
**Cultural and historical digital libraries dynamically mined from news archives**

# User Requirements Specification

| | |
|---|---|
| **Project Reference No.** | **FP7-215874** |
| **Deliverable No.** | **D2.2** |
| **Workpackage no:** | **WP2: User Requirements Specification** |
| **Nature:** | **R (Report)** |
| **Dissemination Level:** | **PU (Public)** |
| **Document version:** | **2** |
| **Date:** | **16/07/2008** |
| **Editor(s):** | Eirini Mergoupi-Savaidou (NKUA/h), Akrivi Katifori (NKUA/i), Aristorelis Tympas (NKUA/h) |
| **Document description:** | This deliverable contains the analysis of user needs and the detailed user requirements and specifications for the Papyrus platform |

# History

| Version | Date | Reason | Revised by |
|---|---|---|---|
| 01 | 11/07/2008 | First complete version | Akrivi Katifori (NKUA/i), Eirini Mergoupi-Savaidou (NKUA/h), Aristorelis Tympas (NKUA/h) |
| 02 | 16/07/2008 | Second Complete Version | George Kountourakis (ATC), Nikos Sarris (ATC), Akrivi Katifori (NKUA/i), Nadjzeya Kiyavitskaya (UTR), Luisa Mich (UTR) |

# Authors List


| Organisation | Name |
|---|---|
| NKUA/i | Akrivi Katifori, vivi@di.uoa.gr |
| NKUA/h | Eirini Mergoupi-Savaidou, savaidou@phs.uoa.gr |
| NKUA/h | Aristorelis Tympas |

# Table of Contents


# List of Figures

# List of Tables

# List of Abbreviations and Terms

| | |
|---|---|
| AFP | Agence France Presse |
| DW | Deutsche Welle |
| IPTC | International Press Telecommunications Council |
| OCR | Optical Character Recognition |
| RID | Requirement Identification |
| UID | Use Case Identification |

# Executive Summary

This document outlines the analysis and presentation of the user and system requirements for the Papyrus project. All issues related to the Papyrus system and functionality are described in detail in order to provided a concise and detailed recording of the Papyrus use cases.
The process for analysing user requirements was based on interviews and questionnaires based on how cross-discipline research works for the two domains, News and History, identified for the pilot platform. The results were analysed and, along with several interviews in focus groups, used to determine the Papyrus use cases, the functional and non-functional requirements.


# 1. Introduction

One of the critical success factors in the development of a high quality software product is the deep understanding of the user's real requirements, as opposed to their perceived requirements. This is where many projects fail; they do not correctly specify what the system should do.

The user requirements stage is crucial for the development of any application. Especially for innovative tools where the users do not have a clear idea of what they would like the system to do for them, there is the risk of creating a tool that does not meet actual and important user requirements.

User Requirements Capture is the process by which user desires, needs and expectations are gathered in order to establish what the users will actually use the software for and recorded in a way that will be meaningful both to users and developments.

User Requirements Capture may be a difficult process mostly because:

- The developers do not have a clear idea of the users' domain and needs
- The users do not have a clear idea of what the new technology may offer them
- More than one user groups may have to be taken into account for the same application

Once identified, the user requirements effectively lay the foundation for developers, testers, and implementers to begin determining the functionality, responsiveness, and interoperability required of the system.

## 1.1. Definitions

User requirements analysis provides precise descriptions of the content, functionality and quality demanded by prospective users. For the identification of user needs the user perspective must be assumed and result in:

**Functional requirements**

The goals that users wish to accomplish and the tasks they intend to perform with the new software must be determined. By recognising the Functional Requirements, we understand the tasks that involve the abstraction of why the user performs certain activities, what his constraints and preferences are, and how the user would make trade-offs between different products- software applications. The important point to note is that WHAT is wanted is specified, and not HOW it will be delivered.

**Non-functional requirements**

Specification of non-functional requirements includes the categorization of the users (professionals and personal users), the description of user characteristics such as prior knowledge and experiences, the special needs of professional (journalists, editors, etc) and personal users (news audience), subjective preferences, and the description of the users' environment, in which the product or service will be used. Legal issues, intellectual property rights, security and privacy requirements are also an issue.

## 1.2. Methods for Capturing User Requirements

For carrying out the process of requirements identification and analysis a variety of tools were used in a complementary way. These include:

**Scenarios/Use Cases**. Detailed realistic examples of how users may carry out their tasks in a specified context with the future platform. These examples are compiled and refined with the co-operation of the users in order to provide a vivid representation of the envisioned use of the system and provide insight as to the exact methods users employ to accomplish specific tasks.

**User Surveys.** A questionnaire distributed to a sample population of users. Surveys can help determine needs, current work practices and attitudes towards the new system concepts. It provides a way to record the opinion of a large sample of users, but in some cases the answers provided are not as detailed as it would be desired.

**Interviews**. Users are interviewed, in most cases based on a predefined questionnaire. Although the interviews are carried out based on a series of fixed questions, users are prompted to elaborate on their responses as well other related interesting issues that may not be included in the interview questionnaire.

# 2. User Study – Current Practices and Needs

This section presents an analysis of the Papyrus users current practices and needs for four main user categories:

- History researchers
- Students
- Journalists
- Amateurs and other professionals

User research methods have been identified through the use of questionnaires, interviews, as well as group discussions with users. They are presented in detail, focusing more on the main user group, history researchers. Research processes and methods are highlighted through use scenarios.

## 2.1. User Group

Papyrus is a cross-disciplinary platform that will attempt to facilitate access to experts of one domain to content created and archived by experts of another domain. To showcase this idea, the domains of History and News have been selected.

As a result, our main user group of interest for the context of Papyrus is history researchers, who study and interpret primary source material[1], in our case News Archives, in order to compile historical essays. However, not only professional history researchers are actively involved in historical research. Social scientists aiming to cover specific needs of the society for the creation of guidelines, very frequently perform extensive historical research.

To a smaller extent, other groups perform historical research as well. An important one is students of history courses. Especially in the context of preparing term papers, they 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.

In the context of Papyrus, we performed 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 following sections describe in detail the user study method as well as the results.

## 2.2. Method

In order to capture user needs as well as gain a deep understanding of how various user groups perform historical research, we had to employ various means. These included:

- Focus group meetings
- Questionnaires

[1] **Primary source material** is an original source of information that has been created by an authoritative individual with direct knowledge of the fact s/he describes or impresses. On the contrary, **secondary source material** discusses information originally presented elsewhere (primary sources or other secondary sources). For example, an article in the history of stem cells in the journal of History of Biology is a secondary source for a historian who researches the history of stem cells. Primary source material in this case may be references found in news agencies archived items.

- Interviews

### 2.2.1. Focus group meetings

Several group discussions were held in the past few months in order for projects participants to better understand user needs. These meetings were mostly held in Athens as well as in Eindhoven with the participation mostly of researchers from NKUA/i and NKUA/h, but also journalists and personnel from ATC and researchers from UTR.

History researchers as well as participating journalists provided their own point of view for Papyrus and their expectations and suggestions were discussed in detail. The results of these discussions have been taken into account for the analysis of the following sections.

### 2.2.2. Questionnaires and Interviews

To record historical research methods and needs, a questionnaire was created (see Section 6). This was distributed by partners to the personnel of Institutions and organizations, as well as individuals that perform historical research. Among the participating organizations the following may be noted:

- Deutsche Welle
- National and Kapodistrian Univeristy of Athens - Department of Philosophy and History of Science
- Universita Di Bologna
- Deutsche Nationalbibliothek
- Agence France Presse

And more than 20 other universities and organizations.

Based on the same questionnaire, interviews were conducted as well, in order to get more detailed information as to research methods and needs. The total number of questionnaires and interviews per user group is presented in Table 2-1.

**Table 2-1. Questionnaires and interviews conducted for the Papyrus user study.**

| | **Historians** | **Students** | **Journalists** | **Amateurs and other professionals** | **Total** |
|---|---|---|---|---|---|
| **questionnaires** | 38 | 6 | 34 | 5 | 83 |
| **interviews** | 29 | 2 | 3 | 2 | 36 |

## 2.3. Historical Research Methods & Current Practices

This section presents the results of the user study of user practices and research methods in relation with historical research. For each of the user groups, general user practices, ideas and suggestions are presented and general research methods are described, along with user scenarios.

### *2.3.1. History Researchers*

Historical research is a process that historians follow in order to approach and interpret the past by answering new queries or rearticulating queries that have been previously explored. Although each historian employs his/her own method of research, there are also common practices which describe the stages of this practices. This allows us to identify distinct stages of their research. This section focuses on the historians' research method and it provides details on the use of both printed and digital material while it also highlights the research process with a use scenario.

#### 2.3.1.1 Historian User Study Group

For the needs of our user study, 29 history researches from various European Universities and Institutions where interviewed and 38 filled in questionnaires.

This user group is comprised by 33 males and 34 females with the following age distribution:

| **Age Range:** | 18-25 | 26-35 | 36-50 | 51- |
|---|---|---|---|---|
| **Number of Historians:** | 3 | 24 | 23 | 17 |

Their educational level was the following:

- Bachelor: 4
- Master: 6
- PhD Student: 12
- PhD holder: 45

The languages used when performing research are presented in Table 2-2.

**Table 2-2. Languages used by historians when performing research.**

| Language | User Percentage |
|---|---|
| **English** | 97% |
| **French** | 57% |
| **German** | 45% |
| **Italian** | 25% |
| **Greek** | 24% |
| **Dutch** | 22% |
| **Spanish** | 16% |
| **Russian** | 9% |
| **Swedish** | 4% |
| **Turkish** | 4% |
| **Other** | 13% |

It is interesting to note that 88% of the researchers perform multilingual research. The number of languages used in their research is presented in Table 2-3

**Table 2-3. Number of languages used for research**

| Number of Languages | User Percentage |
|---|---|
| **1** | 12% |
| **2** | 12% |
| **3** | 31% |
| **4** | 28% |
| **5** | 12% |
| **6** | 3% |
| **7** | 1.5% |

As we may note, the use of 3-4 languages is very common for the majority of history researchers.

As to computer skills, 97% of the researchers have more than 5 years of experience with using a computer and 95% feel comfortable with its use. As to the time they spend in front of the computer and using the Web, Table 2-4 summarizes the results.

**Table 2-4. Historians use of the computer and the Web**

| | Under 1 hour | 1-3 hours | 4-6 hours | More than 6 hours |
|---|---|---|---|---|
| **Time spent in front of the computer per day** | 0% | 12% | 24% | 57% |
| **Time spent in the Web** | 12% | 51% | 21% | 16% |

As one of the researchers pointed out, the use of the Web depends on the current phase the research is. For example, when in the writing phase of a research paper, the Web is used less than when looking for material on a new research topic.

#### 2.3.1.2 Historical Research Process

Briefly, the historical research on a topic that would involve primary and secondary source material could proceed as following:

**Locating and reading the relevant secondary bibliography for contextualizing the topic under research**

First of all, historians start their research because they have one or more particular queries to answer. Since they need to locate and study the primary source material, as well as to contextualize it with the already existing historical knowledge, they are looking for secondary source material. Secondary source material informs about the existence of primary sources and also provides knowledge about the historical context in which they have been produced. The content of the primary sources is never taken for granted by historians unless it is corroborated by other evidence. Part of the historians' job is to recognize the connotations, the rhetoric, the personal views and biases hidden in the primary source material. Thus, they look for valid secondary sources in order to interpret the primary sources.

The broad diffusion of the web in recent years has enabled historians to search for their secondary sources in the websites of educational and cultural institutions that happen to provide electronic resources. Therefore, one of the first steps taken by historians is to visit electronic databases, electronic journals and digital libraries in order to locate valid secondary material that is related to the topic they want to study. They usually browse the index of such websites; they also use names of scholars as keywords that are working on the same subject or type keywords in a simple and/or advanced search where the title, the author, the date of publication, the discipline and the subjects seem to be the information most commonly used to get the results they want. When the historians

locate the relevant bibliography of secondary sources, they select some or all of it in order to familiarize themselves with the various perspectives that have been used in order to study the topic they are interested in. Various websites offer also information about historical topics, yet sometimes it is difficult for the historians to immediately know which of those provide expert and verified knowledge.

**Locating and exploring the primary source material**

Based on their own historical knowledge and the information drawn from the secondary bibliography, historians then try to locate their primary source material. They usually visit pubic and private historical archives, newspaper archives, libraries and museums in order to find archival material that is useful for their research. They also look in the web for digital libraries and archives that offer access to relevant primary source material. Once they visit relevant portals, historians usually start to explore them by browsing the index in order to get a first idea about the content. They may also type keywords related to their topic in order to check if relevant primary sources exist in these libraries and archives. Once they make sure that relevant material may be available, they proceed in articulating queries within the material in order to obtain useful results.

**Articulating queries in the primary source material in order to get the relevant results**

Looking into physical archival material is often a chaotic process for researchers, since they need a lot of time and effort in order to locate the relevant information to their topic of interest. Most prefer to look into digitized archival material that has been proved more efficient to historical research. After reviewing the structure of the content of the digital libraries and archives, historians usually type keywords in order to get results about the topic of their interest. In the beginning they type keywords that are related linguistically and conceptually to their research topic, and are usually selected through the broader reading they have made in the secondary literature.

The first search is usually exploratory covering a wide area of the research topic. When they get results from the first search, they usually read some of the texts in order to get more information and identify more keywords that will enrich their inquiry. Subsequently, they try to refine their searches according to specific domains/questions and time periods of their research topics. Seeking to answer particular questions, they use keywords or phrases in combination in order to get a number of returned results that concern each specific question. In order to restrict the time period they search their topic by date.

**Reading the primary source material**

Historians usually read the whole bulk of their primary source material unless there is a time-restriction that forces them to make a selection of documents, usually in respect to the relevancy of their titles. This method, however, is error-prone because titles are often misleading. In the case of digitized primary source material, historians proceed in more refined searches in order to isolate the articles that are more relevant to the conceptual categories they have in mind. Lack of time may sometimes lead historians not to read the whole content of texts attentively. In such cases, they pay attention to the first paragraphs expecting to find a brief overview of the whole text in order to decide if this is worthy to be read. They also go through the text trying to locate specific words of interest, the conceptual context of which may provide them with useful information for their study.

