# BROWSING NEWS ARCHIVES FROM THE PERSPECTIVE OF HISTORY: THE PAPYRUS BROWSER HISTORIOGRAPHICAL ISSUES VIEW

*M. Platakis, Ch. Nikolaou, A. Katifori, M. Koubarakis, Y. Ioannidis*

{platakis,charnik,vivi,koubarak,yannis}@di.uoa.gr
Department of Informatics and Telecommunications, University of Athens

## ABSTRACT

News Archives constitute an important source for historians, both for research and educational purposes. However, access to their material is not easy due to the special characteristics of the archival content as well as the possible difference between the historian's vocabulary and that of the Archive. In the context of the Papyrus EU-funded project, the requirements of historians have been investigated and taken into account for the creation of a specialized web-based tool, the Papyrus Browser. This paper focuses on the description of the requirements that lead to the design of this tool and provides a detailed description of its main view, the Historiographical Issues View. Design and implementation issues are discussed, as well as plans for future work on the tool.

## 1. INTRODUCTION

News Archives constitute an important source for historical research. The News not only contain a recording of events, as they take place, but also present opinions and arguments that may lead to important conclusions on the public opinion and social issues and tendencies of certain periods in general. News Archives are a powerful tool in the hands of history researchers and they form the primary material for research related to history and social sciences and the humanities. Furthermore, for the education of history there is a particular need for history majors, as well as engineering ones [1] to be acquainted with the process of using archival sources to perform historical research.

However, research in archival material is not an easy task [2] [3] [5]. Historians traditionally had to go through hundreds or even thousands of pages in order to identify relevant information. This was a time-consuming work and as a result researchers could use only part of the material.

Digitization has offered a new means to retrieve the archival material. Many archives today gradually make available their material on-line, in most of the cases in image format. Video, sound and image content formats are also available, apart from text.

Even in digital archives however, there are several difficulties for historians. On one hand, the nature of the textual material makes difficult its full transformation from scanned images to full text. Furthermore, there is a lack in metadata in both textual and multimedia, as keywords given initially by the archivists mostly served archiving needs and did not take into account the need for document retrieval by archive visitors.

Another issue very relevant in the case of archival research is cross-disciplinarity. The News Archive content features its own vocabulary, which may vary within the time span that the archive covers.

Lastly, current digital tools providing access to archive content, like ECHO[1], offer only keyword search and allow simple browsing mainly to collection categorizations.

The aim of the EU-funded Papyrus project[2] is to provide to history researchers, advanced and amateur, the necessary tools to explore archival content contextualized with already existing historical essays. The use case to be implemented in the project is the exploration of News Archives through an ontology representing History of Science and Technology. The Papyrus Browser, which will be presented in the following sections, is the end user visual exploration tool that has been created in order to provide to history researchers of different expertise levels cross-discipline access the multimedia news archived content.

The novelty of the Browser is not only the support for cross-discipline access to ontologies used to annotate multimedia content. It is also effective as a tool for end-user browsing of ontologies, in contrast with existing ontology browsing and editing tools that are created with the ontology designer in mind (see for example web-protégé[3])

The following sections briefly describe user needs (Section 2), functionality (section 3) implementation issues (Section 4) and challenges (Section 5), whereas section 6 concludes the paper and outlines future work.

## 2. HISTORY RESEARCHERS’ NEEDS

[1] http://echo.mpiwg-berlin.mpg.de/home
[2] http://www.ict-papyrus.eu
[3] http://protegewiki.stanford.edu/index.php/WebProtege

The Papyrus ontology browser has been designed after taking careful consideration of the user requirements as recorded in the user requirements analysis performed within the context of the project [8] as well as in prior user needs studies undertaken by our group [4, 5]. According to these requirements, browsing indexes and categorizations of archival material is as important to users as keyword search. It also provides a good general overview of the available material as well as of the vocabulary used for archiving this content.

History researchers proceed in specific steps when attempting to gather the material needed to investigate a specific subject. These steps include (in any order):

- Collecting relevant secondary material, in this case essays of other history researchers on the same or similar subjects
- Collecting primary material, in this case news archive content related to their research subject.

To collect the appropriate secondary material, the users need to formulate their research topic in terms of the historiographical issues and concepts involved, recorded in the vocabulary of historians. Three different groups of ontology concepts are in use in this case [6]:

- General historiographical issues that express history research topics, like “Change in Science”, or “Determinism”.
- History of Science and Technology domains, like “Biotechnology” and “Renewable Energy”.
- Specific concepts that may be the subject of the research, like “Stem cell” or “Wind mill”

To collect the necessary archival content, they need to identify the appropriate vocabulary used in archives to annotate this content.

As a result, history researchers when using digital tools to retrieve archived multimedia content, should be able to effectively bridge their domain, the domain of History, with the archive domain, in our case, the News one.

