Supporting Research in Historical Archives: Historical Information Visualization and Modeling Requirements

Akrivi Katifori, Elena Torou, Costas Vassilakis, Constantin Halatsis
vivi@di.uoa.gr, etorou@di.uoa.gr, costas@uop.gr halatsis@di.uoa.gr

***Abstract*— The on-going progress in the area of digital libraries has lead to the beginning of a digitization effort in Historical Archives, as well. The requirements of historical research, which works with histories of entities and incomplete information, create the need for supplementary tools to support users in handling the digitized content. This work is based on a user study of historian information retrieval methods in order to create a set of tools for the context of historical archives, which will facilitate historical data storage, management and visualization.**

***Index Terms*— historical archive, visualization, information retrieval, ontology, heuristics**

I. Introduction

The recent progress in the area of digital libraries and the semantic web has lead to new ways of digitizing, organizing and presenting library material, enhanced with the incorporation of semantics. More and more organizations, libraries and document repositories opt for digitizing their material, either for internal use or for publishing it through the web. The great variety of digitized material has brought new needs and several research issues have arisen.

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 (at least for some portion of the material), 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 [3]. 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 [4] [7], as it is typically compiled by archivists to suit archiving purposes. As a result, even browsing 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. Time-varying information and entity evolution are also frequently encountered in historical archives, and the modeling, processing and visualization of these aspects are not adequately addressed insofar [8] [9].

As HAs constitute a very important source for historical research, in this work we attempt to investigate the historical researchers' needs and propose a set of tools to assist them in their research in digitized archives. These tools are based on an ontology [6], a construct that presents an overview of the domain related to a specific area of interest and may be used for browsing and query refinement. The proposed set of tools enables the user to visualize temporal information, indicating which data varies with the passage of time and enabling users to efficiently track entity evolution along the time axis. A set of heuristics is also made available to assist researchers in locating ontology entities that correspond to different periods of the same real-world entity.

The rest of the paper is organized as follows: Section 2 presents a user study of historian search methods, section 3 proposes a set of tools to support information retrieval and, finally, the last section concludes the paper and outlines future work.

II. A Study of Historian Research Methods

Historians and researchers collect and process historical data in order to document historical facts and produce information connecting them. 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.

Tibbo in [7] presents the preliminary results of a user study concerning the way historians locate primary resource materi-

als 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. Digital library and archive tools made available to historians could be significantly improved by taking into account the methodologies employed by historians when searching within historic information. These methodologies include certain tasks, both regarding location of information and its visualization, which can be facilitated by historic archive tools. . As suggested in [7], there is a lack of user studies on this issue. In order to identify the historic 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 the use of semi-structured interviews with historians. Both these approaches and their results are briefly described in the following sections.

A. *Study of Queries to a Historical Archive*

The set of queries that users have made to an HA, each requesting documents relevant to a specific subject, can provide useful information on the historians' interests in relation with the HA contents and the typical query categories posed to an archive. We analyzed approximately 100 user queries made to the Historical Archive of the University of Athens and grouped them by query topic, as shown in Table I.

TABLE I. TOPICS OF QUERIES MADE TO THE ARCHIVE

Topic	Percentage
Person Biographies	24%
Historical Evolution of Institution/Organization	18%
Ceremonies	15%
General Socio-political issues	13%
Economic issues	10%
Administration of Institution/Organization	8%
Request for artistic or photographic material (photographs of persons, portraits, monuments, etc)	6%
Books	6%

As seen from Table 1, evolution-related queries, either person biographies or institution histories are predominant among the queries, re-enforcing thus the notion that time and entity evolution is of great importance in the context of an archive.

Approximately 32% of the queries had a user-specified time period or time-point in varying granularities to restrict the search; granularity, ranged 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 on April 21st, 1896"). In most queries, however, year-level granularities were used ("Information for the Chair of Physiology of the Medical School from 1931 to 1939"). Therefore, providing support for entity evolution and time-restricted queries would be important in the context of an HA.

