followed by all researchers:

1. They identify and isolate keywords in the topic of their research. These keywords are very often entities like persons, places or organizations and in many cases the search is restricted by a time point (date, year, etc) or period.
2. They focus on one keyword at a time and look for material in the primary and secondary sources available.
3. They separate compound terms like «Department of Chemistry» into to individual terms («Department», « Chemistry»), in order to isolate two terms and investigate each one of them separately so as to introduce new related concepts.
4. They attempt to perform searches combining more than one of the identified keywords, for example name – date, or place – name – date.
5. 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 also use the word "chemical".
6. 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; e.g., for the "Department of Chemistry", they would introduce "study programme", "professor" or "book".
7. They organize terms in a hierarchical taxonomy by using a mental model on depicting the related terms closer than the others.
8. They use connections between related terms, by connected terms 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).

The phases presented above are the steps of a mental technique that historians use, while searching in a digital or in a printed historical material. These findings can be used for education purposes, since they can be incorporated in a methodology for conducting research on historical information. Furthermore, these findings can serve as the basis of user requirements, when building tools to support historical research, since such tools should help researchers perform these phases more efficiently in terms of completeness (all steps are performed and all possible options are available to the researcher to try) as well as in terms of time (document retrieval should be performed more rapidly).

5. Experiment evaluation

According to the researchers' responses during the interview, their majority prefers to search in printed collections, rather than in digital ones. This mainly stems from their experience with searching in digital collections, where the results were poor, since (a) a lot of documents irrelevant to their queries were retrieved (low precision [12][12]) and (b) many important relevant documents were missed (low recall [12][12]). According to the users, the metadata that is used by digital libraries or digital historical archives [5][6], do not cover the needs and requirements of the historical research; especially referring to questions related to the entity evolution [8] [10] like the evolution of an institution or a person. Additionally, in digital search, they used fewer combinations of keywords. Another interesting aspect is the fact that even if an advanced search was available in the digital tools, researchers confined themselves to use simple search only, neglecting the advanced one. For instance, in the case of Google Scholar [13] where it is possible for the user to pose a query with more detail, e.g. requesting articles by a specific author or published in a specific period, researchers only used the simple search, where terms are given in a "flat" fashion (e.g. if an author name is given, documents containing the designated name in the author list, paper body, footnotes or references are returned). Researchers stated that this is owing to the fact that in many cases they have missed important documents due to metadata incorrectness (e.g. the author's name has been used instead of his/her surname; the year of publication of a conference's proceedings has been used instead of the year that the conference was held) or incompleteness (e.g. the year of publication has not been recorded at all). As for the synonyms and concepts related to the query terms, in digital search, they used less synonyms or related concepts (in some cases none at all) and limited themselves to the keywords presented in the topic.

6. Assisting historical research

In order to assist historical research, a number of requirements for the functionality of the tools, that will be available to researchers, were identified [17]. The results were based both on methodology results and the experiments evaluation [8][16][17]].

7. Conclusions

This thesis has proposed in order to presents a user study aiming to record the historians' information retrieval methods in the context of a Historical Archive. The study was conducted by examining the research methodologies they use while they investigate a historical archive. Through gaining insight to the practices employed by researchers, requirements for information organization and tool support so as to facilitate historical research within digitized repositories of primary and secondary sources can be formulated. Based on an initial set of these requirements, regarding the terms they use, the frequency that each term appears in each IR task and the importance of time in historical research, a prototype tool architecture has been drafted [7] and an initial ontology schema has been designed [11]. 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.

Acknowledgments. This dissertation has been concluded under the PENED 2001 framework.

References

[1] Investigative Techniques Glossary, http://www.pbs.org/opb/historydetectives/techniques/glossary.html

[2] Primary and Secondary sources, http://ipr.ues.gseis.ucla.edu/info/definition.html

[3] 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

[4] Mark Vajcner, The imprortance of context for digitized archival, Journal of the Association for history and computing, Volume XI, Number 1, April 2008

[5] Ian H. Witten, David Bainbridge « How to build a digital library», Morgan Kaufman publishers, 2003

[6] Victoria Irons Walch, Marion Matters,Standards for Archival Description: A Handbook, The Society of American Archivists, 1994 http://www.archivists.org/catalog/stds99/toc.html

[7] A. Katifori, E. Torou, C. Vassilakis, C. Halatsis, Supporting Research in Historical Archives: Historical Information Visualization and Modeling Requirements, Proceedings of IV 08

[8] Torou, E., Katifori, A., Vassilakis, C., Lepouras G., Halatsis, C., Creating an Historical Archive Ontology: Guidelines and Evaluation, Proceedings of the First IEEE International Conference on Digital Information Management (ICDIM 2006) December 06-08, 2006, Bangalore, India

[9] Kathleen Owings Swan and Mark Hofer, The historical scene investigation (HSI) Project: Facilitating historical thinking with web-based, digital primary source documents, Journal of the Association for history and computing, Volume XI, Number 1, April 2008

[10] Katifori A., Torou E., Vassilakis C., Lepouras G., Halatsis C., Daradimos E., Historical Archive Ontologies – Requirements, Modelling and Visualization, Proceedings of the RCIS 2007 Conference

[11] Torou, E., Katifori, A., Vassilakis, C., 2007a, University of Athens Historical Archive Ontology Version 1, http://oceanis.mm.di.uoa.gr/pened/?category=pub#ontos

[12] Wikipedia, 2008. Precision and recall, http://en.wikipedia.org/wiki/Precision_and_recall

[13] Google inc, 2009. Google Scholar advanced search. http://scholar.google.gr/advanced_scholar_search

[14] Kobsa, A. 2004. User Experiments with Tree Visualization Systems. In IEEE Symposium on Information Visualization (INFOVIS'04), 9-16

[15] Katifori, A., Torou, E., Halatsis, C., Vassilakis, C., Lepouras G., A Comparative Study of Four Ontology Visualization Techniques in Protégé: Experiment Setup and Preliminary Results, IV 2006

[16] Katifori, A., Torou, E., Vassilakis, C., Lepouras, G., Halatsis, C., Selected Results of a Comparative Study of Four Ontology Visualization Methods for Information Retrieval tasks, Proceedings of IEEE RCIS 2008

[17] 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