‘Media in Transition 6: stone and papyrus, storage and transmission’
International Conference
April 24-26, 2009
Massachusetts Institute of Technology

Paper Title:

# ***Tradeoffs in seeking to automate historical research in digitized media archives: Historians of media meeting media informaticians***

Authors:
Akrivi Katifori, Eirini Savaidou, and Aristotle Tympas
University of Athens, Greece
vivi@di.uoa.gr, savaidou@phs.uoa.gr, tympas@phs.uoa.gr

# Abstract

This paper reports on the possibilities and the risks that the team of its authors has encountered while working on *Papyrus*, an interdisciplinary research project of the European Union that brings together historians of science and technology who specialize in media history, and, computers scientists who specialize in the computational technology involved in storing and assessing media archives. The ongoing digitization of media archives (textual and visual), including the European news agency archive that our team specializes in, affects the way historians and other humanists work with these archives. Using the history of science and technology as a test case, we have been experimenting with ways to integrate the availability of these digital archives in the research (and teaching) of a historian's work.

Several of the flows involved in a historian's research in media archives are being affected by the availability of these archives in digital form. In this paper we present our experience with automating the flow between the 'History Ontology' and the 'News Ontology' of Papyrus, meaning the flow between the recording of events as perceived by the journalist and the use of this recording to study history by history researchers. We attempt to identify user needs and current practices related to these flows, both concerning traditional as well as digital archives and use this experience to support them with appropriate digital tools.

Our case study focus on the digital archives of two major news organizations Deutsche Welle and Agence France Press and two media-based historical research themes: research on the history of biology and biotechnology and on the history of the science of climate change and renewable energy.

# Introduction

The ongoing digitization of media archives (textual and visual), including the European news agency archives that our team specializes in, affects the way historians and other humanists work with these archives. Using the history of science and technology as a paradigmatic case, we have been experimenting with ways to integrate the availability of these digital archives in the research (and teaching) of a historian's work.

Papyrus is an interdisciplinary research project of the European Union that brings together historians of science and technology who specialize in media history, journalists and, computers scientists who specialize in the computational technology involved in storing and assessing media archives. It attempts to investigate issues related to the use of media for historical research and to propose ways to support this research. Several of the difficulties involved in a historian's research in media archives are being affected by the inability to access digitized archives in an efficient way, in a way that takes advantages of their availability in digital form.

In this paper we present our experience with producing a relatively automatic flow between the 'History Ontology' and the 'News Ontology' of Papyrus. This is a flow between the questions of the researching historian and the narrative of the journalist who wrote the news text (or directed the news agency audio-video item). Several important tradeoffs seem to be involved in such automation, as a historian may generally substitute research on a substantially larger quantity of archival material for physical access to a portion of this material. Physical access to an archive can definitely bring quality, but, so can the relatively automatic access to quantity of archives, which brings additional contextualization and ongoing refocusing of research within the reach of a historian.

We have tried several cases in order to study the implications of this tradeoff and its material embodiment in the design of a digital media tool for historical research like Papyrus. In this paper we introduce to observations and interpretations regarding comparisons of three media-based researches: research on the history of biology and biotechnology, on the history of the science of climate change and renewable energy, and, research on the history of computing science and technology (a case of media-based history of media).

Papyrus is an EU funded research project that started in March 2008. It intends to provide a dynamic digital library that will usher in user queries in the context of a specific discipline, help to look for content in an area-domain of that discipline and to return the results presented in a way useful and comprehensive to the user. To be able to achieve this, the source content has to be 'understood', which means analysed and modelled according to a domain ontology. The user query also has to be 'understood' and analysed following a model of this different discipline. Correspondences will then have to be found between the model of the source content and the realm of the user knowledge. Finally, the results have to be presented to the users in a useful and comprehensive manner according to their own 'model of understanding'.

In this paper we briefly present the Papyrus users and their needs as well as our attempts to model the two domains, News and History, and our proposed solutions for supporting historical research through them.

## Users

The users themselves are our primary and main focus within the Papyrus project.