B. *Structured Interviews*

A user group of 15 historical researchers was interviewed for gaining insight on how they search for historical information. The user group consisted of 5 men and 10 women. 4 of them are employees of the HA, and 11 are historical researchers who have visited the HA of the Athens University more than 3 times. The participants were chosen to be familiar with digital IR technologies, so as to provide a more complete view on information both on digital and printed sources.

The interview has 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, generic concepts that s/he researches and what is his/her general method to retrieve information both in digital and in printed sources.

The second part of the interview was composed of seven IR tasks. Four of these were typical queries to the HA of the University of Athens, whereas the remaining three were based on queries made to the HA, but transformed to facilitate our task of recording information on how different search types were conducted.

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

1. They identify and isolate keywords on the topic of their research. These keywords are very often entities like persons, places or organizations.
2. In many cases, the scope of the search is limited to a specific time point (date, year, etc) or period.
3. Researchers initially search using a single keyword at a time; afterwards they perform searches combining more than one of the identified keywords, for example name – date, or place – name – date. Finally, researchers use synonyms and derivatives of the keywords (e.g., for the topic "history of the department of Chemistry", besides the word "Chemistry" they would also use the word "chemical"), while they may introduce new concepts they consider related; e.g., for the "Department of Chemistry", they would introduce "study programme", "professor" or "book". Related terms may be derived from generic or more specialized concepts relevant to the initial ones (e.g. for "Undergraduate Student" they may use "Student" or "PhD Student"); 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 – see Fig. 1). Introduction of synonyms and related terms is referred to as "search enrichment process"; synonyms and related terms can be combined with the initial search keywords.

It should be noted here that the process of enrichment of the initial terms with related ones, as derived from the study of the hypothetical information retrieval tasks proposed to the interviewed historians, is in accordance with the model of the

human mental lexicon described in [6]. It is suggested that concepts in our brain are represented in a semantic network of words, as in Fig. 1. 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.

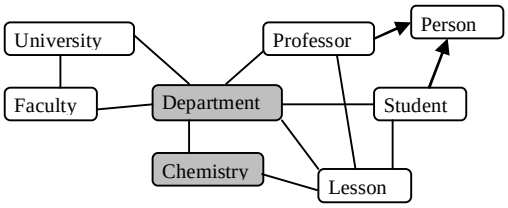

Fig. 1. Example of a semantic net, derived from a researcher's answer for the query "History of the Department of Chemistry". Note the generalization relation for "Student" and "Professor", as she explained that after looking for "Students" or "Professors" she would search for other "Persons" related to the specific department.

Therefore, in this case it seems that an ontological structure, which in fact models the creation of a semantic network incorporating taxonomies, is highly suitable for storing metadata for primary historical sources, while it can also be used as a repository for historical information, both for research and educational purposes. These results are in accordance with the ones reported in [16].

Besides requirements for information modeling and processing, the semi-structured interviews provide insight on the requirements for the visualization of information and the user interface in general. The user should be able to easily limit the scope of the visualized information to a designated period or even time point; concept hierarchy should be presented clearly and the ability to move upwards and downwards this hierarchy should be also provided. Synonyms for concepts in the domain of discourse should be also made readily available to users. Finally, since entity evolution is an important task in historical archive research, special provisions for designating time-varying data and visualizing entity evolution chains should be made.

III. A Tool Set to Support Historical Research

Taking into account the digital HA characteristics and historian needs, we propose a set of tools to aid historical research in the context of an HA. These tools are based on the ontology of the organization the material of which the archive contains.

An ontology was chosen as the means for organizing knowledge since it offers a number of advantages. First, it allows meta-data to be stored separately from the documents themselves (but still linked to them); this was essential in the context of the HA since archive material may either be unavailable in electronic format or available as scanned images only. This metadata may be then exploited in information retrieval tasks, including formulation of initial queries, broadening and narrowing the scope of searches [2] etc. Ontologies may be also correlated with existing categorization schemes [1], making them more understandable and more efficient. Second, ontologies allow for expressing rich semantics in a formal way, facilitating thus both browsing by historians and the operation of heuristics. Finally, recent research has shown that historians and archivists can easily learn and adopt an ontology-based representation of historical information within an archive and perform IR tasks through it [8], [9].