**Collecting and saving the relevant primary sources**

Once they have reached a certain point of selection of primary sources, historians try to store them in their personal archive. In case of printed material, they often take notes, make photocopies, use scanners and/or take photos with digital cameras. In regards to audiovisual material, they use audio and video recording, although such methods are rather rare, since historians use audio and video less comparing to texts. In respect to the digitized material, downloading and saving are very common methods for storing primary sources, since historians want to have access to them independently of the web accessibility. Access to references is very important for the process of historical research, since historians consult their sources all the time in order to base their argumentation on historical evidence. Storing the primary sources is therefore necessary in order for historians to recur to them whenever they need them.

Table 2-5 summarizes the historians' answers on how they collect material for their personal archives.

**Table 2-5. Methods used by historians to save interesting retrieved primary source material**

| | OFTEN | RARELY | NEVER |
|---|---|---|---|
| **Notes** | 85% | 9% | 0% |
| **Photocopies** | 84% | 10% | 1% |
| **Digital Camera** | 40% | 31% | 19% |
| **Downloads** | 69% | 10% | 21% |

Interviews, also an important source of information for historical research are usually recorded in audio and in some cases partially transcribed.

### Organizing the material

Part of the historical research also involves the organization of primary source material in order to be able to study it thoroughly and reference it when compiling the historical essay.

Most historians organize their material either in physical or in digital form. Making files and folders, and using sticking notes to categorize the physical material by date and topic are common methods for historians. Placing it in chronological order especially is very important. Other categorizations include:

- Topic/Theme
- Type (text, images, audio, video)
- Related secondary material

One historian specifically mentioned that he would use in the categorization of material the set of the keywords that he would have articulated to retrieve it, for example: all the material for "scientific responsibility" or "history of cloning".

For the physical organization of material they frequently use photocopies in binders and use sticky notes to categorize it.

Yet, in recent years digital technology came to offer a wide range of better and more efficient tools for the organization and categorization of both physical and digital material. Many historians now construct their own electronic databases where they categorize their primary source material per date, source, author, topic, question etc., and insert information, notes and comments that are searchable electronically by keywords. In addition, such categorization by digital tools offers historians the possibility to easily get statistical results for their research. In case of digitized archives, there is also the possibility of linking electronic database with the primary source material by using hypertext.

Apart from constructing their own electronic databases, historians also use reference managers which provide organization and categorization of the primary and secondary sources in various fields, automatic organization of the references in respect to the style used by each publication and hypertextuality to the content of primary sources themselves. Electronic databases and reference managers offer historians the possibility to survey the variety of information of the primary source material in an organized and efficient way. Chronological categories help historians to see the development of an event, concept, etc. through time, while the thematic categorization, which has been based on the questions each historian has, enables them to look for common patterns they need in order to classify the information in conceptual groups that will help them elaborate their historical argument.

### Writing the essay

After surveying the whole range of the necessary information in the databases and after searching for specific topics and printing the results, historians start thinking about writing their essay.

There is no strict method on how the final product of historical research, the essay, is compiled. Several historians mentioned that, as they write, they continuously go back to the collected material, in order to identify patterns and refine their perspective and theories, as well as extract quotes and specific references.

As a result, the more efficient the organization of the collected information has been the easier it is to structure of the historical argumentation and to elaborate on a perspective.

#### 2.3.1.3 Historian User Scenario

For example, let's consider Catherine, a historian of technology, who is currently doing her PhD on the social impact of the transition from steam to electric urban railways from 1890 to 1910 through Greek daily press, would proceed as following:

1. Catherine would first look for secondary literature in order to contextualize her research question. She would search for bibliography on the Greek daily press, the general history of Greece in the late 19th and early 20th century, infrastructures, electrification, transport means, railways, etc. by using library catalogues, or online academic databases, such as the one of History of Science, Technology and Medicine. Catherine would select the relevant secondary source material and read it in order to get a first idea about her research topic.

2. After reading the secondary literature, Catherine would then proceed in selecting and locating the primary source material. Since her research concerns the social impact of the transition from steam to electric technology in respect to railways through the Greek daily press, she should select some newspapers in order to find news items, opinion articles, parliament reports, illustrations, advertisements, etc. on this topic. This selection would be done according to her research requirements and questions. Thus, for example, if one of the preliminary questions was to study her research topic in three newspapers of different political orientation, Catherine would have known by the secondary bibliography the political stance of each Greek newspaper of the early 20th century, and according to the availability of such material in the physical and digital libraries and archives, she would choose which three to use. To her convenience, she would first check the digitized material that is accessible on line. For the time being, the National Library of Greece has proceeded in the digitization of eight newspapers from 1880s to 1980. She would visit the website of the National Library of Greece and go to the index in order to see which of the digitized newspapers refer to years 1890-1910. If the digitized material did not cover the selected years, or there were only one or two digitized newspapers for this time period, Catherine would visit physical libraries and archives in order to identify the rest of the newspapers to use. Of course some extra criteria should have been taken into consideration for the comparability of newspapers, such as their format, circulation, readability, etc.

3. Supposing that there is at least one digitized newspaper which is applicable for her research topic, she would then proceed to identify the relevant articles. Visiting the digital archive of the National Library of Greece, she would discover that there are two options in order to find articles on the steam and electric urban railway.

First, there is a calendar browser for each newspaper, where the researcher has the ability to browse newspapers by issue. She could then use this calendar day by day and look for articles on steam and electric urban railway by reading each page of the newspaper issues, just as she would do with the newspaper(s) found in printed or microfilm form in a physical library.

The other option provided by the website of the National Library of Greece is a search engine that employs the OCR (Optical Character Recognition) tool in order to search for words within the PDF images. She would use this tool typing keywords such as "railways", "steam railways", "electric railways" "electrification", etc. in order to find relevant articles. Because most historians are suspicious of machines doing their work, and because often such search tools do not work properly, Catherine would probably like to check whether the OCR tool is reliable. She would thus choose to read the whole newspaper issues for a particular time period and then make a search with the OCR tool for the specific time period typing keywords found in the articles she read, in order to see whether the results of each method are the same. The outcome of such a trial would be taken into account in respect to her preliminary requirements. If she would not like to miss a thing and found that the OCR tool misses

some cases, she would proceed in reading the whole newspapers in order to find the relevant material. If she would like to get a number of relevant articles as case studies and the OCR returned a satisfactory number of results, she would use it. If, on the other hand, Catherine had chosen to employ statistics, for example getting 5 articles per month, she could use OCR and according to the number of returned results, she could enrich her sample by reading some newspaper issues in order to cover the number of articles needed.

4. Once she would have collected the primary source material, i.e. the three newspapers' articles on steam and electric railways, she would start reading them. By reading this material she would perhaps identify new keywords that could expand her search.

For example, at that period, the urban railway of Athens had the euphemistic nickname "Therion" ("Beast" in English), that was often used by journalists when referring to it. Getting this information by the primary source material, the historian would search also for articles containing this word. It is possible, though, that some of the returned results from such search might be irrelevant, since the word "therion" could also be found in articles referring to actual beasts. Thus, by reading the returned articles, she would isolate those related to her topic. As she would continue her reading, she would perhaps identify more linguistic or conceptual correlations that, if taken into account, could lead her to extend or refine his search in the primary source material and get more relevant results valuable for her research.

She could also get information that would enrich her research questions. For example, if the historian typed the keyword "transportation means", in order to get results for urban railways, she would discover that the returned results do not include such topics. By reading some of the articles on steam and electric railways, she would identify that in the early 20th century, transportation means were mentioned as "communication means". By getting this information about conceptual change in terminology, Catherine could refine her search and enrich her historical interpretation.

5. When reading the primary source material, she would download and save in his personal archive whichever article of the digitized newspapers would be useful for elaborating her argumentation. Similarly, she would photocopy, scan or take photos with a digital camera of the useful articles of the newspaper found in printed or microfilm form.

6. Once she has collected her primary source material to use for her research, she would proceed in organizing it into physical and digital files and folders, electronic databases and reference managers. These methods would help her categorize the material into particular topics, in order to recur to it easily or in order to start structuring her argumentation, since some of the topics could correspond to her historical questions she would like to answer. For example, she would make folders for the articles of each newspaper, or of each year of the two decades, or of each technology (steam and electric railway). Furthermore, she could make different folders for different type of articles, such as opinion articles, news items, advertisements, etc. Or she could make different folders for articles involving different kind of discourses, such as articles negotiating technological and scientific issues, articles on modernity, articles comparing the status of these technologies to other European countries, etc. All these topics could correspond to historian's particular research questions.

Thus, organizing articles according to such topics, the historian could elaborate her questions in a more efficient way by having all her relevant primary sources concentrated. As long as the process of reading and searching keep on and the information taken enriches her ideas about topics, the historian would refine the organization of the primary source material according to her new and more elaborated questions. Once a good point of organization reached, the historian would start writing, although the processes of searching, organizing and writing could keep on going together.

#### 2.3.1.4 Use of primary and secondary sources

Questionnaires and interviews have produced specific statistical data for certain aspects of historical research related to the sources employed.

All historians stated that they employ both primary and secondary sources for their research. Results are presented in Table 2-6.

**Table 2-6. Historians' use of primary and secondary sources.**

| Use of sources | ALWAYS | OFTEN | NEVER |
|---|---|---|---|
| **Primary** | 67% | 31% | 0% |
| **Secondary** | 85% | 13% | 0% |

As to where the historians find their material, results in Table 2-7 suggest that libraries are the most common choices along with Historical Archives and Newspaper archives to a somewhat smaller extent. All other sources are not so generally employed. It seems that they are used for researching very specific topics.

**Table 2-7. Archives and institutions visited by historians when performing research**

| Places where research material is acquired | OFTEN | RARELY | NEVER |
|---|---|---|---|
| **Historical Archives** | 78% | 13% | 4% |
| **Newspaper Archives** | 43% | 40% | 15% |
| **Radio Archives** | 3% | 22% | 66% |
| **TV Archives** | 7% | 30% | 54% |
| **Libraries** | 91% | 7% | 0% |
| **Museums** | 28% | 51% | 13% |
| **Other institutions (organization archives, societies, etc)** | 21% | 18% | 61% |

The results concerning the form and the type of the material they employ when researching, the results are presented in Table 2-8. Textual material is by far the most commonly used both in printed and digital form, followed by images. Apart from Text, which is slightly more often used in printed form, all other material types are used preferably in digital form.

**Table 2-8. Comparative use by historians of material of various types in physical and digital form.**

| | OFTEN | | NEVER | | NEVER | |
|---|---|---|---|---|---|---|
| | physical | digital | physical | digital | physical | digital |
| **Text** | 97% | 90% | 3% | 7% | 0% | 1% |
| **Images** | 36% | 73% | 52% | 18% | 9% | 9% |
| **Audio** | 6% | 19% | 36% | 27% | 49% | 46% |
| **Video** | 7% | 21% | 34% | 34% | 52% | 39% |

The more details on the use of textual material are provided in Table 2-9. Books and journals are the most common textual material employed by historians. Both printed and digital material seems to be employed in comparable measure, with the digital one slightly less used in comparison with the printed one.

**Table 2-9. Comparative use by historians of different textual material types in digital and printed form.**

| | OFTEN | | RARELY | | NEVER | |
|---|---|---|---|---|---|---|
| | **physical** | **digital** | **physical** | **digital** | **physical** | **digital** |
| **Books** | 96% | 43% | 2% | 37% | 2% | 15% |
| **Journals** | 88% | 70% | 6% | 19% | 2% | 5% |
| **Newspapers** | 55% | 54% | 33% | 31% | 8% | 10% |
| **Encyclopedias / Dictionaries** | 52% | 61% | 39% | 24% | 5% | 5% |
| **Other (manuscripts, notes, etc)** | 32% | 30% | 15% | 21% | 53% | 49% |
| **Correspondence[2]** | 13% | 3% | 2% | 0% | | |
| **Reports[1]** | 9% | 7% | 0% | 0% | | |
| **Websites** | | 95% | | 2% | | 3% |

The following section presents in more detail the method and use of digital tools for historical research.

#### 2.3.1.5 Use of digital tools for historical research

Digital libraries and archives are becoming a rather common and useful tool for the historians, since they offer easy access to primary and secondary source material that is either difficult to find or fragile to use. Traditional historical research through physical (manuscript, microfilm, analogue) primary and secondary source material has been complemented in recent years by the use of digitized material. In most cases this is available through the web, thereby offering historians a wide range of facilities in accessing, surveying and storing it. Various academic and cultural institutions have proceeded in the creation of digital libraries and archives that are accessible through the web, so as to enable historians to conduct more easily and efficiently their research for the purpose of contributing to the advancement of historical research in general.

Digital libraries and archives have almost eliminated one of the main restrictions in historical research, which is the physical distance between the scholar and the information. Inasmuch computer networks are widespread and in use in houses, institutions, offices, etc., researchers from all over the world may have easier access to digital libraries and archives through the web, where they can find and explore their desirable primary source material, without needing to actually visit the institution where it is held. Given the variety of sources needed for historical research, such an option saves historians time and money, since there is no need for moving from place to place in order to locate material that is usually spread in various places and institutions. In addition to wide, easy and inexpensive access to the digital archival material, the use of the web provides also 24/7 access to digital libraries and archives.

Historians, especially those working on modern history, may employ various types of archival material such as texts, images, audio and video, in order to enrich their research. The variety of research material offers historians the opportunity to get a better idea of the time period or the subject they work on. They can thus come closer to reality and reconstruct the past in a more complete and

[2] Correspondence and reports were among the prominent choices included in the “other” material. Percentages for NEVER are not included as the lack of mention of these two options does not necessarily mean that historians do not use them but they may be attributed to the users not noting down this option in the questionnaire.

comprehensive way. Therefore, it is crucial for historians that, through digitization and availability of various types of primary source material, they have the ability to study them simultaneously and interchangeably improving the quality of their research and perhaps broadening the potentials of their interpretation.

In the context of university related work, historians usually undertake historical research in order to write a monograph or an academic paper, prepare a history course or write essays and dissertations. Central to historical research is the articulation of questions that historians try to answer by employing primary and secondary source material as historical evidence on which they base their argumentation and interpretation. However, dealing with such material is often a hard task, since most of the time it is difficult for researchers to survey and analyze the whole bulk of it in an efficient and convenient way. Search engines and digital tools that apply to electronic resources of digital libraries and archives offer historians the ability to articulate queries by typing relevant keywords and phrases in order to get results that may provide answers to their questions. They can thus gather a bulk of primary or secondary sources in a massive and easy way, and subsequently explore them by topic or chronologically, depending on the query they have articulated to the search engine.