We may distinguish four main user groups for the Papyrus platform depending on their authorization level for content access:

- **Administrators**. Administrators have full access rights in the system and are able to manage users and assign rights.
- **News Content Managers**. News Content Managers are authorized users that are responsible for adding new content as well as editing the existing one. Furthermore, when the Papyrus platform is deployed within a News organization, the management of news content will be accomplished through the tools already available for handling the digital material of the organization.
- **Ontology administrators**. Ontology administrators are domain experts authorized to edit the News and History Ontologies. Users with these rights may be history domain experts for the history ontology and journalists and news archivists for the news ontology.

- **End Users**. End users include all those who may take advantage of the Papyrus system for research or recreational learning activities. In respect to the area of use, they may come with an interest in any of the four following areas: Social Sciences and Humanities in general, History (e.g. of science and technology), Journalism (e.g. science and technology journalism) and Science (mathematical, physical and life sciences and engineering). We have also discerned three levels of use: the Advanced/Professional level, the Intermediate level and the Beginner/Amateur level.

Our main user group of interest for the context of Papyrus is the end users, who undertake historical research in the News Archives to satisfy specific science, social or entertainment needs. History researchers study and interpret primary source material, in our case News Archives, in order to compile historical essays.

Social scientists aiming to cover specific needs of the society for the creation of guidelines, very frequently perform extensive historical research.

Students of history courses, especially in the context of preparing term papers need to research specific historical topics but are too inexperienced to directly access primary sources and review them critically.

Also, journalists often have to research past events to locate useful material that could contextualize their news piece and provide a wider perspective.

Lastly, there are a lot of amateur historians that undertake historical research with a lot of dedication purely for entertainment purposes, but also in the context of their profession.

## Specifying and Supporting the Papyrus User Needs

In the context of Papyrus, we undertook a user study in order to capture the user needs of all the aforementioned user groups and record practices and problems of every one.

The user study included the circulation of 83 questionnaires among the communities of professional historians, as well as history students and journalists and also conducting 36 interviews to receive more detailed feedback and ideas on current practices and needs.

As far as the users' domains of interest are concerned, we have chosen Biology/Biotechnology, and Climate Change/Energy/Wind Power. These were found to be really prominent as research themes among interviewed historians.

The user study concluded that there is a strong interest within all the communities that perform historical research for a tool like Papyrus. Users were very positive in the idea of having a tool that would provide access to the primary source material of the digital news archives through related secondary material organized in a structured way. The structure in the secondary material, the historical knowledge is a feature particularly important to students and journalists, as they need to be able to view overviews and gain quick insight on particular research themes.

Papyrus will attempt to support the recorded user needs by providing appropriate semantic representations of the two domains, News and History. For the representations two ontologies will be used, the News and the History ontology and through mappings and correspondences between them the researcher will be able to access the digital archive material not only directly but also through the history ontology which provides context and related information.