The proposed toolset for supporting historical research consists of (Fig. 2):

- *An ontology*, which is used to organize the knowledge contained in the HA, by classifying, indexing, linking and providing meta-data for the documents.
- A pre-defined set of parameterized heuristics, which may be used by historians to assist their common research tasks. At this stage, emphasis has been put on the task of discovering entity evolution.
- A visualization methods library, providing intuitive and interactive presentations of both knowledge recorded in the ontology and heuristics results.

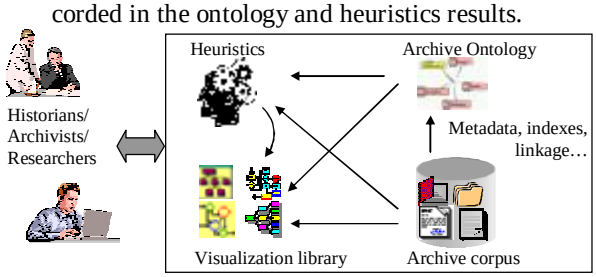

Fig. 2. Toolset components and architecture

The following paragraphs describe the modules in more detail.

A. Historical Archive Ontology

As a first step and with the co-operation of a historian in our team, we created an ontology of the HA of the University of Athens [12], in order to test its efficiency as a tool for historians and use it as a basis for experimenting with the proposed toolset.

The creation of an ontology for an HA is a complex process, since the digitized archive documents are not in text format, making automatic concept extraction impossible, and the concepts that must be captured may vary among different time periods. For compiling the ontology, the user-centric methodological approach presented in [4] was used.

After creating the first complete version of the archive ontology, we prepared an experiment in order to evaluate (a) the researchers' reaction to this representation of historical knowledge and (b) the efficiency of existing ontology visualizations for historical information retrieval tasks. Four different Protégé [10] ontology visualizations were used in this evaluation and the preliminary results may be found in [9]. These results suggested that special attention should be

given to the modeling and visualization of the ontology features related to entity evolution.

Regarding the modeling of an HA ontology, we note here that this ontology is *temporal*, i.e. besides the classes, entities and their properties and relationships, it needs to express the way that these constructs evolved along the time axis. Modeling such an ontology using existing tools was cumbersome and resulted into an excessively complicated model, mainly due to the fact that most environments for ontology management (e.g. [10] [13]) deal with *ontology snapshots*, i.e. they model and manage only the most recent version of ontologies. In order to facilitate the storage and management of time-varying information in an ontology, we extended the Protégé ontology editor, integrating into it data types for historical properties, which are properties capable of storing the different values that this property has been assigned, along with the corresponding time period. For example, the property Name in the University entity can be a historical property and its value for a specific instance may be {(Othonian University, [1837-1911]), (National University, [1912-1922]), (National and Capodistrian University of Athens, [1922-now])} to reflect the changes in the name of the specific University. This provision alone, however, is not sufficient for modeling all possible evolutions since, in some cases, entities have undergone more radical changes, which cannot be modeled through simply recording different values for a slot: for example, a laboratory may have evolved to a museum, which is a completely different entity type. To allow the explicit recording of such evolutions within the ontology, slots "previous" and "next" have been included in the ontology meta-schema, where links to immediate predecessors and successors, respectively, can be stored. The extended version of Protégé is more extensively described in [11] and [8].

B. Visualization

The implemented toolset provides users with the following visualization options:

1. Restrict the display to classes, entities and relationships pertaining to a specific time period.
2. Visualize the entity timeline, i.e. the entity's evolution along the time axis. Both the evolution information recorded in the ontology and the evolution heuristics results are accommodated in this feature. More details on evolution detection and visualization are given in the next section.
3. Co-display the timeline of multiple entities; this has proven to be a very useful tool for historians, enabling them to draw conclusions on the interdependencies of historical entities and events.