Digital means also offer methods for a more facilitated research within the material. In contrast with the physical material, the content of digitized material can be explored in a more quick and efficient way, once appropriate tools have been developed. For example, the OCR tool enables researchers to search for specific words or phrases within a vast amount of texts, without needing to read –or even go through- the whole content. This method may be proved very useful since it saves time, enables comparative research and concentrates the totality of results –provided that the tool is actually evolved and trustworthy. Other digital means, such as zooming and image processing (adjusting contrast, brightness, sharpness etc), may also facilitate research, since in many cases texts cannot be read because the material is old or damaged.

Historical argumentation cannot but count on references, thus it is very important for historians to access and handle such references in a convenient way that would facilitate their research. Digital systems and tools that have been evolved in recent years offer historians efficient ways to collect, concentrate and store the digitized material in their personal archive, recurring to it whenever they need it. Comparing to physical archival material, the digital one is much more convenient to get stored, since the space it occupies is considerably less, reduced to the space of digital means such as a hard disc, a memory stick, a CD or DVD, etc. Owing to the facilities of collecting, storing and recurring to the digitized archival sources, historians tend to prefer working with them instead of the physical archival material.

One of the most challenging tasks for historians is the organization of their primary and secondary source material. According to traditional methods, historians are used to categorize their archival material in files and folders and take notes on the material itself or in separate documents in order to isolate the information that is most important for their research. The use of digital tools has contributed a lot to the efficient organization and categorization of archival material by enabling historians to deal with it in a better and easy way. Electronic databases offer the possibility to organize historical information chronologically and thematically and easily get statistical results that are useful for historical argumentation. Moreover, reference managers, such as ENDNOTE, categorize the references used by historians in order to write an essay, an academic paper or a book, and may transform automatically their style in respect to each publication's requirements. Such a method is very useful since it brings together references once and for all, saving historians time from converting them every time they need to publish their work. Reference managers enable also researchers to insert notes and keywords, and use hypertext and links to the sources, so that they can easily retrieve information from the primary and secondary source material and explore it in a dynamic and complementary way comparing to one another instantaneously.

As it is evident a growing number of researchers realize the advantages that the use of digital tools may bring to all the steps of the research process, in the context of our study we asked the participants to provide feedback on the main Papyrus objectives and express their own thoughts and suggestions as well. Statistical results on the fixed questions are provided in Table 2-10.

**Table 2-10. Answers provided by historians on the usefulness of suggested tools**

| | YES | MAYBE | NO | I don't know |
|---|---|---|---|---|
| Would you find useful a tool that provides access to primary source historical material through related secondary material? | 80% | 17% | 0% | 3% |
| Would you find useful a tool that presents secondary source historical information through a categorization of concepts | 69% | 25% | 6% | 0% |
| Would you find useful a tool that combines primary source material into a coherent text? | 34% | 28% | 24% | 13% |
| Would you find useful a search tool on audio and video primary source material? | 58% | 27% | 9% | 6% |

The access to the primary source material through related secondary one was found in general very interesting by researchers, as many commented that it would facilitate and accelerate on-line research. Furthermore, they believe that such a tool would help the researcher "save time since in one search attempt he could have results for both primary and secondary sources" as well as to "save the material categorized". As another researcher suggested, it would be useful in order to directly verify the primary source information and to get a better overview of the sources that are related to a specific secondary source.

Most researchers agree on the usefulness of a concept categorization for organized access to secondary sources, especially when these concepts are historical and philosophical notions such as: historical time, modernity, public/ private sphere, nation state, etc. A researcher suggested that especially for scientific concepts it would be very useful to see a link from each concept to its history and their possible different semantics. She felt that a tool that offers a structured view of important concepts would be very useful for interdisciplinary studies. Change in the terminology is also considered very important. The researchers would like to see it appropriately represented in such a categorization.

However, a few of researchers expressed reluctance to accept such a tool. One, for example stated: "I don't like someone else telling me what concepts fit in specific categories. I'd like to do it by myself. I could accept it for some types of research where knowledge is quantitative." Others felt that though useful, especially for journalists, it could be misleading for historians if not carefully constructed: "It is useful for contextualizing a journalistic research, but on what concerns historical research if it provides information according to a specific/restricted historical interpretation or historiographical approach, it could be restricting and misdirect the research."

As a result, the main point to be taken into account here, which directly concerns the Papyrus ontology, is that a concept categorization as a point of access to secondary sources would be useful to get a quick overview and contextualize the research. But these concepts would be useful only if they can guarantee certain completeness. This means that a lot of attention has to be paid to the criteria of categorization and selection.

On the question if they would find useful the automatic creation of texts based on primary sources and how they would imagine such a tool, historians were in general negative. The idea of a tool that creates a coherent historical text automatically was rejected for various reasons.

Firstly, it would be misleading and falsely orient historical research and interpretation. As one historian stated: "Such kind of tool might ignore information that I could find important and interesting." Most researchers felt that it would restrict historians in respect to the articulation of historical queries and that it would provide predefined historical context that would need a lot of work to validate

Furthermore, such a tool would not be scientifically reliable: “A software that would collect and combine historical data and create a "coherent" text with particular criteria and information that are under question would neither be efficient or legitimate as it would necessarily function on preselected criteria and instructions. It is questionable whether this software could respond to standards and demands of all the kinds of academic research, as well as to the standards and demands of various academic and epistemic disciplines in respect to their own types of research.”

In general, historians felt that the interpretation of primary sources is their job and a machine could neither perform it satisfactorily nor do they need it to. They would find a text synthesis tool possibly useful in the case that it provided summaries of primary sources or extracted specific information and combined it in a text through which they could, if needed, easily go back to the primary source: “It could be helpful in factual things like the first mention of a concept (like “relativity theory”), dates, locations with a connection to the source not only for a primary but also for secondary sources. Also it would be nice for overviews/summary. Historical interpretation is not feasible. It is the historian’s work to put things into context. The tools however should make the sources as available and accessible as possible and allow the historian to make his own interpretations”.

The tool could provide, if given a specific query, a collection of texts found in primary and possibly secondary sources, related to the query, always with a link to the related source. Such a tool would be particularly useful for students as several historians suggested.

Moreover, many researchers expressed interest in better access to audiovisual material through search. Although most scholars stated that they do not use this kind of material very often, they felt that this was partly the case due to its low availability. They said that they would spend more time studying this kind of material if easy and reliable access could be guaranteed

#### 2.3.1.6 Searching in Digital Libraries and Archives

Questionnaire and interview results have provided insight as to how historians approach search and browsing in Digital Libraries and Archives. Table 2-11 summarizes results on several actions historians perform when searching

The majority of the historians rarely read the instructions provided for the Digital Library search tools, as they feel these tools are useful only when the interface is not clear enough as to its use and available functionalities.

**Table 2-11. Historians actions related to the use of digital library tools**

| | ALWAYS | RARELY | NEVER |
|---|---|---|---|
| **Read Instructions** | 21% | 57% | 15% |
| **Browse Index** | 52% | 37% | 3% |
| **Search with keywords** | 90% | 5% | 0% |
| **Search with phrases** | 49% | 30% | 10% |
| **Use simple search** | 84% | 10% | 2% |
| **Use advanced search** | 66% | 25% | 0% |
| **Search within results** | 37% | 37% | 12% |

According to the data obtained, 92% of the historians reported that their first step when searching in a digital library is simple keyword search, followed by advanced search and browsing indexes to get more specific results. Whole phrases are used when a whole title of a document is researched, for example.

Advanced search is also employed when metadata like the document author are known or the researcher wishes to constrain the search in specific time periods. One of the users noted that when the digital library has limited material the advanced search becomes useless. On the other hand, for large digital libraries as well as in situations when she knows exactly what she is looking for, she

prefers advanced search. With simple search in some cases she only checks how many results she gets.

Historians reported a variety of problems when searching a Digital Library with a search tool.

**Table 2-12. Problems historians usually encounter when searching in digital repositories**

| Problems | Researchers |
|---|---|
| **Difficulty in browsing** | 31% |
| **Inconsistent/inadequate material categorization** | 42% |
| **Too many irrelevant results** | 52% |
| **Not enough digitized material** | 52% |
| **Limited access to digitized material** | 52% |
| **Inability to accommodate alternative spellings of keywords** | 30% |
| **Inability to accommodate research in multilingual material** | 21% |

The existence of too many irrelevant results when searching seems to be one of the major problems that historians reported along with the lack of digitized material and the insufficient access to it. Some users comment that the great amount of results they sometimes get, most of which in many cases irrelevant or marginally relevant, are distracting and time-consuming, as the researcher may get lost in trying to locate useful material in a great list of irrelevant one and, as a result, ignore useful sources due to lack of time.

As multilingual search is very important to historians, the lack of this functionality when searching was also found important along with the inability to accommodate alternative spellings of keywords.

#### 2.3.1.7 News media and news content as useful sources for historical research

News media have been regarded as rich archival sources for historians and social scientists not only because of the variety and diversity of their content, but also because of their nature as means of information and communication. The content of news media covers a wide range of social, political, economic and cultural topics of daily life, presented through various types of material. For example, in newspapers, from the late 19th century already, one can find political, economic and social news, editorials and opinion articles, parliament reports, official notices, press conferences, interviews, illustrations and political cartoons, commercial advertisements, obituaries, etc., all of which offer valuable information about the past. News media can also be regarded as offering a vivid image of society, since the daily events are described very lively by the journalists, who are contemporary actors without having any future knowledge that could filter their interpretations, as is the case in other types of archival material. They could, thus, be regarded as primary rather than secondary source material, since all accounts have been written or recorded first-hand.[3]

Although news media are being acknowledged as archival material for historical research, historians are sceptical about their use in history. Media studies have paid attention to the double role of news media both as instruments of influence and expression of public opinion. “Public opinion” is usually perceived as a criterion for the understanding of certain (public) views expressed by specific social groups with specific objectives, which also addressed less privileged groups through news media. Therefore, news media are considered to be the main regulator of the way in which public opinion is expressed and integrated in the public sphere.[4] On the other hand, newspapers, radio and TV broadcasters are often biased in their selection, presentation and interpretation of events, since they have traditionally been oriented to specific political directions and ideologies, occasionally serving as organs of political propaganda. Thus, their factual validity is questionable and this is the reason why

[3] For a comprehensive report on the uses of newspapers by scholars, see [1]

[4] The daily press was the first media of mass communication that contributed to the development of public opinion. See [2]

they should be employed with caution as tools for historical research. Yet, historians are well aware that cautiousness should be applied to any kind of historical source [3]. Therefore, there is no obvious reason why news media should not be taken into consideration as sources of material that, if comprehended and interpreted properly, could be engaged in historical research.

The content of news media, and especially that of newspapers, has been mostly used by historians of various fields as a complementary archival material for their historical research. Employing such material, historians of political, social, economic and cultural history can gather information about political and ideological trends, patterns of social attitudes, images of race, class, gender and national identities, urbanization, cultures of consumerism, local communities, minority groups, lives of individuals, etc [1]. Studies in the fields of history and sociology of science have also demonstrated the utility of news media content in studies about the representations and perceptions of science in society. Social scientists, employing both quantitative and qualitative methods, have studied the social representations of science and technology and their changing positions in modern society by using the content of specific newspapers for long time periods.[5] Some historians of science, on the other hand, have studied how newspaper accounts reflect popular perceptions of science, of scientific disciplines and theories, as well as of scientific discovery [1].

A collective project that focuses on various issues concerning science in the daily press has recently developed by historians of science working in the "European periphery". Within the context of the research group STEP (Science and Technology in the European Periphery), historians of science from Greece, Spain, Portugal and Denmark have produced comparative historical work on the public images of science and technology through the daily press in the late 19th and early 20th centuries[6][7][8]. Studying the various types of newspapers' material, from news items and opinion articles to advertisements, as well as the rhetorical discourse employed in them, these research teams have explored various historiographical approaches dealing with the communication, appropriation, popularization and consumerism of science, as well as the role of news media in such processes within local contexts. Furthermore, the historians have emphasized the importance of the daily press as a privileged means of archival material in order to show the crucial role science and technology played in the formation of modern societies, as well as to legitimize its complementary –or even autonomous– use as primary source material for the history of science [9]. According to them, the archival material of news media in general may be proved advisable for the history of science since not only can it provide information on scientific and technological issues, but also reveal the role of social actors, political agendas, public perceptions, rhetoric of journalistic discourse, as well as the role of news media themselves in disseminating scientific and technological knowledge and in forming public images of various disciplines, theories and artefacts.

### *2.3.2. History Students*

History research is not limited to professional historians but to a great degree concerns students who have to take history courses. These may include two general categories of students, students of history or social sciences related departments as well as students of science and technology category. In this section we briefly present practices and needs of these two student categories

#### 2.3.2.1 Students of history and social sciences departments

The first category of students is the one that contains future history researchers. In this case students actually need to be trained not only in history but also in the methods of making history and performing research.

They are assigned with essays in the context of university courses and in some departments a final dissertation is obligatory in order to receive their BA degree. Professors of history departments usually encourage students to write original essays, instead of copying secondary sources for their assignments, and use as much as possible primary source material in order to get engaged with actual historical research. However, locating primary sources and employing them in doing historical

[5] Some representative works in this field have been done in the context of *Public Understanding of Science* project. See, indicatively: [4][5]

research is not an easy task for students since such processes exceed their current expertise and knowledge as researchers. Even if they visit an archive, either printed or digital, they would have problems locating relevant material.

In order to gain further insight as to how history students perform historical research, 2 students have been interviewed and 6 filled in questionnaires. All of them are undergraduate students in the History and Philosophy of Science of the National Kapodistrian University of Athens in Greece. 3 were male and the rest female, all between 18-25 years old, apart from one who was 26-35. All had several years of computer experience and felt comfortable with the use of a computer and all stated that they spend 1-3 hours daily in front of the computer and again 1-3 hours on the Web.

The main languages they use in their research are Greek and English. Moreover, 3 of them have been using French as well.

As seen from Table 2-13 students are base their work more on secondary sources than on primary ones. They most frequently visit libraries to find material for their work, followed by archives, news and history ones as shown in Table 2-14.

**Table 2-13. History Students' use of primary and secondary sources**

| Use pf sources | ALWAYS | OFTEN | NEVER |
|---|---|---|---|
| **Primary** | 38% | 50% | 12% |
| **Secondary** | 50% | 38% | 12% |

**Table 2-14. Archives and institutions visited by history students when performing research**

| Places where research material is acquired | OFTEN | RARELY | NEVER |
|---|---|---|---|
| **Historical Archives** | 75% | 0% | 25% |
| **Newspaper Archives** | 75% | 0% | 25% |
| **Radio Archives** | 0% | 12% | 88% |
| **TV Archives** | 0% | 12% | 88% |
| **Libraries** | 88% | 12% | 0% |
| **Museums** | 25% | 50% | 25% |

Table 2-15 suggests a strong user preference for textual material, followed by images. Digital and textual material seems to be employed in equal measure with slightly greater preference for digital. Audiovisual material in both forms is rarely employed in both forms.