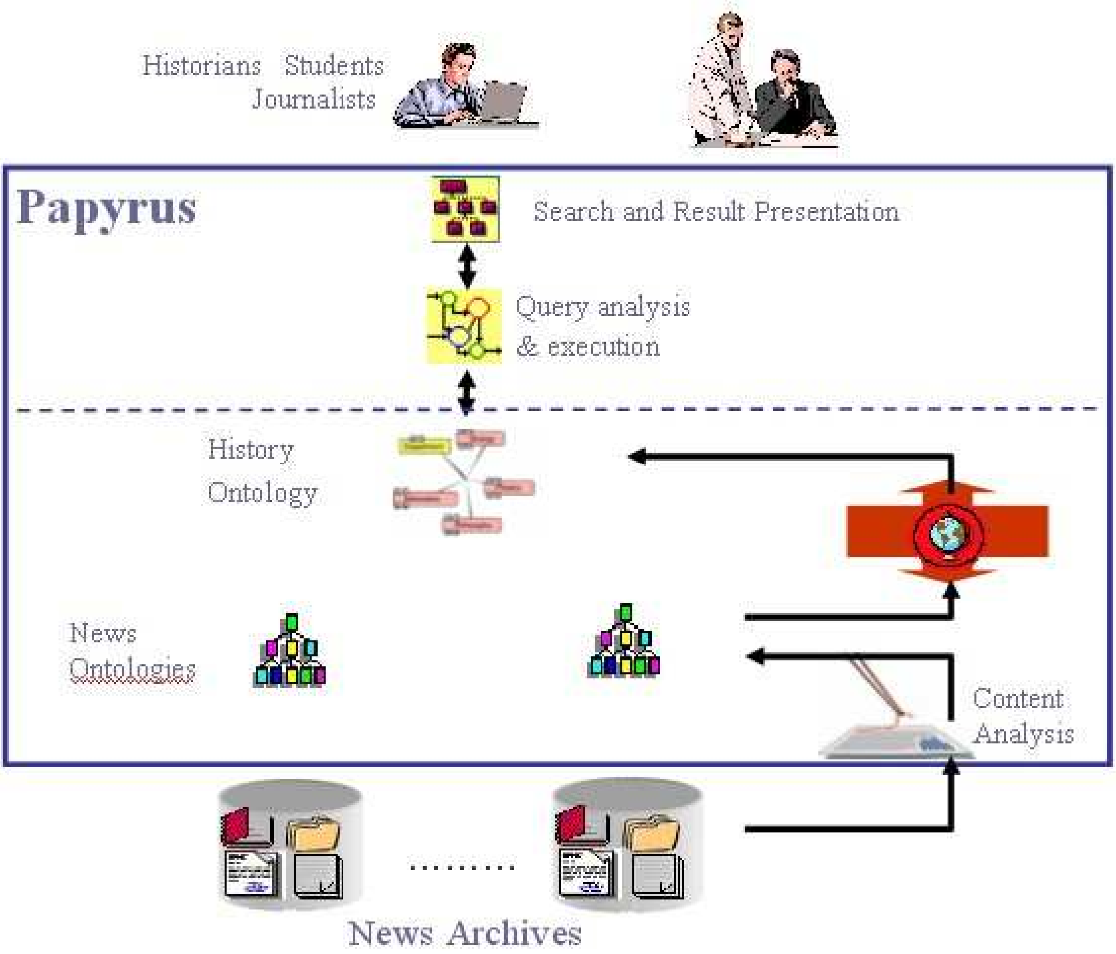

**Fig. 1. Overview of the Papyrus System**

This process may be illustrated by the following representative Papyrus use scenario.

## *Using Papyrus to support historical research in news archives – A user scenario*

Antonia, an advanced researcher on history of science and technology, is preparing an academic paper on the history of cloning. By using Papyrus, she is expecting to enrich her research with information about the social meaning and the discussions on ethics raised by cloning experiments, as well as with information about the way media covered this topic.

1. Antonia enters Papyrus and starts her research by typing the keyword “cloning”. Papyrus returns results about this query in two windows, the one beside the other, which correspond to the history ontology and the news ontology. These results are super-concepts and subconcepts, such as “experiments and experimentation” (super-concept), and “human cloning”, “animal cloning”, “gene cloning”, “therapeutic cloning”, “experimental cloning” etc (subconcepts).

Antonia chooses to see the definitions of these super- and sub-concepts. She can also view an essay on the history of biotechnology, which is provided as a property of the term “Biotechnology, in order to understand to what extent and from which point henceforth has cloning been part of biotechnological research. An essay on the history of a chosen domain (in this case on the history of biotechnology) is an overview article based on a set of articles that include as many of the super-concepts as possible and it includes references to these articles. The articles are also written by specialists in the history of this domain.

2. Antonia asks for related concepts to “cloning”. Papyrus presents her with related concepts from history and news ontologies, grouped according to their super-concepts. For example, related concepts in the history ontology may be “biotechnology”, “genetics”, “genetic engineering”, which are grouped under the super concept “academic disciplines”. In the news ontology, related concepts to “cloning” may be “health”, “treatment”, “therapeutics”, which are also related to the history ontology super concept “research and development’”. Then Antonia can view the groups of concepts and gets an idea of the variety of historical topics cloning is related to.

3. Antonia is interested mostly in ethics and the social impact of cloning experiments, so she can explore more deeply the “science and ethics” group of concepts from the history ontology. She selects some of these concepts, such as “human experimentation” and “bioethics”, in combination with “cloning” and asks to survey news items that deal with the ethical dimensions of cloning.