Ideally, users should be able to start from the history domain, where they could identify relevant concepts and related information. Through them they should be able to move to the news domain where they could review related vocabulary and concepts, and finally, reach the relevant multimedia content, the news items.

After having identified these needs, Papyrus tried to support them with the creation of a specialized browsing tool for bridging the two different domains, History and News.

## 3. FUNCTIONALITY

The heart of Papyrus is its two ontologies, the History and News one [7]. The History ontology models secondary historical material (as available in historical essays) in a hierarchy of concepts important to historical science whereas the news ontology semantically annotates the archived news content. Appropriate mappings define the relations between the two domains, history and news.

The Papyrus browser is a tool that allows the exploration of news content, as expressed by the news ontology, from the History point of view. It is not a simple web-based ontology browser. It is a specialized tool which combines two different domain ontologies as well as the content they describe.

As it was defined early on, the main objective of the browser was to provide the end users the possibility to have both ontologies available at the same window and through them reach the news content. The Papyrus Browser is envisioned to be the tool for researching effortlessly both primary (news ontology and content) and secondary (history ontology) material.

The final design of the browser has been the result of several focus group meetings between history researchers and computer scientists. Early on, the historians stressed the need of being able to transparently follow the connection between historiographical issues and history ontology classes and instances, and from them move to news ontology concepts and instances and then to the news items themselves. (For more information on the history ontology structure refer to [7])

As a result, the following steps have been designed and implemented:

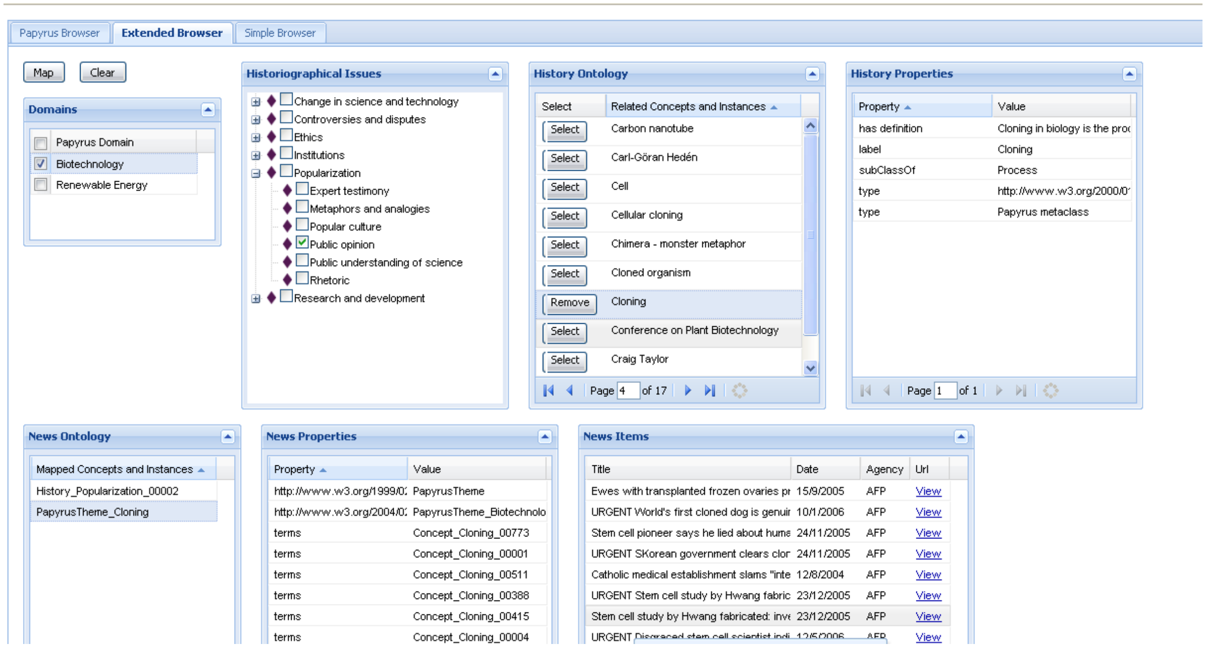
1. Select historiographical issues and domains of interest.
2. Select history ontology classes and instances that are of interest.
3. View related news ontology concepts and instances related to the ones selected in step 1 and 2, and select the ones that are of interest.
4. View news items, the actual multimedia news content, related to the concepts selected in step 3.

The user is presented with all this information at the same browser window. Textual or multimedia information like class and instance descriptions or the news items themselves, are presented in a separate pop-up window.

A general view of the Papyrus browser is presented in Figure 4.1.