**Table 2-15. Comparative use by history students of material of various types in physical and digital form.**

| | OFTEN | | RARELY | | NEVER | |
|---|---|---|---|---|---|---|
| | physical | digital | physical | digital | physical | digital |
| **Text** | 75% | 100% | 25% | 0% | 0% | 0% |
| **Images** | 50% | 63% | 13% | 38% | 38% | 0% |
| **Audio** | 0% | 0% | 50% | 63% | 50% | 38% |
| **Video** | 0% | 0% | 63% | 50% | 50% | 38% |

As for the textual material, more details are provided in Table 2-16. There seems to be a slight preference for printed material over digital one. Along with books, the most popular choice for retrieving material is undoubtedly the Web. Only journals are more often read on-line.

**Table 2-16. Comparative use by history students of different textual material types in digital and printed form.**

| | OFTEN | | RARELY | | NEVER | |
|---|---|---|---|---|---|---|
| | **physical** | **digital** | **physical** | **digital** | **physical** | **digital** |
| **Books** | 100% | 63% | 0% | 38% | 0% | 0% |
| **Journals** | 63% | 75% | 13% | 13% | 25% | 13% |
| **Newspapers** | 50% | 38% | 25% | 50% | 25% | 13% |
| **Encyclopedias/ Dictionaries** | 75% | 63% | 13% | 38% | 13% | 0% |
| **Websites** | | 100% | | 0% | | 0% |

As to using the search tools of digital libraries and repositories in general, the statistical results on various options are presented in Table 2-17.

Students in general do not read instructions. 63% of the journalists stated that they rarely read the help or instructions for the use of the search tools, 13% often do and 25% never do. Simple search with keywords is used most frequently (88% of cases), whereas advanced search and search within results less (38% for each of them). Using phrases instead of keywords is a popular choice as well.

**Table 2-17. History students actions related to the use of digital library tools**

| | OFTEN | RARELY | NEVER |
|---|---|---|---|
| **Read Instructions** | 12% | 63% | 25% |
| **Browse Index** | 50% | 50% | 0% |
| **Search with keywords** | 88% | 12% | 0% |
| **Search with phrases** | 63% | 37% | 0% |
| **Use simple search** | 88% | 12% | 0% |
| **Use advanced search** | 38% | 50% | 12% |
| **Search within results** | 38% | 50% | 12% |

Table 2-18 presents the answers students provided for the most common problems they face when searching for digital material. Limited access to material was a very important problem for them. Taking into account that as students they lack both the experience to locate repositories with material available on-line and subscriptions to non-free material, it is difficult for them to access relevant material in most cases.

The problem of insufficient material categorization was also important to them. They would like to have better categorization of the information in the form of basic upper level concepts and their gradual specialization of these concepts. They often feel that keyword search is not very reliable and that it produces many irrelevant results. As a result, most of them feel that structured metadata may offer better access to the material.

**Table 2-18. Problems history students usually encounter when searching in digital repositories**

| **Problems in finding digital material** | **Students** |
|---|---|
| **Difficulty in browsing** | 50% |
| **Inconsistent/inadequate material categorization** | 63% |
| **Too many irrelevant results** | 63% |
| **Not enough digitized material** | 25% |
| **Limited access to digitized material** | 63% |
| **Inability to accommodate alternative spellings of keywords** | 50% |
| **Inability to accommodate research in multilingual material** | 0% |

Students were also asked on the perceived usefulness of certain tools to assist historical research. Results are presented in Table 2-19. As already mentioned, a structured categorization of concepts seems very useful to students. A professor suggested that “it could save them time and provide information that they are not familiar with”. Furthermore one student mentioned the need for detailed timelines as well as the presence of links on the concepts of the categorization to more material or details.

Referring to providing access to primary source material through secondary material, they felt it would it would be very helpful as primary source material is “difficult to search and access” and “it may offer a historical context”.

For the automatic creation of historical text they were not so enthusiastic. One commented that “it could be useful in a first level (amateur) since the user could get a first idea about the topic s/he is looking for. But it would be useless in a professional/ academic level, because in order to validate it the researcher should in any case read all the sources”.

**Table 2-19. Answers provided by students on the usefulness of suggested tools**

| | YES | MAYBE | NO | I don't know |
|---|---|---|---|---|
| Would you find useful a tool that provides access to primary source historical material through related secondary material? | 63% | 37% | 0% | 0% |
| Would you find useful a tool that presents secondary source historical information through a categorization of concepts | 75% | 25% | 0% | 0% |
| Would you find useful a tool that combines primary source material into a coherent text? | 25% | 63% | 0% | 12% |
| Would you find useful a search tool on audio and video primary source material? | 50% | 12.5% | 12.5% | 25% |

#### *2.3.2.1.1 History Students research process and scenario*

Let's consider the example of a student of the history of science and technology department who is asked to write an essay on the evolutionary theory-creationism public controversy for the last decade in the United States.

Without Papyrus, the student would go to an online search engine like Google and type the keywords "creationism", "evolutionary theory", "United States", "controversy" in order to find relevant texts to his research topic.

She would perhaps come across websites that discuss such issues, but their historical validity would be rather doubtful.

The student would also visit library catalogues and electronic databases of academic institutions in order to find secondary source material that would be useful for the contextualization of her research topic. Yet, although she could perfectly contextualize her research by reading the secondary bibliography, her essay would lack original argumentation by not employing primary source material.

In the best case, the student would visit the digital archive of newspapers, periodicals or news agencies in order to find some primary sources that would provide evidence for her essay, but probably she would not be able to deal with the bulk of articles, neither could she find correlations among concepts.

In any case, to do a historical research with such kind of material would be rather difficult and time consuming with poor results in the most cases.

### 2.3.2.2 Students of engineering and science departments

Students majoring in engineering and science departments usually have to take several courses in the humanities and the social sciences. The list includes courses in history, sociology, anthropology, philosophy, linguistics, etc. Depending on the department curriculum, a student may have to (or may elect to) take, for example, a general course in history and/or a special course in history. The Department of Informatics and Telecommunications at the National and Kapodistrian University of Athens offers a typical example. As many other high ranking computer science and engineering departments in Europe and the US, it requires that undergraduate students take general history courses (including courses in European and Balkan history) as well as a special course in the history of technology and science, entitled 'History of Computing and Telecommunications'.

An historian of computing-telecommunications who happens to teach such a course usually asks students to write a term paper on the history of computing-telecommunications. Let us, for example, assume that students are asked to write a term paper in the history of artificial intelligence. At present, these students write their assignment by relying only on secondary sources. In the best case scenario, they read a book or articles on the history of artificial intelligence, written by historians. In the worse case scenario, they locate sites with time-tables of the history of the artificial intelligence in the internet and use them to respond to their assignment. It would be unreasonable to expect that engineering and science students would have to undertake archival research in order to respond to this assignment. Such archival research could, however, improve the quality of this term paper dramatically, because students could check what other historians wrote (or the time-lines available on the internet) by their own research.

### *2.3.3. Journalists*

The work of the journalist involves mostly the creation of News by recording current events. In order to proceed with this task, however, the professional in this field often needs to research issues related to a specific event. This research may involve simple things like the background and biographical information of a person to be interviewed or more complex ones like important issues and recent advances in biotechnology, needed, for example, to introduce a news item on a new discovery that lead to a breakthrough in cloning. In some cases, even, when the journalist has to produce a piece, either in text or audiovisual, that will present the history of a specific issue of general interest, s/he has to do research in the history and evolution of the specific domain. Furthermore, although the work of a journalist is mainly to record events, there is also creative work on his/her part on the perspective from which they will be presented as well as on their interpretation and the possible conclusions drawn from them. In this section we briefly present journalist practices when creating News, focusing mostly on the research aspect that may be needed on their part and defines their needs in relation with Papyrus.

#### 2.3.3.1 Journalist User Study Group

In order to gain insight as to how journalists proceed when they wish to retrieve and analyze material on past events, 3 journalists where interviewed and 34 filled in questionnaires. This study has helped us understand the journalist user needs in relation with historical research in the context of their profession. 22 of these journalists work for Deutsche Welle, the German media organization, either for TV or the on-line media and the rest for the news agency AFP (Agence France Presse) and various newspapers or magazines.

This user group is comprised by 29 males and 8 females with the following age distribution:

| **Age Range:** | 18-25 | 26-35 | 36-50 | 51- |
|---|---|---|---|---|
| **Number of Journalists:** | 2 | 10 | 20 | 5 |

The journalist educational level is the following:

- High school: 5
- Bachelor: 10
- Master: 22

34 of the participants stated that they can perform research in English, 24 in German, 8 in French and a few in other languages like Greek, Arabic, Italian, Dutch and Spanish. All of the journalists perform research in more than one language.

As to computer skills, all 37 journalists have more than 5 years of experience with using a computer and they feel comfortable with its use. They spend more than 4 hours in front of the computer and again, several hours in the Web.

#### 2.3.3.2 Journalist Research Methods and Practices

Journalist needs and practices in terms of resources as well as work method are in great part defined by the type of media that publishes their work. Journalists working for printed media usually have important differences in the way they work than the ones working for the Radio or Television.

To produce a piece for a specific event, the resources needed rise significantly when this piece will be audiovisual. Let's consider for example the coverage of an event like a science convention. In terms of resources, in order to produce a news article for a printed media like a newspaper, what is needed is for the journalist to attend the specific event, take notes using a notebook or even a tape-recorder to record an interview to be transcribed later and, in some cases some photos. On the other hand, an audiovisual piece for television requires the use of very expensive equipment for capturing video content. This equipment must be operated by specialized technicians and as a result the cost of shooting footage for a news piece is prohibitive for allowing the television journalist the same freedom to collect material and record events. Television is the most expensive way to convey news and as a result the journalist's time to create the piece is limited in every aspect, both when doing research and when actually shooting the piece.

When asked to provide details on the research method they follow before creating the news items, interviewed journalists expressed a different view towards the research needed in comparison with history researchers.

Firstly, as journalists they do not usually perform an extensive research on archive material and other sources in order to acquire deep knowledge on the presented topic or study deeply past events and historical facts. In most cases they need general and summarized information on the specific topic, in order to quickly get an idea on the perspective they should present the topic and what is the most important information or aspect to offer to the public. As a TV journalist stated, a news item is mostly based on a predefined "script", created after a quick study of material that can be made available relatively easily.

Furthermore, journalists in general are not experts in any particular field. Their role is to record and present current events and in most cases their knowledge on the topics they present is limited. Furthermore, the presentation of events is made, however involuntarily, through a specific perspective, which is partly influenced by their own views and opinions and partly by implicit rules each media organization may have on the perspective through which news are to be presented, for example, its political directions.

On the whole, when a journalist performs archival research related to the topic to be presented, s/he is not so much interested in covering each and every aspect of the "truth", but rather to understand to some extent the topic and possible historical implications. The goal is to locate in as little time as possible, as much relevant information and material in the archive that could be re-used.

Research in a news archive is, on the other hand, a rather common task for a journalist, as past material may be proven very useful when presenting a piece. This material may be needed in cases where there are news related to a past important event (e.g. the terrorist attack of the 11$^{th}$ of September) and the images of the event itself would be useful, when the journalist wants to make a parallelism of a current event with a past one (e.g. a crash in the stock market) or even when s/he wishes to add some scenes from a city street in the 30s when talking about an event that took place at that time.

When assigned with a subject s/he is not very familiar with, the journalist in most cases performs a quick research related to it in order to locate the best places to visit and maybe shoot scenes and the appropriate people to interview. This is increasingly accomplished through research in the Web, as well as asking the advice and directions of colleagues or experts in this particular topic for advice and directions.

If the journalist has access to a repository of archived news items, either through his/her agency or organization or an external one, it is usual to perform searches within it to locate material to be re-used in the new news piece. Journalists may perform these searches either digitally, if this service is provided, or by asking the help of the archive personnel to locate it. As it has been commented, archivists are very effective in locating material related to a specific topic. They have a lot of experience in this task and in some cases it is their responsibility to archive the material as well and as

a result archivists have a deep understanding and a general idea of the available content and its categorization in the archive.

All journalistis stated that they employ both primary and secondary sources for their research (Table 2-20).

**Table 2-20. Journalists' use of primary and secondary sources**

| Use pf sources | ALWAYS | OFTEN | NEVER |
|---|---|---|---|
| **Primary** | 32% | 65% | 3% |
| **Secondary** | 30% | 65% | 3% |

As to where they find their material, results presented in Table 2-21 suggest that Newspaper and TV archives as well as libraries are the most common choices. Radio archives do not seem a prominent choice; however, this is may be due to the fact that only few of the journalists of the sample have worked for Radio. These answers suggest that journalists employ a variety of sources to acquire material for their research.

**Table 2-21. Archives and institutions visited by journalists when performing research**

| Places where research material is acquired | OFTEN | RARELY | NEVER |
|---|---|---|---|
| **Historical Archives** | 11% | 54% | 35% |
| **Newspaper Archives** | 73% | 19% | 8% |
| **Radio Archives** | 14% | 46% | 41% |
| **TV Archives** | 51% | 30% | 19% |
| **Libraries** | 19% | 70% | 11% |
| **Museums** | 3% | 57% | 41% |

As to the form of the material they employ when researching and its type, the results are presented in

Table 2-22. As we may note, journalists seem to employ all material types with textual material being the most prominent type. There is however a notable difference in the use of physical and digital material with the digital one being the most popular option, especially for audiovisual material. This is to be expected, as going through audiovisual material in physical form may include the painstaking process of carrying audio or video tapes to specialized devices, reviewing these tapes and then taking them back to their appropriate place. Furthermore, this kind of material is far less accessible in physical form than it is in digital form.

**Table 2-22. Comparative use by journalists of material of various types in physical and digital form.**

| | OFTEN | | RARELY | | NEVER | |
|---|---|---|---|---|---|---|
| | **physical** | **digital** | **physical** | **digital** | **physical** | **digital** |
| **Text** | 76% | 90% | 21% | 7% | 3% | 0% |
| **Images** | 17% | 80% | 53% | 20% | 30% | 0% |
| **Audio** | 11% | 51% | 42% | 38% | 47% | 11% |
| **Video** | 19% | 64% | 49% | 22% | 32% | 14% |

Additional details on the textual material are provided in Table 2-23. Both printed and digital material seems to be employed in comparable measure. Newspapers seem to be a very popular choice for journalists as a source for material for their research. However, the most popular choice for retrieving material is undoubtedly the Web.