4. One of the topics that concerns Antonia for her research is modes of popularization that media used in respect to cloning. She selects concepts of the group of popularization in combination with “cloning” and sees that media used “therapeutic” and “experimental” cloning not only to make a language distinction between them but also as a contrasting pair in order to suggest positive or negative connotations of cloning. Antonia inserts “therapeutic cloning and experimental cloning” as a key phrase and asks to see specific news items that include this contrasting pair of concepts, in order to watch more closely the language and conceptual context of media discourse when dealing with cloning.

5. According to the period she focuses on, Antonia refines her search about news items on cloning by defining a specific time period.

6. Antonia views news items of interest using the news content presentation tools offered by P, created to accommodate the different requirements of textual and audiovisual material.

# Tradeoffs in creating and matching the history and news ontology

In order to implement the system that will support the aforementioned scenario, we had firstly to model the core of the system, the two ontologies.

## *Tradeoffs in the design of the Papyrus History Ontology*

The History Ontology is both the point of access of the end users to the primary source material and a means for acquiring information by itself. Its role is also to provide a structured view of the secondary source material. Up to now, there is no history ontology on which the Papyrus project could build, and as a result an effort has been made within Papyrus to create an appropriate general and sufficiently expressive model to represent history-related concepts. The outcomes of a detailed requirements analysis and discussions with the history partners of the consortium have produced a general idea of the main points to be taken into account for ontology modelling.

Through systematic interaction between the Papyrus historians and the informaticians, we have agreed to base the Papyrus History Ontology on two formal classifications of history of technology and history of science subjects. They are offered by the journals Technology and Culture and ISIS, the journals of the Society for the History of Technology [4] and the History of Science Society respectively [5].

For the purposes of Papyrus, we combined the two subject classifications, we selected a set of inclusive subjects, and we clustered them in the following six sets:

1. change in science/technology
2. institutions
3. research and development
4. controversies and disputes
5. popularization
6. ethics

Table 1 includes the full list of the subjects selected and the subject clusters that they contain.

Table 1. General History Ontology Subjects

| |
|---|
| **1. change in science/technology:** change in science, change in technology, environmental history, discipline formation, discovery (in science), artifacts, experiments and experimentation, academic disciplines, scientific communities, professions and professionalization |
| 2. **institutions**, universities and colleges, societies, institutions, academies, (international) congresses, conferences, and meetings, research institutes, research schools, research stations, laboratories, prizes, awards, Nobel Prizes |
| **3. research and development**, technological innovation, impact of technology, technology assessment, public policy, government sponsored science, patents, big science, science and industry, technology and industry, entrepreneurs and entrepreneurship |
| **4. controversies and disputes**, determinism, progress (ideas of), revolutions in science, globalization, modernization, international cooperation, futurism, utopias, authority of science, technocracy, controversies and disputes, political activists, non-governmental organizations, risk assessment, biological diversity, safety, limits of science |
| **5. popularization**, popular culture, rhetoric, metaphors and analogies, public opinion, public understanding of science, expert testimony |
| **6. ethics**, science and ethics, technology and ethics, privacy, private life, interprofessional relations |

These subjects have been organized into a topic categorization to be used as the point of access for the end user in the P System. These subjects are related to sets of concepts that are important to historians. These concepts have been identified by historians and at a later stage structured into an ontology. As a basis for the ontology the CIDOC-CRM has been used and extended.

The CIDOC Conceptual Reference Model (CRM) [6] provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. It includes a proposed model not only for culture heritage objects but also for their history. As a result, it introduces concepts like Actor, Period and Place, all useful to represent the model history. Furthermore, it distinguishes between Conceptual Objects and Physical Things and offers an elaborate structure where a more History-oriented ontology could be built upon.

The process of extending this model has been a laborious one and the Papyrus History Ontology is still a working draft. Several issues have arisen and dealt with, which were mostly related to:

- understanding the needs of historians
- recording important concepts
- organizing these concepts into a meaningful ontological structure

Co-operation between experts in history and in informatics has been essential in order to design and implement the appropriate level of formalism in the ontology that will make it readable by both machines and humans.