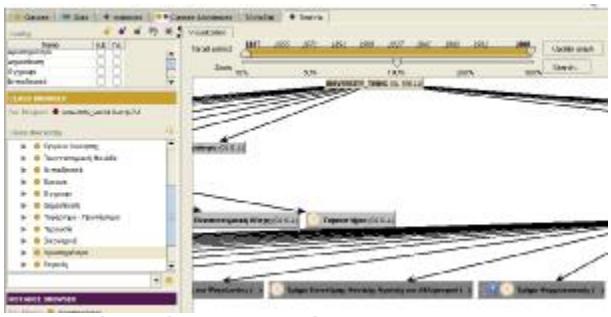

Fig. 3. The visualization main window.

A main result in the ontology visualization experiment described in [9] was that ontology visualization alone is not enough for supporting complex queries. However, a visualization with interactive features, combined with a powerful search mechanism may be very effective. To this end, the Protégé [10] Class Browser, an easy to use and familiar to users indented-list visualization, was combined with a tree layout visualization with enhanced functionality in order to support the temporal characteristics of the ontology as well as the visualization of entity timelines (Fig. 3). Both the indented list and the tree layout visualization paradigms inherently support hierarchies, satisfying thus the user need to organize concepts hierarchically and navigate along hierarchy links, moving from generic to specific concepts or vice versa. Using both paradigms was considered necessary, since no single visualization method could be effective enough and accommodate all requirements. More specifically, as identified in [8] and [9], tree-link visualizations tend to leave a lot of unused space on the top (or left) of the screen and overpopulating the other side, making thus ineffective use of the available space and necessitating a considerable amount of scrolling. On the other hand, the indented list paradigm cannot visualize effectively multiple inheritance, while role relationships (including previous and next in evolution) are only shown in item details panes and not as visual links [8] [9]. Provisions for synchronizing the focus of the two panes are also made.

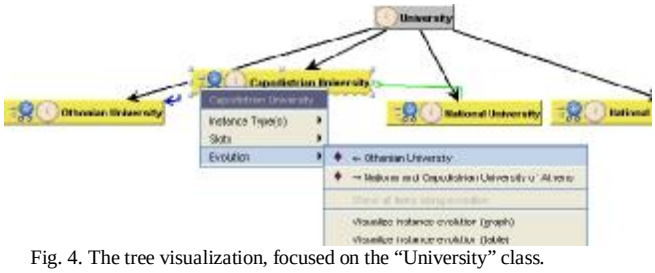

Fig. 4. The tree visualization, focused on the "University" class.

The visualization allows the user to select the classes and instances that will be initially displayed, through a configuration panel. A class or instance containing temporal slots (providing thus the ability to model evolution) is denoted using a static clock, whereas classes or instances that have actually evolved (i.e. they have a temporal property that has changed its value, or they point to a previous or next entity) are marked with a moving clock. Isa-links are always visible (black arrows), while links to the previous/next instances can

be configured to be visible when an entity is selected; having previous/next links always visible has been found to clutter the display and hinder user tasks [9]. Navigation within the graph may be performed either using the graph scrollbars or through the context menu of the selected instance (Fig. 4). The context menu allows the user to view the values of the selected entity's slots, as well as its type and evolution links. By clicking on an instance or class within the context menu (value of a slot, entity type or previous/next item), the clicked instance/class becomes selected within the graph and the graph scrolls to make the selected entity visible. If the entity reachable through some relationship is not currently displayed on the graph, clicking on it automatically leads to its inclusion on the graph and focusing on the selected node, thus the user may dynamically extend the graph as an integral part of the exploration process. Navigation through the context menu has been favorably commented by users in a preliminary evaluation, since it allows them to reach related entities without having to visually search through the graph.