**Table 2-23. Comparative use by journalists of different textual material types in digital and printed form.**

| | OFTEN | | RARELY | | NEVER | |
|---|---|---|---|---|---|---|
| | **physical** | **digital** | **physical** | **digital** | **physical** | **digital** |
| **Books** | 54% | 40% | 39% | 44% | 7% | 16% |
| **Journals** | 49% | 65% | 43% | 27% | 8% | 8% |
| **Newspapers** | 73% | 81% | 27% | 12% | 0% | 7% |
| **Encyclopedias/ Dictionaries** | 57% | 60% | 32% | 22% | 11% | 18% |
| **Websites** | | 83% | | 10% | | 7% |

All journalists answered that they use the Web for their research. They frequent websites like Wikipedia, digital libraries like the US Library of Congress, Press Archives, on-line newspapers, etc. The 90% of the journalists also perform multilingual searches frequently.

Table 2-24 presents the results obtained on the use of various search tools of digital libraries and repositories in general. Reading the instructions, as it may be expected is not something most users usually do. 63% of the journalists stated that they rarely read the help or instructions for the use of the search tools, 20% often do and 17% never do. Whether to read or not depends to a great extent to the simplicity of the interface and whether it inspires confidence to the user that s/he will be able to use it to find the material needed.

**Table 2-24. Journalist actions related to the use of digital library tools**

| | OFTEN | RARELY | NEVER |
|---|---|---|---|
| **Read Instructions** | 20% | 63% | 17% |
| **Browse Index** | 36% | 54% | 13% |
| **Search with keywords** | 90% | 3% | 7% |
| **Search with phrases** | 43% | 40% | 17% |
| **Use simple search** | 93% | 4% | 3% |
| **Use advanced search** | 47% | 47% | 6% |
| **Search within results** | 37% | 47% | 16% |

Simple keyword search appears by far to be the most prominent technique for journalists to look for material. Whole phrases are also used in search but less often.

Browsing indexes to find material is also employed quite often. Its importance has been commented in more detail by some of the journalists. A good index offers a useful overview of the available material, facilitating thus the navigation and the retrieval of relevant content. It provides a quick way to examine the topics covered in the available material which is essential for the assessment of the relevance of the material to the current research topic.

All of the journalists expressed problems when using digital tools to retrieve material for their work. These are summarized in Table 2-25. The issues of digitization and access to digitized material are important for at least half of the journalist in our sample. These problems are expected to be partly

solved as a result of the continuing digitization effort which is currently in progress in the E.U. and by ensuring access to these material for researchers.

**Table 2-25. Problems journalists usually encounter when searching in digital repositories**

| Problems in finding digital material | Journalists |
|---|---|
| **Difficulty in browsing** | 22% |
| **Inconsistent/inadequate material categorization** | 63% |
| **Too many irrelevant results** | 77% |
| **Not enough digitized material** | 40% |
| **Limited access to digitized material** | 53% |
| **Inability to accommodate alternative spellings of keywords** | 20% |
| **Inability to accommodate research in multilingual material** | 47% |

However, the rest of the problems mentioned are related to more complex issues. 63% of the participants reported that they experience problems with inappropriate categorizations of the material, as they may be unclear, inaccurate, too generic or inconsistent. As a result they rarely can really benefit from the advantages that browsing a collection of material may offer. The difficulty in browsing 20% of the journalists reported may also suggest the inadequacy of most interfaces to ensure comfortable and effective navigation in the categorization and the material itself.

Two also very important problems reported are those related to the different forms a term may take within the same language or in different ones. More specifically, a word may have alternative spellings that in most cases are due to grammar changes in the language with the passage of time. Especially in the context of research in archives and material that covers a large time span in general, this issue is very important in order to retrieve as many relevant results as possible. Furthermore, as most of the journalists, especially the non-native English speakers, already stated, in most cases have to perform research in multilingual material. This may become problematic when the journalist is not sure about the translation of a search term to another language or the change in semantics this term has undergone in different countries and languages.

Lastly, the most important problem mentioned by journalists when accessing digital libraries and archives is the great number of irrelevant results that they get when searching. 77% of the users reported this as an issue and it is among the main goals of Papyrus to alleviate it to some extent.

Papyrus proposes to facilitate browsing and searching the material and offer effective presentation ways of the end result. When asked on the usefulness of main goals of Papyrus, journalists were generally positive. The results are summarized in Table 2-26.

**Table 2-26. Answers provided by journalists on the usefulness of suggested tools**

| | YES | MAYBE | NO | I don't know |
|---|---|---|---|---|
| Would you find useful a tool that provides access to primary source historical material through related secondary material? | 57% | 38% | 0% | 5% |
| Would you find useful a tool that presents secondary source historical information through a categorization of concepts | 62% | 35% | 0% | 3% |
| Would you find useful a tool that combines primary source material into a coherent text? | 43% | 41% | 5% | 8% |
| Would you find useful a search tool on audio and video primary source material? | 73% | 22% | 0% | 0% |

More specifically, when asked on the usefulness of a tool that provides access to primary source historical material through related secondary material, 57% of the users answered yes and 40% maybe. Access through structured secondary material to the primary source one is expected to narrow down irrelevant results and provide a "summary" view of the topic in question as well as organized access to the primary source relate to it. However, as one journalist commented, the usefulness of such a tool depends largely upon the quality of the categorization. "After all, my research interest might be completely different from given categories and the tool would thus complicate my research instead of simplifying it." Other users mentioned the importance of such a tool as "it proves often useful but really difficult to retrieve primary sources".

On the question whether they would find useful a tool that presents secondary source historical information through a categorization of concepts, again journalists were positive. 63% answered "yes" and 37% "maybe". As one suggested, structured secondary source material will provide to the journalist a quick overview of the topic in question as well as links to related primary source material if in need to elaborate on this: "Journalists are total generalists and a superficial and quick overview of their topic will be very helpful". Furthermore, they believe that the History ontology will provide a detailed and logical categorization of concepts that will facilitate browsing and retrieval of related primary source material.

The creation of a tool that automatically combines primary source material into a coherent text was faced with skepticism by many journalists. 40% answered "yes", 43% "maybe" and 7% "no". Although one would expect that an ideal tool that automatically selects related primary sources and creates an article on the issue in question would be considered very useful by journalists, this does not seem to be the case. First of all, they cannot easily trust software that "makes up its mind" as to what material to use and what to ignore and they feel that not only it cannot perform this task correctly but also that it would be bad practice to let it do so. They want to reserve for themselves "judging which sources to put into which context". Furthermore, an important obstacle for such a tool to be useful for journalism is the existence of implicit rules of a Media organization as to the character and directions the material they publish or broadcast should have. These rules and tendencies are usually unofficial and very difficult to record.

The journalists' suggestion for a tool that builds text based on related primary source material would be to offer a kind of summary or overview of the material, always with links to the primary sources. It could for example contain the paragraphs in the primary sources found relevant that directly seem to refer to the current query. This text would to some extent simulate the practice of several journalists

to copy text from relevant sources, either electronically or by hand, in their notes and then use it to prepare their news item.

Finally as concerns the importance of facilitating keyword search on audiovisual material, journalists were very enthusiastic, as 83% answered "yes" and 17% "maybe". They commented that this would be very useful because these sources are often less available than textual resources and more difficult to search.

On the way they save, categorize and prepare the material they collect in order to proceed with creating the news piece, results are summarized in Table 2-27.

**Table 2-27. Methods used by journalists to save interesting retrieved primary source material**

| | OFTEN | RARELY | NEVER |
|---|---|---|---|
| **Notes** | 84% | 13% | 3% |
| **Photocopies** | 42% | 48% | 10% |
| **Digital Camera** | 20% | 30% | 50% |
| **Downloads** | 61% | 30% | 9% |

Many of the journalists stated that they keep in their collected material whole paragraphs. In some cases, when they keep material in word documents, they highlight it.

The following sections focus on news creation processes and scenarios in order to illustrate the journalist research methods and present them in the context of their everyday work.

#### 2.3.3.3 Audiovisual news item creation process and user scenario

This section illustrates the creation of news for audiovisual media by presenting the general process of news creation as well as a user scenario of how this process is applied be journalists in Germany's international broadcaster Deutsche Welle.

##### *2.3.3.3.1 Audiovisual news item creation process*

This section presents the general process followed by television journalists in order to create a news item based on new original material as well as old material stored in a News Archive. For the context of Papyrus, focus has been given to the research performed in the news archives for contextualizing the piece to be created. However, the whole news creation process is briefly illustrated, as it is important to understand the journalist's way of work and particular needs.

The creation of a news item is accomplished in the following steps:

1. **Topic selection**. The journalist selects or is assigned a topic to due a news piece.
2. **Research**. The journalist performs research on the given topic in order to collect related information on the given topic. This point will be further analyzed later in this section.
3. **Topic preparation**. The journalist prepares a per-defined "script" of the way the piece will be presented, having a general idea on the archived and new material to be used, as well as to the way they will be combined in the final piece.
4. **New footage creation**. In this step the journalist along with the team of technicians shoots the material in the pre-selected locations. This material may include interviews of key persons for the specific topic, video of specific locations or events and even brief interviews of passers-by.
5. **Editing the piece**. After the new material has been created, it passes the editing phase when it may be combined with old material, possibly retrieved from the archive.
6. **Adding the voice over**. The finished piece is combined with the appropriate sound, voice and sometimes music

7. **Broadcasting the piece**. The finished piece is combined with other pieces and in most cases a presenter and it is broadcasted in the context of a TV programme.
8. **Archiving**. The broadcasted piece is moved to the archive where it remains as long as it is defined by the organization policies. Archiving is performed firstly for legal reasons and secondly to be re-used for future news items if necessary.

#### *2.3.3.3.2 Audiovisual news item creation - Deutsche Welle user scenario*

Germany's international broadcaster Deutsche Welle focuses on people around the world with an interest in Germany and Europe, in particular opinion leaders and the so-called "information elite", providing them with news and information content, inter-cultural communication, and service-oriented information.

At present, about 1,200 employees from more than 60 countries work for Deutsche Welle, making it a true multinational and multicultural organisation. DW maintains external bureaux in Washington, Moscow and Brussels, and co-operates closely with the network of correspondents operated by the ARD and ZDF (Zweites Deutsches Fernsehen – "Second German Television").

Deutsche Welle broadcasts via satellite an up-to-date TV programme that is produced in Berlin. The news and information programme is transmitted around the clock: twelve hours each in German and English, the language changes every hour. In Europe, North, Central and South America, North Africa and the Near and Middle East also programme windows in Spanish (2 hours per day) and Arabic (3 hours per day) are broadcast.

The structure of the programme is as follows: news on the hour, a magazine show or documentation on the half hour. All broadcasts take thirty minutes. A TV journalist working in Deutsche Welle usually has to create 3-5 minute news items of audiovisual content.

Erik is TV journalist in Deutsche Welle, and has lately been assigned to a department that focuses on world-wide economy issues. He has to create a news piece relevant to a talk to be given by a famous economic analyst related to the world-wide consequences of the stock market crash of 1987.

His idea is to present through the presentation of the event of the talk to make a reference to the 1987 crash at the same time making a comparison with the 1929 stock market crash. Before preparing the piece, he decides to search for more information.

As he has not covered an issue related to this event in the past and he doesn't know much about the specific topic, he starts by a quick search in the Web for this issue in order to familiarize himself with the issue and have some insight as to how to proceed. Web search produces results but he is not sure on what to focus exactly and which would be the best perspective to present the events. In order to get advice, he contacts a colleague who has been working in the same department for a longer period of time to get advice. The colleague proposes an economics professor for a brief interview, if the main speaker of the event is not available. Also, he mentions that about 20 years ago he had made a piece on the 1987 stock market crash, which should currently be found in the Deutsche Welle Archives. With this information in mind, Erik performs a search in the Archives through the provided digital tool.

The Deutsche Welle Archives contain videotapes of TV shows broadcasted during the operation of the organization. The news pieces created by the journalists are not kept in individual tapes but rather as they were broadcasted along with the show in the context of which they have been created. Each tape is accompanied by a set of metadata, coming from two different sources, the show producing department and the archive itself.

The producing department metadata are added after the editing of the show. They include metadata for the whole show as well as for the individual audiovisual pieces comprising it. For the whole show date, duration and the specific sequences of films and presenter time are kept, among others, whereas the individual pieces include provenance, whether the piece has been bought or it comes from the Archive, the journalist that created it, possibly information on its content, etc. The metadata added by the archive personnel include content descriptions in the form of keywords. There is also a standard set of keywords as well as a categorization according to the program type.

On the whole, search within the Archive material can be done through an advanced search tool that allows searches by combining:

- Keywords
- Broadcasting period/dates
- Program type

- Person names (presenter, journalist, interviewees)
- Broadcasting length
- Title
- Duration

Erik uses the search tool to locate relevant material. He combines keywords related to the issue, like "stock market crash 1987" as well as the information provided by his colleague on useful existing news items in the archive. Keywords in this case are combined with the journalist name as well as the period when the particular show was broadcasted. His search produces some possibly useful results, including the 1987 news piece on the crash of that period as well as some pieces containing footage from 1929, but he would like something more specific if possible.

To this end, he visits the Deutsche Welle Archive to ask for her help to retrieve more relevant material. The archive personnel, more experienced in this kind of searches, is able to retrieve other relevant news items with footage that could possibly be re-used.

Once he has all the material in his hands as well as a concrete idea as to what people should be interviewed as well as appropriate places to shoot scenes, Erik prepares a first draft of the text for the voice over of the piece, i.e. the piece "script".

The next step is to arrange possible interviews as well as when and where he will shoot scenes and to collect this material. When all the material is collected, he is ready to edit the material in order to create the final piece. The process of interviewing and shooting new material may have changed his initial idea for the news item "script", so he may have to adjust it accordingly. Either himself, or with the help of an editing technician proceeds with the editing and then adds the voice over.

The edited piece is then forwarded along with others that will be broadcasted during a specific show in the studio where the presenter will present them. The whole show is recorded and the tape then goes to the archive.

Currently Deutsche Welle is undergoing an effort to digitize the whole process. Erik would benefit from this change as it would simplify his work. When he is away shooting scenes for a piece, he will be able to directly store in his laptop the audiovisual material, access the archive and collect archive material, and then edit everything into the final piece to be sent to his department to be broadcasted.