This functionality of the browser has been implemented with 7 panels:

**History Domains**. These are the domains that have been selected for the context of Papyrus, Biotechnology and Renewable Energy



**Domains**

- ☐ Papyrus Domain
- ☑ Biotechnology
- ☐ Renewable Energy

**Historiographical Issues**

- ☐ Change in science and technology
- ☐ Controversies and disputes
- ☐ Ethics
- ☐ Institutions
- ☐ Popularization
  - ☐ Expert testimony
  - ☐ Metaphors and analogies
  - ☐ Popular culture
  - ☑ Public opinion
  - ☐ Public understanding of science
  - ☐ Rhetoric
- ☐ Research and development

**History Ontology**

| Select | Related Concepts and Instances |
|---|---|
| Select | Carbon nanotube |
| Select | Carl-Göran Hedén |
| Select | Cell |
| Select | Cellular cloning |
| Select | Chimera - monster metaphor |
| Select | Cloned organism |
| Remove | Cloning |
| Select | Conference on Plant Biotechnology |
| Select | Craig Taylor |

Page 4 of 17

**History Properties**

| Property | Value |
|---|---|
| has definition | Cloning in biology is the proc |
| label | Cloning |
| subClassOf | Process |
| type | http://www.w3.org/2000/0 |
| type | Papyrus metaclass |

Page 1 of 1

**News Ontology**

| Mapped Concepts and Instances |
|---|
| History_Popularization_00002 |
| PapyrusTheme_Cloning |

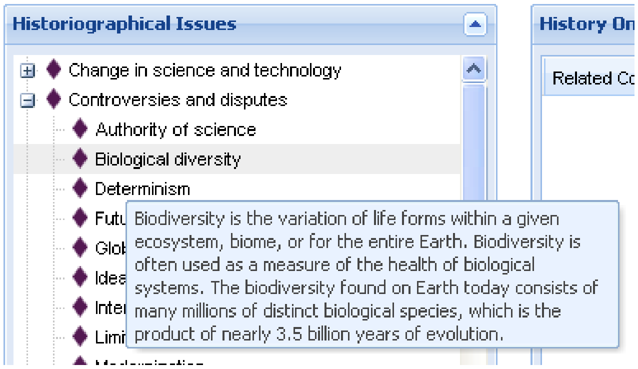
**News Properties**

| Property | Value |
|---|---|
| http://www.w3.org/1999/0 | PapyrusTheme |
| http://www.w3.org/2004/0 | PapyrusTheme_Biotechnolo |
| terms | Concept_Cloning_00773 |
| terms | Concept_Cloning_00001 |
| terms | Concept_Cloning_00511 |
| terms | Concept_Cloning_00388 |
| terms | Concept_Cloning_00415 |
| terms | Concept_Cloning_00004 |

**News Items**

| Title | Date | Agency | Url |
|---|---|---|---|
| Ewes with transplanted frozen ovaries pr | 15/9/2005 | AFP | View |
| URGENT World's first cloned dog is genuir | 10/1/2006 | AFP | View |
| Stem cell pioneer says he lied about hum | 24/11/2005 | AFP | View |
| URGENT SKorean government clears clor | 24/11/2005 | AFP | View |
| Catholic medical establishment slams "inte | 12/8/2004 | AFP | View |
| URGENT Stem cell study by Hwang fabric | 23/12/2005 | AFP | View |
| Stem cell study by Hwang fabricated: inv | 23/12/2005 | AFP | View |
| URGENT Disgraced stem cell scientist indi | 12/5/2006 | AFP | View |

**Figure 4-1 General view of the Papyrus Browser**



- Change in science and technology
- Controversies and disputes
  - Authority of science
  - Biological diversity
  - Determinism
  - Futu
  - Glob
  - Idea
  - Inte
  - Limi

Biodiversity is the variation of life forms within a given ecosystem, biome, or for the entire Earth. Biodiversity is often used as a measure of the health of biological systems. The biodiversity found on Earth today consists of many millions of distinct biological species, which is the product of nearly 3.5 billion years of evolution.

**Figure 4-2 Detail of the Historiographical Issues panel. The description of the class appears on mouse over.**

**Historiographical Issues**. This panel presents the hierarchy of the History ontology historiographical issues concepts [7] (Figure 4.2).

**History Ontology**. This panel contains a list of history ontology concepts and instances filtered by the selected domain(s).

**News Ontology**. This panel contains a list of related news ontology concepts and/or instances according to the selected historiographical issues and history ontology classes/instances.

**News Items**. This panel lists the news items that are related to the selected news ontology concepts.

**History and News Properties panels**. These two panels are used to provide more information on the selected classes and instances of the two ontologies.

The user may explore the presented ontology hierarchies and concept lists and select the ones that are of interest. This way, the researcher may move on the same window from secondary historical information, to the primary archival material and thus contextualize the archival research.