The ability to focus on a specific time period is provided through a slider, through which the user may designate the beginning and end of the period of interest (target period slider in fig. 3). Entities with which are not known to be valid during the selected period (either because their validity period has not been set or because it does not overlap with the period of interest) can be configured to be either grayed-out or to be not included in the graph.

The prototype also incorporates a search facility (fig. 5), which allows users to locate entities whose content matches given keywords. The search panel's functionality allows again the user to focus the scope of the search to a selected time period; the scope of the search can also be set to either the whole ontology or to the items displayed in the main visualization window. Double-clicking on a search result (or selecting the appropriate function from the context menu) causes the main visualization window to focus on the selected entity.

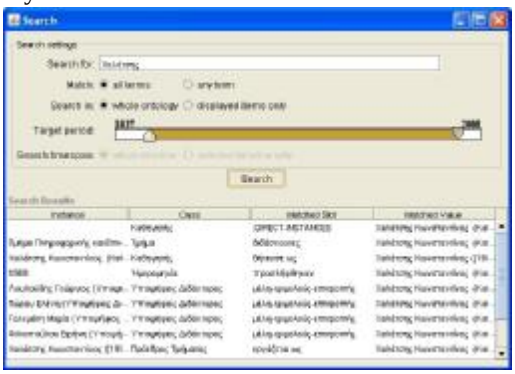

Fig. 5. Search panel.

C. *Detecting and visualizing entity evolution*

An important feature of the plug-in is the visualization of the entity timeline (fig. 6), which is invoked through the context menu (cf. fig. 4). The timeline may include entity splits (e.g. the Othonian University splits into the National University and Capodistrian University) or merging of multiple entities into a single one. Similarly to the visualization window, the entity timeline window includes a slider through which the user may set the period to be visualized. For easier access to time-varying information, the entity timeline window allows the user to display the values of time-varying slots into the entity icon (slot "name" of entity "Othonian University" in fig. 6).

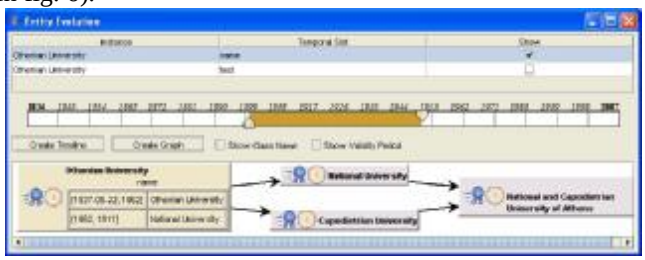

Fig. 6. Entity evolution visualization

The default functionality of the entity evolution visualization window is to visualize evolution information explicitly recorded in the ontology (previous/next slots). One of the tasks of historian researchers is however to correlate facts in order to reach conclusions regarding the evolution of entities; these conclusions may be later recorded in the ontology in the form of facts. To facilitate this task we have implemented a set of heuristics that simulate to some extent the way the historians locate this information. This proposal is complemented with related visualizations, to further assist the user in viewing the entity evolutions and navigating across involved entities. This functionality is integrated with the visualization library and attempts to track entity evolution by exploiting the meta-data embedded in the ontology meta-model.

After the user has selected an entity, the heuristics tool proceeds with locating other instances which may represent the same real world entity (for example a student who later became a professor) (fig. 7), identifying thus possible evolutions. The search is carried out based on a set of pre-defined rules which the user may customize. An example of such rule is checking the similarity [14] of the "Name" slot values of compared instances; this is a key common practice used by historians for locating instances corresponding to the same real-world entity (e.g. person, department etc).

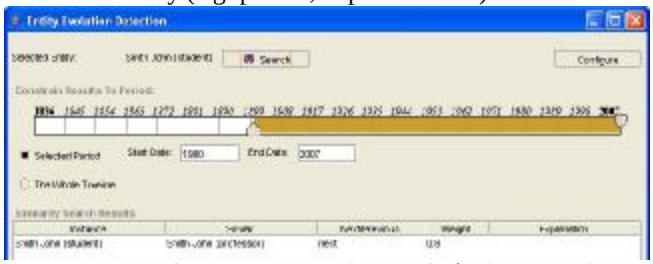

Fig. 7. Entity Evolution Detection Window. Results for the entity "John Smith" are presented.