The role of Papyrus in this process will be very important in the research step of news creation as it will quickly provide an overview of the historical implications of a specific topic as well as useful insight as to which information should be presented and from what perspective.

#### 2.3.3.4 Textual news item creation process and user scenario

This section presents the general process followed by journalists that write news articles in order to create a news item based on new original material as well as old material stored in a News Archive. For the context of Papyrus, focus has been given to the research performed for contextualizing the piece to be created. However, the whole news creation process is briefly illustrated, as it is important to understand the journalist's way of work and particular needs.

##### *2.3.3.4.1 Textual news item creation process*

The creation of a news item is accomplished in the following steps:

1. **Topic selection**. The journalist selects or is assigned a topic to due a news piece.
2. **Research**. The journalist performs research on the given topic in order to collect related information on the given topic. This point will be further analyzed later in this section.
3. **Material collection**. In some cases the journalist may need to perform on site visits or interviews in order to have a clearer idea of the topic.
4. **Writing the article.** Taking into account the selected material, as well as his/her own views and observations, the journalist compiles the article.

5. **Editing the article**. The article is edited in order to be "polished" and ready for publication.
6. **Article publication**. The finished piece is published in a newspaper or magazine or available for use by either Media if it was created by a News Agency.
7. **Archiving**. The article is stored in the archive where it remains as long as it is defined by the organization policies. Archiving is performed firstly for legal reasons and secondly to be re-used for future news items if necessary.

#### *2.3.3.4.2 Textual news item creation - A user scenario*

Mary is a Greek journalist working for a newspaper. She writes articles for a weekly column presenting science related issues to the wider public. This week she is assigned an article on cloning, as there was an announcement on progress in that field. Her aim is to present the announcement along with a brief overview of the history of cloning and its main research breakthroughs.

Her first step is to perform a web search with "cloning" and other related keywords she can thing of, like DNA, genome, cellular biology, bioethics, etc. She also visits on-line encyclopaedias like wikipedia.

As she still has many unanswered questions, she decide to call a friend in the University to direct her if possible to a biology researcher who could explain to her the basics she needs and also direct her to the correct material to read. After several phone calls she manages to have some ideas on which material to focus, so, again through the web, she collects downloaded files and prints them in order to examine them.

The next step is review the news item from the News Agency on the research discovery that has been the incentive of her topic. Apart from this, she attempts a search in the News Agency archives her newspaper co-operates with in order to possibly locate other news items that may be of interest.

When she has all the material available and roughly categorized, she starts taking notes and even coping and pasting interesting paragraphs to compile the first draft of her article. After several passes she manages to create the final version of the document, which is then passed over to the editors.

The study of journalist current practices for creating news items and their method for researching relevant material has lead to the identification of specific user needs that are summarized in section 2.4.

### *2.3.4. Amateur historians and representatives of other professions*

Historical research is not limited to history students and researchers or journalists. Historical research in some cases concerns other professionals as well as amateurs.

There are several professions that may involve at some point researching the past. These may include writers of any kind, from novels to scripts, when the story they create is taking place in an era other than the contemporary one. It may also include other professions related with the Arts. Cinema, comic, theatre, even painting and sculpturing at some cases call for looking for material on how people dressed, for example, two centuries ago or what were the means of transportation in the 1880s. Such research is necessary in these cases to provide the air of authenticity and faithful recreation of an era that otherwise would be lacking from the finished piece.

Amateurs, on the other hand, may undertake historical research for many reasons. Studying the past and its secrets holds a fascination for many people and the availability of material in the Web means that users may instantly research a topic or past event simply by typing a few keywords in Google. Seeing a film or reading a novel that takes places in the Middle Ages for example may trigger an interest for reading about this era to understand what the historical truth behind the specific artistic creation is. In some cases they research the past in the context of hobbies like miniature modeling and wargaming and role-playing games.

The defining characteristic of this user group is that research in the past is not an everyday activity but one that they undertake sporadically. However, when they undertake it, either for professional needs or out of pure enthusiasm, they want to be as thorough and effective as possible. This may be a problem for them, of course, because they lack the method and knowledge to perform organized

historical research: They do not know where and how to access primary source material, how to distinguish reliable secondary sources and how to combine all this material to get a coherent view of the topic they are interested in.

To better understand this user group, 5 users filled in questionnaires and 2 were interviewed. All perform historical research for entertainment purposes, either purely to learn as much as possible for a specific topic or to write novels or prepare material for games. The group is comprised of 5 males and 2 females of various educational levels. All have more than 5 years of computer expertise and spend daily more than 4 hours in front of the computer and on the Web. They use their own language and English as the languages in which they perform research.

As seen from Table 2-28, amateurs mostly employ secondary sources. 57% never employ primary sources. As one stated, "I would gladly use them I knew were to find them".

**Table 2-28. Amateurs' use of primary and secondary sources**

| Use pf sources | ALWAYS | OFTEN | NEVER |
|---|---|---|---|
| **Primary** | 14% | 29% | 57% |
| **Secondary** | 43% | 29% | 29% |

On the material they often use, results are summarized in Table 2-29. Is evident from these results that material in physical form it rarely used in comparison with material in digital form. Especially for audiovisual material, it seems that is only rarely used in physical form, as opposed to the digital form.

**Table 2-29. Comparative use by amateurs of material of various types in physical and digital form.**

| | OFTEN | | RARELY | | NEVER | |
|---|---|---|---|---|---|---|
| | **physical** | **digital** | **physical** | **digital** | **physical** | **digital** |
| **Text** | 57% | 100% | 29% | 0% | 14% | 0% |
| **Images** | 0% | 100% | 43% | 0% | 57% | 0% |
| **Audio** | 0% | 71% | 43% | 14% | 57% | 14% |
| **Video** | 0% | 71% | 43% | 14% | 57% | 14% |

As with the rest of the user groups, instructions and help are not often taken into account when searching for material. Simple search with keywords is definitely the most prominent one and search within the results is hardly ever used.

Websites were the main mains for material that amateurs reported to use.

**Table 2-30. Amateurs actions related to the use of digital library tools**

| | OFTEN | RARELY | NEVER |
|---|---|---|---|
| **Read Instructions** | 0% | 43% | 57% |
| **Browse Index** | 43% | 43% | 14% |
| **Search with keywords** | 100% | 0% | 0% |
| **Search with phrases** | 43% | 57% | 0% |
| **Use simple search** | 100% | 0% | 0% |
| **Use advanced search** | 71% | 29% | 0% |
| **Search within results** | 0% | 14% | 86% |

Table 2-31 summarizes the problems encountered by amateur researchers. Except from accommodating different spellings of keywords and multilingual search, the rest of the problems seem to be important for the majority of the users.

We may note that 86% of the users commented the lack of digitized material, which is to be expected as most of them did not know where to look for primary source material and digital repositories. Furthermore, a problem for them was to understand terminology and browse existing material. As in many cases they are not familiar with the topics they have to research, they are often confronted with "cryptic" or insufficient categorizations which discourage them from further exploring the particular issue.

**Table 2-31. Problems amateurs usually encounter when searching in digital repositories**

| **Problems in finding digital material** | **Amateurs** |
|---|---|
| **Difficulty in browsing** | 43% |
| **Inconsistent/inadequate material categorization** | 86% |
| **Too many irrelevant results** | 71% |
| **Not enough digitized material** | 86% |
| **Limited access to digitized material** | 71% |
| **Inability to accommodate alternative spellings of keywords** | 0% |
| **Inability to accommodate research in multilingual material** | 14% |

As the rest of the user groups, amateurs were also positive with the majority of the proposed tools, except the text synthesis one. Users felt that such a tool may ignore interesting information.

**Table 2-32. Answers provided by amateurs on the usefulness of suggested tools**

| | **YES** | **MAYBE** | **NO** | **I don't know** |
|---|---|---|---|---|
| Would you find useful a tool that provides access to primary source historical material through related secondary material? | 86% | 14% | 0% | 0% |
| Would you find useful a tool that presents secondary source historical information through a categorization of concepts | 86% | 14% | 0% | 0% |
| Would you find useful a tool that combines primary source material into a coherent text? | 29% | 71% | 0% | 0% |
| Would you find useful a search tool on audio and video primary source material? | 71% | 29% | 0% | 0% |

As it was evident from the interviews and questionnaires, amateurs perform research mostly using search engines like Google. They are very interested in viewing primary sources as well but in most cases are not certain how they can retrieve them: When they manage to find digital repositories is mostly by chance.

They start their research with search engines where they insert keywords related to their topic and then explore the retrieved websites. They usually download interesting material for later use.

As an example we present the case of George, an employee of a software company who in his spare time is an amateur novelist as well as story teller for role playing games. Both these hobbies include story making and he often wants to create stories that take place in the past, for example the Middle Ages, the 1800s, 1900s, etc. He usually comes up with a concept of a character or event as well as a place where the story will play out.

After an initial search based on his general idea of the setting for his story, he locates a place that seems to have an interesting history and, using keyword search, he tries to investigate every historical aspect of that place (religion, politics, groups that were active there in the past, myths and stories, important monuments and personalities, etc.). George downloads and prints interesting material to study it further. Sometimes his starting point may be an historical event that he wants to relate with his story.

Having collected all this material he re-examines it to get ideas to enrich his story. He would like to have had primary source material, like pictures or drawings of the place in the time period that interests him and archival material, in order to add to the story details that will make it more authentic.

## 2.4. User Needs Summary

In the aforementioned analysis of the user practices and problems when performing historical research, several user needs have been mentioned by the four user groups in the study. In this section they are summarized in Table 2-33.

**Table 2-33. Summary of user needs**

| | Users | Need |
|---|---|---|
| 1 | Historians, | Access to valid secondary sources |
| 2 | Historians | Access to secondary sources related to the primary ones they retrieve, in order to validate them |
| 3 | All | Effective concept categorization and indexes |
| 4 | Historians | Summaries of primary source material |
| 5 | Historians | Save related secondary and primary material related to a specific query in their personal computer while maintaining its structure |
| 6 | All | View concept/term definitions |
| 7 | Historians, students | View concept/term changes with the passage of time |
| 8 | All | Easy access to digitized primary sources |
| 9 | Historians, students | Easy access to references of secondary material related to a specific topic/concept |
| 10 | Historians, students | Links between primary and secondary material |
| 11 | Journalists | Overview of the secondary material in a structured form |
| 12 | Journalists | Locate quickly some relevant secondary source material |
| 13 | Journalists | Locate quickly some related primary source material to re-use |
| 14 | All | Clear and easy to use interfaces |
| 15 | Historians, students | Effective keyword search that will take into account changes in terminology |
| 16 | Historians, | Multilingual search |

| | | |
|---|---|---|
| | Journalists | |
| 17 | Historians, Journalists | Effective navigation in the categorization of concepts |
| 18 | Historians, Journalists | Ways to quickly review and reject irrelevant results |
| 19 | Journalists | Effective search on audiovisual material |
| 20 | Journalists, students | Text with relevant/important paragraphs from primary and secondary sources |

An important issue discussed during some of the focus group meetings was the enrichment of the history ontology through the contributions of the whole historian community. This is important according to the participating users, as it could make Papyrus a point of reference for the scientific community as well as contribute to the rapid population of the ontology with new material.

As the historians suggested it would be very useful for Papyrus users to be able to submit new concepts, along with their definitions, sub-concepts, etc. These could be reviewed by a committee and if appropriate, added to the history ontology.

The practices and needs discussed in this section have lead to the definition of the user requirements presented in the following one.

# 3. Papyrus Use Cases and Requirements

## 3.1. Papyrus Use Cases

### *3.1.1. User Roles*

From the study of user needs discussed in Section 2, we may distinguish three main user roles for the Papyrus platform:

- Administrators
- Content Managers
- End Users

**Administrators** have full access rights in the system and are able to manage users and assign rights.

**Content managers** are authorized users that are responsible for adding new content as well as editing the existing one. “Content” in this case includes primary and secondary source material available through Papyrus. More specifically, in our case where the domain of history and news has been selected to showcase Papyrus as a cross-domain digital library, primary source material refers to the news content whereas the secondary one to the ontologies and historical information they may include.

More specifically, different content managers may be assigned to the management of different content types (news items, News ontology, History Ontology) with the appropriate rights assigned to them by the administrator.

**End users** include all interested individuals that may access the Papyrus system to assist them in their research. These may be professional history researchers, students, journalists, sociologists, amateur historians and in general every user who is interested in a specific history topic, either professionally or for entertainment purposes.

### *3.1.2. Use Cases*

The aim of Papyrus is to create a cross-disciplinary digital library and show-case it with the domains of News and History. History researchers, either professionals or amateurs will be able to access the primary source content (News multimedia content) through a structured view of the secondary material (history).

In Figure 3-1, an overview of Papyrus is presented. On top of the multimedia content, the primary source material, there are News ontologies that serve as its categorization. New multimedia content is mapped to the News ontology off-line with a semi-automatic content analysis process. The News Ontology is managed by authorized content managers.

The History ontology represents the secondary source material in Papyrus and is managed by authorized users. This ontology is mapped to the News one. End users looking for either secondary or primary source material on particular historical research topics may browse the ontologies, navigate between them and access through them the primary source material. They may also query the history ontology directly and then either refine their search or view the results in various presentation methods. They may save the results of their queries along with the query that produced them in an appropriate structure.

End users may also submit new material for the history ontology, which the content managers may review and reject or approve.