## 4. IMPLEMENTATION

The user interface of the Papyrus Ontology Browser is implemented on the Google Web Toolkit (GWT)[4] 1.7. GWT allows developers to quickly build and maintain complex yet highly efficient JavaScript front-end applications in the Java programming language. As already mentioned, most web ontology browsers and editors currently on development also employ GWT. We use remote procedure calls (RPC) to retrieve data from a  Sesame Server[5] playing the role of the Knowledge Base[6]. Being a web application, the browser operates on Apache Tomcat server 6.0.20[7] and has also been tested on Jetty[8] 6.1.11 servlet container.

On the client side, our browser uses many of the rich widgets available from the GWT-Ext[9] 2.0.6 library, such as the panel, the tree panel, the grid panel and more. Eclipse[10]

[4] http://code.google.com/webtoolkit/

[5] http://www.openrdf.org/

[6] More details about the knowledge base can be found in D5.2 section 3.1

[7] http://tomcat.apache.org/index.html

[8] http://www.mortbay.org/jetty/

[9] http://gwt-ext.com/

[10] http://www.eclipse.org/

Ganymede (Version 3.4.2) was the chosen Java development IDE which integrates smoothly with GWT via an appropriate plug-in offered by Google. The development took place under Windows Vista Business SP1. The browser executes smoothly on every latest generation web browser (Mozilla Firefox, Google Chrome, Opera, Internet Explorer, Safari).

## 5. CHALLENGES

Designing and implementing the historiographical issues view of the Papyrus browser has been a challenging process. In the requirements stage a crucial step was to define the boundaries between the two ontology domains, history and news. Historical research combines research in the primary archival sources with the study of existing historical essays. In depth discussions with expert researchers, with experience in education, has lead to the design of the historiographical issues view. This, however remains one of the possible Papyrus Browser views and the challenge is to evaluate it properly as a stand-alone view and in relation to other possible views.

Furthermore, history is a demanding science based on time-consuming and in-depth research of the archival content. History terminology is related to issues of philosophy and it is subject to time and cultural changes, as well as differences in point of view, and, as a result, is not trivial to model and represent.

Making ontologies accessible to even non-computer experts has been another task that required careful design of the user interface. The complexities of a big ontology structure had to be hidden from the end users by extensive use of labels as well as making parts of the ontology invisible.

The current version of the browser does not yet represent concept evolution or multilingualism issues, both of them very important for modeling historical information within the history ontology. The representation of both multilingualism and evolution within the same visualization tool is a challenging task.

## 6. CONCLUSIONS AND FUTURE WORK

The Papyrus Browser presented in this work is a specialized tool designed to support the very important user need for cross-discipline access to archived multimedia content. Furthermore, it is a tool designed with the non-computer expert in mind to be used as an end user tool for semantic web applications.

The Historiographical Issues view of the Papyrus Browser in its current form implements the basic functionality for exploring both the ontologies simultaneously, according to the user requirements as defined by the participating history researcher groups. It is currently being expanded with more advanced functionality. There are already features that it is decided to be implemented but it is foreseen that after the initial evaluation of the tool new ideas will be recorded on its improvement.

The immediate issues to be explored are:

- Specialized visualization of the ontology temporal and evolution characteristics. To this end we plan to take advantage of our existing research on ontology visualization for history ontologies [4].
- Incorporation of ontology keyword search functionality being implemented in the context of Papyrus

The work on the Papyrus Browser will continue throughout the following period in close cooperation with history researchers in order to receive immediate feedback from the users on improvements and additions.

## 7. REFERENCES

[1] A. Katifori, A. Tympas and E. Mergoupi-Savaidou “Making History Courses Relevant and Attractive to Engineering and Sciece Majors by Bringing Archival Research Within Their Reach: The PAPYRUS Initiative,” in *Proceedings of the International Technology, Education and Development Conference (INTED)*, 2009.

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

[3] H. R. Tibbo, “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] A. Katifori, E. Torou, C. Vassilakis, C. Halatsis, “Supporting Research in Historical Archives: Historical Information Visualization and Modeling Requirements”, Proceedings of IV 08

[5] E. Torou, A. Katifori, A., C. Vassilakis, C., G. Lepouras, and C. Halatsis, “Capturing the historical research methodology: an experimental approach”, Proceedings of International Conference of Education, Research and Innovation (ICERI 2009), Madrid, November 16-18, 2009

[6] A. Katifori, A. Tympas and E. Mergoupi-Savaidou, Tradeoffs in seeking to automate historical research in digitized media archives:
Historians of media meeting media informaticians, Media in Transition 2009, Boston Massachusetts, http://web.mit.edu/comm-forum/mit6/papers/Katifori.pdf

[7] Papyrus Deliverable D3.1 – Ontologies for news and historical content, http://www.ict-papyrus.eu/files/Papyrus-D3.2-v1.93.pdf

[8] Papyrus Deliverable D2.2 – User requirements specification, http://www.ict-papyrus.eu/files/Papyrus-D2.2-v02.1.pdf