Two of the most important points that had to be taken into account during the modelling phase are the following:

**Concepts change with the passage of time**. An important point stressed out by the majority of interviewed historians is the evolution of the terminology in History. A term may undergo conceptual changes in different time periods as well as in different languages and countries. The appropriate modelling of this change is crucial for an effective representation of History in the context of Papyrus. For most Historians the change of a term is a starting point when faced with the study of a particular domain and appropriately modelled it will be very useful for query analysis and enrichment. Concept changes may include:

- Naming. For example “electronic brain” and “computer”
- Semantics/Definitions. A concept may have slightly different definitions in different time periods and places
- Coverage in the press/social impact. The attitude of the public as well as the way a concept is covered in the press may vary in different time periods and different places
- Relations to other concepts. For example, during a certain period a concept may have positive connotations and during another, negative. These connotations may be expressed through relations to other concepts.

Period limits are not exact dates but rather not specific time notations. For example “the beginning of the 80s”, “mid 19th century” or “after the industrial revolution.

The resulting history ontology is general enough to accommodate the modelling needs of most domains of the history of technology and science. For the needs of the Papyrus prototype system, two domains have been chosen as a focus, biotechnology and climate change – energy – wind power.

Once an area-domain was chosen, we undertook research in journals that host articles on the way these technological and scientific areas have been covered by the media. Tens of books and articles were selected for each area, the most relevant being those suggested to us by journals like Science Communication, Public Understanding of Science, Science, Technology and Human Values and Social Studies of Science. An example of the most relevant literature that we were able to identify is that of Miltos Liakopoulos, on the metaphors used in the media coverage of biotechnology [3]. Based on articles like this, we extracted the clusters of concepts-keywords (concepts if seen from the perspective of History Ontology; keywords if seen from the perspective of the news archives content) for each area-domain of technology and science under consideration.

## *Tradeoffs in the design of the Papyrus News Ontology*

The Papyrus News ontology is a tool to be used for news item categorization and retrieval by the News Agencies. As a result, it should conform to existing news ontology standards, with more prominent the one designed by IPTC[1], NewsML-G2. NewsML has been designed by the IPTC (International Press Telecommunications Council) to provide a media-independent, structural framework for multi-media news. The need for NewsML came from the need for better and more consistent ways to structure, describe, manage and associate news content of different media types along their life-cycle, with rapid expansion of the Internet being a strong driving force. At the heart of NewsML is the concept of the news item which can contain various different media – text, photos, graphics, video - together with meta-information that enables the recipient to understand the relationship between components and understand the roles of each component.

A new major version of this standard, named NewsML-G2 [7], has been released in 2008. It is a member of a family of complementary IPTC news exchange format standards - collectively known as G2-Standards, which also offers a standard representation of news events and another for sports results and statistics. This model is made of two parts: a structural model representing news items and news packages, and a basic model of concepts useful for the annotation of general news, e.g. people, organisations and locations. The IPTC Subject News Codes are sets of topics (aka topical subjects) to be assigned as metadata values to news objects like text, photographs, graphics, audio- and video assets.

However, the existing News Ontology was not sufficient for the needs of Papyrus. The history researcher approaches the content of a news agency item from a different angle than the journalist-author of this item. To this end a combination of content analysis techniques and manual modelling taking into account user needs was undertaken, resulting in a richer News Ontology, where named entities have been introduced, as well as important news item concepts. This work has been undertaken for the two main domains that Papyrus will focus on:

- biotechnology
- climate change – energy – wind power

[1] http://www.iptc.org

The result has been an model for a news ontology conforming to aforementioned standards and extended to include relevant concepts to the two selected domains.

## *Mapping history and news domains: the perspective of history*

As explained in the previous sections, the first History Ontology draft resulted from an extensive study of bibliographical sources related on history of technology and science in general and the two selected Papyrus domains and the history of their coverage in the press in particular. On the other hand, the news ontology resulted from manually extending the existing NewsML-G2 model with the results of automatic text analysis on news items of the two domains, including named entities and concepts.