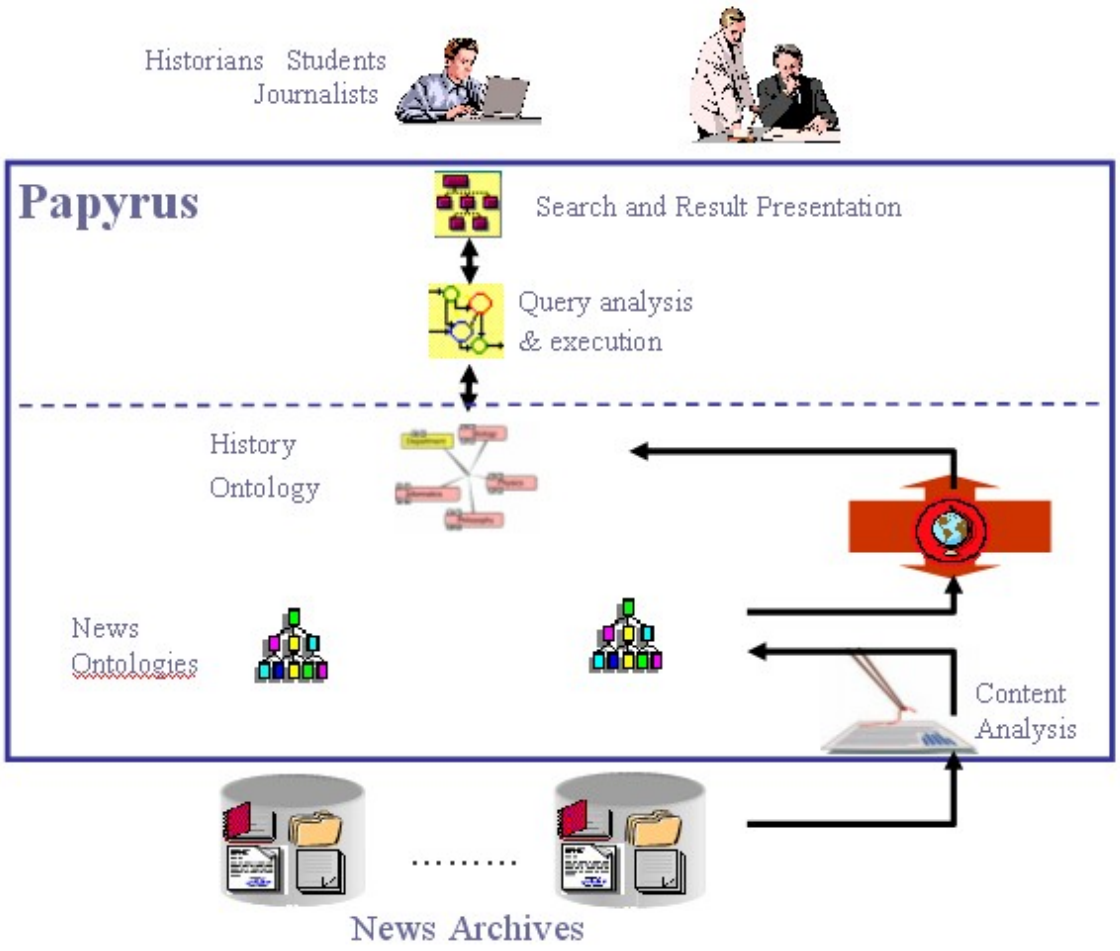


**Figure 3-1. Overview of Papyrus**

This section presents the Papyrus prototype platform use cases as they have been identified after the analysis of user needs. It is certainly expected that the use cases will be refined during the design and implementation processes according to a constant exchange of ideas with the users and will be finalised with the final delivery of the Papyrus system. The identified use cases for accomplishing the Papyrus prototype platform are listed in Table 3-1 UID is the incremental identifier of the use case. For each use case, we present its title, a brief description and the user group that it is applicable.

**Table 3-1. Overview of Papyrus Use Cases**

| UID | User Role | Use Case Title | Use Case Description |
|---|---|---|---|
| 1 | Administrator | Manage User | The administrator may add and delete users, as well as manage their authorization level in Papyrus |
| 2 | Administrator, Content Manager | Login | Authorized users have to login before performing specific actions |
| 3 | Content Manager | Manage content | Content Managers may add or remove text and multimedia content |
| 4 | Content Manager | Manage News Ontology | Content Managers may edit the News Ontology and its instances |

| | | | |
|---|---|---|---|
| 5 | Content Manager | Manage History Ontology | Content Managers may edit the History Ontology and its instances |
| 6 | Content Manager | Content Analysis | Content Managers may initiate the automatic analysis of parts of the stored content. |
| 7 | Content Manager | Map History and News Ontology entities | Content Managers may create mappings between classes and instances of the history ontology and the news one. This may be done manually or semi-automatically. |
| 8 | Content Manager | Manage ontology versions | Content Managers may create versions of the News and History Ontology |
| 9 | All | Browse the History and the News Ontology | All users may browse the ontologies and navigate through the mappings between them and also through them to the multimedia content |
| 10 | All | Query the History Ontology | All users may perform queries to the History ontology to get relevant ontology classes and instances as well as multimedia content through them. |
| 11 | All | View the search results | All users may navigate within the search results, which may be presented in different ways, according to the needs of the user. |
| 12 | All | Save the search results | All users may save the results of the query as well as the query itself for their personal archive |
| 13 | End user | Submit new secondary source material | The non-authorized users may prepare and propose new content for the history ontology to be approved by content managers. |
| 14 | End user | Register | End users may register in order to become content managers or submit new material for the history ontology |

The following tables present each use case in more detail, providing a more detailed description as well main activity steps and its priority.

**Table 3-2. Manage User Use Case**

| | |
|---|---|
| **Use Case ID:** | 1 |
| **Use case name:** | Manage User |
| **User Role** | Administrator |
| **Goals/Description:** | The Administrator may create, edit and delete user accounts. Editing refers to changing their rights in the Papyrus platform. The rights will include: editing the history ontology, editing the news ontology, editing the mappings between ontologies and managing the multimedia and textual content |
| **Scenario example:** | The Administrator is asked to approve the registration of a new account for a Content Manager that will act as editor for the News ontology. He creates the user and sets the appropriate rights for this ontology. |

| | |
|---|---|
| **Priority:** | High |
| **Activity step Description:** | Approve User<br>• The approves the new user account created with the registration<br>Edit User<br>• The Administrator selects the User<br>• The Administrator edits the appropriate rights for the user.<br>Delete User<br>• The Administrator selects the User<br>• The Administrator deletes the user |

**Table 3-3 Login Use Case**

| | |
|---|---|
| **Use Case ID:** | 2 |
| **Use case name:** | Login |
| **User Role** | Administrator, Content Manager |
| **Goals/Description:** | Authorized users have to login before performing specific actions. |
| **Scenario example:** | A content manager logins in Papyrus in order to perform edits on the History ontology. |
| **Priority:** | High |
| **Activity step Description:** | • The user fills in the username and password fields<br>• The system checks the provided information and authorizes or denies access to the system |

**Table 3-4 Manage Content Use Case**

| | |
|---|---|
| **Use Case ID:** | 3 |
| **Use case name:** | Manage Content |
| **User Role** | Content Manager |
| **Goals/Description:** | Content Managers may add or remove text and multimedia content. The content includes audiovisual as well as textual material serving as the primary source material (news) for the Papyrus platform. The user may add new content along with a set of metadata like its title, the date of insertion in the platform, its provenance, etc, as well as remove content if necessary. The user may organize the content in folders if necessary |
| **Scenario example:** | The Content Manager has received a set of news videos she wishes to include in the Papyrus platform. The content is added in the platform along with the appropriate metadata. |
| **Priority:** | High |
| **Activity step Description:** | Add content<br>• The Content Manager selects the content to be inserted<br>• The Content Manager fills in the appropriate metadata<br>• The content is uploaded in the Papyrus platform and marked as new in order to be analyzed, either manually |

| | |
|---|---|
| | or automatically and associated with the appropriate concepts in the News Ontology<br><br>Remove content<br>• The Content Manager selects the content to be removed<br>• The content is removed and all associations to it from the News ontology are removed as well |

**Table 3-5 Manage News Ontology Use Case**

| | |
|---|---|
| **Use Case ID:** | 4 |
| **Use case name:** | Manage News Ontology |
| **User Role** | Content Manager |
| **Goals/Description:** | Content Managers may edit the News Ontology and its instances |
| **Scenario example:** | The Content Manager logins into Papyrus to add in the News ontology a categorization of concepts describing the domain of quantum physics to be added in the ontology. |
| **Priority:** | High |
| **Activity step Description:** | Add Class/Instance<br>• The Content Manager selects the class under which the new class/instance will be added<br>• The Content Manager fills in the class name and adds properties.<br><br>Edit Class/instance<br>• The Content Manager selects the class/instance to edit<br>• The Content Manager adds, removes or edits properties<br>• The system warns the user of the consequences of the changes on classes/instances related to the selected one and affected by the changes.<br>• The user accepts or rejects the changes<br>• The system performs the changes.<br>• The system marks the links to the related news items as well as the mappings to History ontology concepts in order to be revised if necessary.<br><br>Remove Class/instance<br>• The Content Manager selects the class/instance to remove<br>• The system warns the user of the consequences of the removal on classes/instances related to the selected one and affected by the changes as well as the consequences for related news items.<br>• The user accepts or rejects the removal<br>• The system performs the removal. |

**Table 3-6 Manage History Ontology Use Case**

| | |
|---|---|
| **Use Case ID:** | 5 |

| | |
|---|---|
| **Use case name:** | Manage History Ontology |
| **User Role** | Content Manager |
| **Goals/Description:** | Content Managers may edit the History Ontology and its instances |
| **Scenario example:** | The Content Manager logins into Papyrus to add in the History ontology to scientists that are important for the Biotechnology domain. |
| **Priority:** | High |
| **Activity step Description:** | Add Class/Instance<br>• The Content Manager selects the class under which the new class/instance will be added<br>• The Content Manager fills in the class name and adds properties.<br><br>Edit Class/instance<br>• The Content Manager selects the class/instance to edit<br>• The Content Manager adds, removes or edits properties<br>• The system warns the user of the consequences of the changes on classes/instances related to the selected one and affected by the changes.<br>• The user accepts or rejects the changes<br>• The system performs the changes.<br>• The system marks the mappings to the News ontology concepts in order to be revised if necessary.<br><br>Remove Class/instance<br>• The Content Manager selects the class/instance to remove<br>• The system warns the user of the consequences of the removal on classes/instances related to the selected one and affected by the changes as well as the consequences for related News ontology classes and instances.<br>• The user accepts or rejects the removal<br>• The system performs the removal. |

**Table 3-7 Content Analysis Use Case**

| | |
|---|---|
| **Use Case ID:** | 6 |
| **Use case name:** | Content Analysis |
| **User Role** | Content Manager |
| **Goals/Description:** | Content Managers may initiate the automatic analysis of parts of the stored multimedia content. S/he may review the results of the analysis and edit them in order to make sure that the appropriate links between the primary source material (news items) and the news ontology are created. |
| **Scenario example:** | The Content Manager logins into Papyrus to perform the analysis on the new news items recently added in the content repository. After selecting the content he initializes the analysis process and the content is related with appropriate ontology concepts. Then he may review the |

| | |
|---|---|
| | results, accept or reject them and even add explicit links from the News Ontology to the Content |
| **Priority:** | High |
| **Activity step Description:** | Automatic Content Analysis<br>• The Content Manager selects the material to be analyzed<br>• The Content Manager starts the analysis process<br>• The system analyzes the multimedia content and relates it to ontology classes and instances<br>• The Content Manager views the analysis results<br>Editing the links<br>• The Content Manager edits or deletes links between the material and the ontology that have been suggested by the analysis results.<br>• The Content Manager may explicitly create new links if necessary.<br>Approving changes<br>• The Content Manager approves the new or edited mappings. |

**Table 3-8 Map History and News Ontology Use Case**

| | |
|---|---|
| **Use Case ID:** | 7 |
| **Use case name:** | Map History and News Ontology entities |
| **User Role** | Content Manager |
| **Goals/Description:** | Content Managers may initiate the semi-automatic mapping of History and News Ontology classes and instances. Their role is to review, edit and approve mapping suggested by the system or add new ones. |
| **Scenario example:** | The Content Manager logins into Papyrus to perform the mappings between news and history ontology entities. After the part of the ontologies on which this operation will be performed, he initializes the mapping process. Then she checks if the mappings are correct and adds 2 of her own before approving the changes. |
| **Priority:** | High |
| **Activity step Description:** | Automatic Mapping<br>• The Content Manager selects the news and history ontology parts to be mapped. S/he may select the whole ontologies if needed.<br>• The Content Manager initiates the mapping process<br>• The system analyzes the two ontologies and identifies mappings and matching between them<br>• The Content Manager may view the results<br>Editing the mappings<br>• The Content Manager edits or deletes mappings that have been suggested by.<br>• The Content Manager may explicitly create new mappings. |

| | |
|---|---|
| | Approving changes<br>• The Content Manager approves the analysis results. |

**Table 3-9 Manage ontology Versions use case**

| | |
|---|---|
| **Use Case ID:** | 8 |
| **Use case name:** | Manage ontology versions |
| **User Role** | Content Manager |
| **Goals/Description:** | The Papyrus requirements on versioning are not very demanding, however, basic support for storing and managing ontology versions is considered useful. Content Managers may create versions of the News and History Ontology, save them, and re-activate them if necessary. Appropriate mechanisms should be available for taking into account mappings as well as links to multimedia content when different versions are made active. |
| **Scenario example:** | The Content Manager is preparing to make several changes in the history ontology. She creates a new version to work on it until changes are finalized and it can substitute the current version. |
| **Priority:** | Medium |
| **Activity step Description:** | Version Creation<br>• The Content Manager selects the ontology s/he wishes to version<br>• The Content Manager provides metadata for the version<br>• The version is saved.<br>Version Activation<br>• The Content Manager selects the ontology version s/he wishes to version. The version is made activity, showing to the Content Managers possible problems such as mappings to now non-existing concepts or links to missing multimedia content.<br>• The Content Manager may edit the version<br>• The Content Manager may make the version active . |

**Table 3-10 Browse ontologies use case**

| | |
|---|---|
| **Use Case ID:** | 9 |
| **Use case name:** | Browse the History and the News Ontology |
| **User Role** | All |
| **Goals/Description:** | All users may browse the ontologies and navigate through the mappings between them and also through them to the multimedia content. The browsing tools should offer a clear |

| | |
|---|---|
| | view of the class hierarchies, possibilities for focus on specific areas retaining the context at the same time, filtering tools for viewing specific sub-hierarchies or instances with specific characteristics and appropriate representation of time and entity evolution. |
| **Scenario example:** | A History student is preparing a report on the history of alternative energies. He starts by the History ontology, exploring related terms and things and through the mappings to the News Ontology has access to news items on this specific subject. |
| **Priority:** | High |
| **Activity step Description:** | • The user may select ontology nodes to explore<br>• The user may:<br>o Access related nodes<br>o Change the nodes on focus<br>o Filter nodes of interest<br>• The user may move between the ontologies during the navigation |