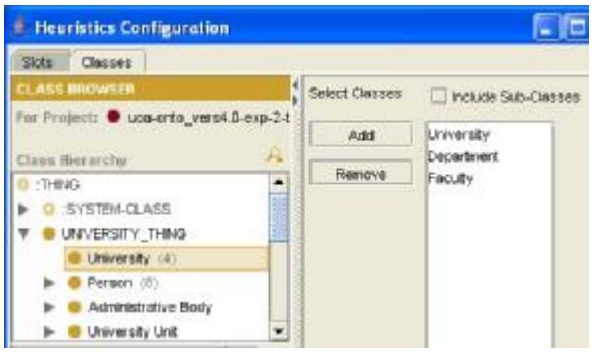

Fig. 8. Heuristics Configuration: Select Classes Pane

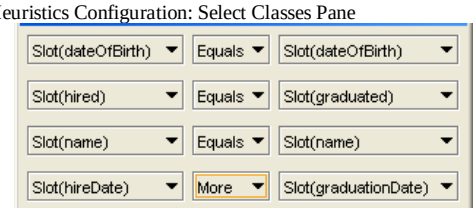

Fig. 9. Heuristics configuration: definition of rules for detecting entity evolution

The user can parameterize the entity evolution detection process through the configuration window (Fig. 8). There the user may designate the top level class (or classes) under which search would be conducted. If the ontology designer has predefined groups of upper-level classes for which evolution among members of their instances is possible, the pertinent groups are retrieved and used as defaults. For example, the evolution group {Laboratory, Museum} indicates that a laboratory may have evolved into a museum (and vice versa). The user may also define rules applied on the slots of the selected entity and those of the instances belonging to the selected classes (Fig. 9) to determine whether a pair of instances actually represents the same real-world entity.

As a next step, the system traverses all instances of the designated classes to locate sets of instances that adhere to the specified rules, i.e. having identical or similar [14] values for specific slots or for which the specified conditions hold. For instance if a search under the class “Person” is conducted, the system could be asked to check the date and place of birth, parent names, etc. Inequality checks can also prove useful, e.g. if the “graduation date” of a “Student” named John Smith is later than the “hire date” of a “Lecturer” John Smith, then there is little probability that the two instances refer in fact to the same person. As another example, consider the case that a specific Geology “Laboratory” ceased to exist as such at a certain date, and some time later a Geology “Museum” was established at the same premises. These known facts can be combined to deduce that the laboratory was transformed into a museum, and may give incentive for further investigation to the researcher. So, for these cases similarity in the name of the instance and the fact that certain dates may be consecutive (*temporal continuity*) or be apart for less than some user-defined threshold (e.g. 2 months), can be in fact strong indications of a relation of succession between the two instances.

For determining similarity between slots, thesauri for synonyms (generic or domain-specific) as well as lexical databases (e.g. Wordnet [15]) are employed. The superconcept/subconcept relationship provided by Wordnet has proven useful in determining whether entities have been merged into a single one or, conversely, split into multiple entities. For example, if a “Natural Science” department ceases to exist and the departments of Biology and Geology emerge, the heuristics can determine that the two newly founded departments are created as specializations of the formerly existing one (which is effectively *split* into two entities), exploiting the information that “Natural Science” is a superconcept of both Biology and Geology.

By applying the heuristics presented above, the system formulates sets of instances that potentially constitute the evolution of a single real-world entity. The relationships between potential predecessors and successors are recorded into slots introduced for this purpose, namely “possible next” and “possible previous”. Each relationship recorded within these slots is tagged with a confidence metric, which is effectively a quantification of the similarity between instances computed by the heuristic.