The resulting news ontology is being studied in order to proceed to the next phase, which is the identification of concrete mappings between the two. Results so far suggest that there will be effective and meaningful mappings.

In order to proceed with this, our history partners, as a first step, have constructed tables like the one in Table 2. The next step is to study and record detailed correspondences between the two domains.

Table 2. General and domain specific concepts in the News and History Ontologies for the subject "change in science/technology"

| **General History Ontology Subjects** | **Domain Specific History Ontology Concepts: Biology-Biotechnology** | **General News Ontology Concepts** | **Domain Specific News Ontology Concepts: Biology-Biotechnology** |
|---|---|---|---|
| **change in science/technology**, change in science, change in technology, environmental history, discipline formation, discovery (in science), artifacts, experiments and experimentation, academic disciplines, scientific communities, professions and professionalization | biology, molecular biology, microbiology, biochemistry, medicine, biomedical technology, biotechnology, genetics, evolutionary genetics, genetic engineering, bioinformatics, nanotechnology, etc | Scientific Discipline, Company, Industry, Event, Activity, etc | life sciences, bioengineering, biomedical engineering, biomolecular engineering, molecular science, molecular medicine, biomedicine, genetic medicine, genetics, applied genetics, genomics, regenerative medicine, pharmacogenetics, pharmacogenomics,<br>gene, genome, genetically modified foods/crops/plants, cloning, stem cells,<br>gene therapy, genetic therapy, gene testing, transgenic organisms, gene cloning |

Our aim is to find the best possible way to implement the correspondences between the two ontologies and make them available in the user interface level. If this is accomplished users may have access seamlessly the digital archives through related contextual material that will place the news articles within their historical context.

## Challenges ahead: Beyond text, to video and audio

As part of the Papyrus project, we are working so as to make the History Ontology usable to research beyond textual archives. Our efforts have been so far concentrated on video and audio news agency archives on wind power.

Upon examining a sample of videos which were mostly Deutsche Welle documentaries, we decided to focus on the additional information that we can get from the videos. “Additional” here refers to information that we cannot already extract from the textual archives.

Specialized multimedia analysis techniques are being investigated in order to be able to automatically extract and provide information related to specific historical research issues. These may include image and video analysis to detect the presence or absence of specific patterns and speech segmentation and recognition to detect speaker changes.

Examples of the possible uses of such technologies for supporting historical research on multimedia archives may be the following:

- Differences in the structure of wind generators (e.g., some windmills have a vertical axis structure of blades while others have a horizontal one).
- Actors or lack of them. In the provided videos we see windmills in operation, but, we seldom see the human actors working with them. Also, we never see the end users of wind power.
- Positive or negative metaphors (e.g. conventional plants chimneys positioned next to wind generators or simultaneous use mills of pre-industrial tradition and current new turbines).
- Gender bias. An important issue is the percentage of female and male speakers in videos related to different topics.

These techniques, if implemented successfully, are expected to bring forward new methods of historical research and to affect the approach of researchers towards new technologies. The use of multimedia archives for historical research will become all the more common in the next years when the digital material will become gradually more available.

## References

[1] Aristotle Tympas, ‘Engaging the History of Technology to Understand Emerging Technical Demarcations and Concepts’, *ASEE/IEEE Frontiers in Education Conference, Session ‘Historical Visions: Enhancing Engineering Education*

*through the History of Technology'*, Saratoga Springs, New York, USA, October 2008.

[2] Athushi Akera and William Aspray (editors), Using History to Teach Computer Science and Related Disciplines, Computer Research Association, 2004.

[3] Miltos Liakopoulos, "Pandora's Box or panacea? Using metaphors to create the public representations of biotechnology", *Public Understanding of Science* 11 (2002): 5 – 3.2

[4] ISIS, Technology In Culture, http://www.ou.edu/cas/hsci/isis/website/thesaurus/TechnologyinCulture.A-F.html

[5] ISIS subjects, http://www.ou.edu/cas/hsci/isis/website/thesaurus/IsisCB.Subjects.A-F.html

[6] The CIDOC Conceptual Referente Model, **http://cidoc.ics.forth.gr/**

[7] NewsML-G2, http://www.iptc.org/cms/site/index.html?channel=CH0111