The user may review the heuristic results, and select items to populate either the next/previous or the probable next/probable previous slots. When heuristic results are displayed, each evolution arc in the visualization depicting heuristic results is labeled with the confidence metric computed for the specific succession.

The prototype plug-in at its current state of development, along with installation instructions may be found in [5].

IV. Conclusions and Future Work

This work proposes a toolset for modeling and visualizing historical information, aiming to assist researchers in studying historic material; the toolset makes special provisions for tracking and visualizing entity evolution, which is an important task for historic researchers. The design and implementation of the toolset has been based on a user study, aiming to record the historians’ information retrieval methods in the contents of an HA and investigate ways to support them. A preliminary evaluation is currently being conducted on the prototype, and full-scale evaluation is being designed, based on queries made to the historical archive in order to assess its effectiveness for supporting historical research and investigate its efficiency in answering complex, time-related queries in comparison with existing ontology browsers.

Issues that are currently being investigated in parallel are the integration of OCR technologies performing word-level recognition for archive documents available only as scanned images, the use of weighted ontologies and spreading activation for assisting information retrieval and the provision of tools to support historians in the creation and maintenance of personal archive of notes and documents.

REFERENCES

[1] Katifori, A., Golemati, M., Lepouras G.: Ontology Aided Information Retrieval in Digital Historical Archives, Proceedings of CSITeA '04, December 27-29, 2004, Cairo, Egypt

[2] Fluit, C., Sabou, M., van Harmelen, F.: Ontology-based Information Visualization, In Visualizing the Semantic Web, Springer Verlag, 2002

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

[4] Torou, E., Katifori, A., Vassilakis, C., Lepouras, G., Halatsis, C., Creating an Historical Archive Ontology: Guidelines and Evaluation, Proceedings of ICDIM 2006, December 06-08, 2006, Bangalore, India

[5] Vassilakis, C., Katifori, A., Torou E., TimeViz: A temporal Ontology Visualization Plugin, http://oceanis.mm.di.uoa.gr/pened/index.php?c=publications#plugins

[6] Noy, N. F., McGuiness, D. L., Ontology Development 101: A Guide to Creating Your First Ontology, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001

[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] Katifori, A., Vassilakis, C., Lepouras, G., Daradimos, I., Halatsis C., Visualizing a Temporally-Enhanced Ontology, proceedings of the ACM AVI 06 Conference

[9] Katifori, A., Torou, E., Halatsis, C., Lepouras, G., and Vassilakis, C., A Comparative Study of Four Ontology Visualization Techniques in Protégé: Experiment Setup and Preliminary Results, proceedings of the IV 06 Conference.

[10] Protégé, http://protege.stanford.edu/

[11] Vassilakis, C., Lepouras, G., Katifori, A., t-Protégé – A Temporal Extension for Protégé, Technical Report TR-SSDBL-06-001, June 2006, available through http://t-protege.uop.gr

[12] Katifori, A., Torou, E., Giannopoulou E., Vassilakis, C., Lepouras G., Athens University Ontology, http://oceanis.mm.di.uoa.gr/pened/index.php?category=publications#ontos

[13] KAON, http://kaon.semanticweb.org/

[14] Hatzivassiloglou, V., Klavans, J., and Eskin, E., Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In proceedings of the Joint SIGDAT EMNLP/VLC, 1999.

[15] Fellbaum , C., (ed), WordNet An Electronic Lexical Database, MIT Press, 1988.

[16] Katifori, A., Torou, E., Vassilakis, C., Lepouras, G., Halatsis, C., and Daradimos, I., Historical Archive Ontologies – Requirements, Modelling and Visualization, in Proceedings of RCIS 07, April 23-26, Ouarzazate, Morroco.

[17] Gazzaniga, M. S., Ivry, R. B., Mangun, G. R., Cognitive Neuroscience, The Biology of the Mind, pp 289-294, W. W. Norton & Company, New Ed edition (April 1998)