**Table 3-11 Query the history ontology use case**

| | |
|---|---|
| **Use Case ID:** | 10 |
| **Use case name:** | Query the History Ontology |
| **User Role** | All |
| **Goals/Description:** | All users will be able to query the History ontology. The query tool should offer simple keyword search as well more advanced options that restrict the search in specific areas of the ontology as well as allow the construction of more complex queries. The user may be presented with suggestions for refining the query and combining it with classes or instances already in the ontology. |
| **Scenario example:** | A History student is preparing a report on the social implications of cloning. He starts by querying the History ontology with the keyword cloning but he is not satisfied with the results. However he is presented with a set of suggestions on possible concepts that may be related with cloning. He selects “bioethics” as one concept related to his topic and performs a new search. |
| **Priority:** | High |
| **Activity step Description:** | • The user may enter a keyword or phrase<br>• The system analyzes and executes the query and returns results as well as suggestions for query refinement.<br>• The user may review the results and perform a new query if needed |

**Table 3-12 View the search results use case**

| | |
|---|---|
| **Use Case ID:** | 11 |

| | |
|---|---|
| **Use case name:** | View the search results |
| **User Role** | All |
| **Goals/Description:** | All users may navigate within the search results, which may be presented in different ways, according to the needs of the user. These options may include faceted result presentation, synthesis of text with relevant paragraphs or summaries, visualization of ontology sub-trees, etc. |
| **Scenario example:** | After starting a search on “Cloning”, several results are returned. The user may select specific categories of the results in order to refine them by selecting for example “Human Cloning” or “Gene Cloning”. |
| **Priority:** | High |
| **Activity step Description:** | • The user may view the results<br>• The user may select one of the different available options to either group the results according to their related ontology classes or instances, view a timeline of terms and related results etc.<br>• The user may request the creation of a text summarizing the retrieved results. |

**Table 3-13 Save the search results use case**

| | |
|---|---|
| **Use Case ID:** | 12 |
| **Use case name:** | Save the search results |
| **User Role** | All |
| **Goals/Description:** | All users may save selected results of their searches along with the specific query that produced them. |
| **Scenario example:** | After a query on “cloning and bioethics” the user has read and found interesting the related concepts and instances produced as well as some of the related news content. S/he is provided the option to save the selected material in a structured way that will allow him/her to later review it. |
| **Priority:** | Medium |
| **Activity step Description:** | • The user selects interesting results of a query.<br>• The results are formatted and downloaded to the user’s personal computer. |

**Table 3-14 Submit new secondary source material use case**

| | |
|---|---|
| **Use Case ID:** | 13 |
| **Use case name:** | Submit new secondary source material |
| **User Role** | End users |
| **Goals/Description:** | End users, in many cases experienced history researchers may be able, if they wish to, to propose history ontology |

| | |
|---|---|
| | entries to be able later to be reviewed and approved by Content Managers. This way Papyrus will be constantly enriched by the History Community and serve as a common repository of historical knowledge. End users may propose additions only for pre-defined parts of the ontology, which are defined by the content managers |
| **Scenario example:** | An experienced history researcher who has performed research through Papyrus realizes that the history of the means of transportation is not sufficiently covered in the existing history ontology. As her phD is exactly on this area, she prepares through Papyrus a categorization of important concepts, along with definitions, related secondary bibliography, identified related news items, etc and she submits it to Papyrus for approval. |
| **Priority:** | Medium |
| **Activity step Description:** | • The user logs in<br>• The user prepares the part of the ontology s/he wishes to submit.<br>• The Content Manager may edit and then approve the submitted content.<br>• The content is added to the ontology. |

**Table 3-15 Register use case**

| | |
|---|---|
| **Use Case ID:** | 14 |
| **Use case name:** | User registration |
| **User Role** | End users |
| **Goals/Description:** | End users may register in order to become content managers or participate in functionality that requires personalized access, like the submission of new material |
| **Scenario example:** | A user fills in her personal information in order to login as a register user and be able to submit material for the history ontology. |
| **Priority:** | Medium |
| **Activity step Description:** | • The user fills in personal information.<br>• The system stores the registration request for the administrator to approve. |

## 3.2. Papyrus Ontology Requirements

The user study performed in the context of Papyrus as well as the analysis of the current state of the art in ontologies, has lead to the identification of several ontology characteristics and important issues that should be taken into account during the ontology modelling activities undertaken in the context of Papyrus.

During the user study performed in the context of Papyrus, main differences between the News and the History ontology have been identified. Firstly the main purpose of the News Ontology is the description of News items and their categorization and, following the nature of news, the attempts to capture the present. The History ontology on the other hand needs to offer a broader and more

general perspective of events and investigates a topic in large time spans, having thus a stronger need for representing time and the changes that come about with its passage.

Another difference identified is that the History ontology has to contain information in a greater level of detail, as history researchers are more interested in an in depth analysis of events. Furthermore, historians mostly study the evolution of specific concepts and the interpretation of News in relation with abstract concepts like “Progress”, “Risk”, etc. Journalists mostly care about current events/news and these concepts may be of interest to some extent but, at least currently, are not used to tag news items. Lastly, historians go from a theory/interpretation to news whereas journalists go from facts to a conclusion.

The following two sections present a brief overview of identified characteristics and requirements for the two ontologies to be developed in the context of Papyrus.

### *3.2.1. History Ontology*

The History Ontology will be essential for the success of Papyrus. This ontology will both be 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 structure 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 should be made within Papyrus to create and appropriate general and sufficiently expressive model to represent History related concepts.

The outcomes of the group discussions with the historians up to now have produced a general idea of the main points to be taken into account for ontology modelling.

Two main ontology class “types” are foreseen:

- **Things** (or Objects). Things model real world objects, persons, institutions, events, etc. and may be described by specific properties (e.g. name, date of birth, etc).
- **Terms**. Terms are more general and abstract concepts that may be of importance to a specific History Domain. Terms are described by a definition, the time they were introduced as well as their relation to Things.

As an example, for the term “cloning”, the historian may want to view:

- its definition
- the history of the term
- related terms (e.g. “Bioethics”)
- related objects/things (e.g. “Dolly”)
- sub-terms (e.g. “gene-cloning”)
- related primary source material (news items)
- references to secondary source material (e.g. papers, books on the history of Biotechnology)
- selected links (material on the website of an Academic Institution)

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 modeling of this evolution is crucial for an effective representation of History in the context of Papyrus. For most Historians the evolution of a term is a starting point when faced with the study of a particular domain and appropriately modeled it will be very useful for query analysis and enrichment.

### *3.2.2. News Ontology*

The Papyrus News ontology should conform to existing news ontologies and namely the one designed by IPTC, NewsML-G2.

**NewsML**[6] has been designed by the IPTC[7] (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.

Everything the recipient might need to know about the content of the news provided can be included in NewsML's structure. For example, NewsML enables publishers to provide the same text in different languages; a video clip in different formats; or different resolutions of the same photograph. NewsML's standardized metadata sets include globally unique identifiers and version numbers that make it easy to track the evolution of a NewsItem over time, publication status (publishable, embargoed, etc.), administrative or descriptive properties such as creator, date of creation, copyright notice, subject and abstract.

NewsML defines default controlled vocabularies to ease the interoperability of implementations but it does not dictate which vocabulary is used on a given metadata property (IPTC Subject News Codes, ISO country codes etc.). Multiple vocabularies can even be utilised within the same NewsItem.

A new major version of this standard, named **NewsML-G2**, 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.

NewsML-G2 has been built around an object model expressed as UML graphs, which may be easily mapped to a formal ontology model expressed in OWL. 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**[8] are sets of topics (aka topical subjects) to be assigned as metadata values to news objects like text, photographs, graphics, audio- and video assets. This allows for a consistent coding of news metadata over the course of time. This 3-levels taxonomy has currently 1,300 terms in it. Each term corresponds to a numeric code and is associated with labels in English, French, German and Spanish. The use of IPTC Subject News Codes is recommended by the IPTC for the classification of NewsML documents.

Within the MESH[9] project, an OWL ontology has been built, extending the IPTC taxonomy with terms and categories in two areas ("Disaster and accidents" and "Unrest conflicts and war"). In the context of Papyrus this ontology will be re-examined and the OWL model for Papyrus will be created.

## 3.3. Papyrus User Scenarios

To further highlight the benefits of Papyrus for historical research, this section presents user scenarios on how it could be used by the various user groups that are involved in this type of research.

### *3.3.1. Historian User scenario in the biology/ biotechnology domain*

A historian who would like to perform research on a scientific/ technological topic by taking into consideration not only the scientific debates but also the philosophical discussions, the political dynamics and the social impact of scientific and technological research, could use the digitized media

[6] http://www.newsml.org

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

[8] http://www.iptc.org/NewsCodes/index.php

[9] http://www.mesh-ip.eu/?Page=Project

archive provided by Papyrus as a primary source material. News items are advisable for such a research since not only can they provide information on scientific and technological issues, but they can also reveal the role of social actors, political agendas, public perception as well as their own role and rhetoric in disseminating scientific and technological knowledge and forming public images of various disciplines, theories and artifacts. Yet, apart from offering a primary source material, in this case digitized news items from news agencies, Papyrus would provide information on valid related secondary literature, as well as a history ontology that would be helpful for the historian to understand the basic concepts of the topic under research and the correlations to each other.

1. If a historian, like Beth, a phD student of Science and Technology, wished to research cloning, she could visit Papyrus exploring both the history and news ontologies and use them interchangeably or simultaneously in order to get information on this topic.

One of the options provided by Papyrus is for the historian to survey firstly the history ontology. By typing the keyword “cloning”, or linguistically related words such as “clone”, the historian could get a definition of cloning, and a categorization of events and concepts that are relative to it. Such categorization would include definitions of sub-terms such as “human cloning” or “gene cloning”, definitions of related terms such as “biotechnology”, “bioethics”, “genetics”, a timeline of scientific research on cloning, and information on related things and objects such as “Dolly”. Papyrus would also provide one or more essays on the history of cloning, the history of the definition and the possible conceptual changes that it has gone through. Such kind of essays would also apply to the concepts and instances of the additional terms and objects provided through categorization, enabling the historian to go as more deeply and broadly as she wants in the theoretical and historical context of cloning.

2. After surveying the theoretical and historical context of cloning provided by the history ontology of Papyrus, Beth will have gotten a wide idea about the area she is going to research. Papyrus, however, can be even more useful for her in order to broaden her knowledge about the discussions on cloning from a scientific, historical, philosophical and sociological point of view, by offering information and links to valid secondary literature. First of all, it will display references to history of science and technology books and articles specialized on cloning, biotechnology, genetics, etc. References may also concern literature on the science and technology studies, the philosophy of science and sociology that could apply to the specific topic. Moreover, Papyrus will also provide links to the websites of professional societies such as the International Society for the History, Philosophy and Social Studies of Biology, where the historian may navigate in order to get more specialized information on her research topic.

3. From the surveying and reading of the history ontology of Papyrus, of selected secondary literature and professional websites, the historian can form an opinion of approaches, issues and keywords regarding the recent history of cloning and as a sub-domain of the history of biology and biotechnology.

In case she wants to employ primary source material from news agencies in order to see how this topic was presented in media and what the political, social, economical and cultural issues are involved in public debates, she could exploit the mappings from the history to the news ontology of Papyrus.

She would make a search using keywords and phrases that she may have isolated as relevant during her study of the history ontology and the other related secondary source material.

In the beginning, she will type keywords of a first set of linguistically related words and relevant concepts to cloning, such as “cloning”, “human cloning”, “clone”, “stem cell”, “DNA”, “genes”, “genome”, etc. By searching with such keywords, and having chosen that she needs to review primary material as well, the historian will get relevant results that will be retrieved through the links and mapping between the history and the news ontology and the news ontology and the content.

The benefit for her in this case is that when querying the history ontology in order to finally reach the primary source material, Papyrus automatically will take into account possible changes in the terminology, either in time or in different countries as well as relations between combinations of the given concepts. It may even allow her, if she wishes, to manually enrich her query by selecting

proposed related ontology concepts (classes or instances) or even limiting the search in particular sub-domains. This way she can be certain that she will get as many relevant results to the given keywords, but also that interesting results won't be left out due to differences in terminology that she has not taken into account.

When she retrieves the related content, she will probably read, listen or watch some of it in order to identify more keywords or phrases that are relevant to her questions to be answered. Combining the new information from the news items with the already acquired knowledge from the secondary literature, the historian will proceed in typing a second set of relevant concepts, in respect to the questions she would like to answer. Thus, for example, if his/ her focus is on the relation of new methods of biotechnology, such as cloning, with ethical issues, the historian will refine her search by typing keywords such as "ethics", "bioethics", "government", "legislation", "religion".

By returning the proper results, Papyrus will have provided the historian with a number of news items on the ethical and juridical debates on cloning, the political and economic decisions made on this issue, the public perception and opinion about cloning and its present and future impact on society, etc. These combined with the historical context as provided by the history ontology will offer a complete picture of the historical topic in question and a strong basis on which the historian may work on and create her own interpretations and conclusions.

An alternative offer to the historian is to use browsing in order to have access to selected primary source material through the related secondary material. This means that by viewing the history ontology, and following the mappings to the news ontology, a list of related news items for each concept explored may be displayed. In this way, the historian can exploit both history and news ontologies of Papyrus at the same time, getting thus a broader idea of the discussion based on evidence from the primary sources.

### *3.3.2. Student User Scenario*

Papyrus would be very useful to history students in many terms. Not only as a tool where the student can acquire information on secondary source material and also use primary source material, but mostly as a tool that could familiarize students with a method in doing historical research.

Here we present two scenarios, one for the student of a humanities related departments and one for the engineering and sciences ones

#### 3.3.2.1 Students of history and social sciences departments

Let's consider again the example of a student of a history of science and technology department who is asked to make an essay on the evolutionary theory-creationism public controversy for the last decade in the United States. As seen in Section 2.3.2.1.1, traditional research methods, even with digital tools are not very effective for undergraduate students. Papyrus, on the other hand, could help the history student retrieve useful information for her research by employing both primary and secondary source material in an interchangeable way:

1. The history student will go first in the history ontology of Papyrus. She will type the keywords "creationism" and "evolutionary theory" and get definitions for these terms, as well as a categorization of relevant concepts. An essay on the history of biological sciences provided by Papyrus would inform the student about controversies on evolutionary theory, including perhaps a timeline with the most significant instances and events from the 19$^{th}$ century and onwards.

2. Reading the texts as well viewing directly the interrelations of concepts, the student will be informed about more concepts related to her topic, such as "intelligent design", "Darwinism", etc., that could further her comprehension about the history of the specific controversy. Papyrus will also help the student to make a quick recontextualization of her topic by providing essays on the history of the controversy between science and religion, part of which the evolutionary theory-creationism controversy is. Moreover, it will also offer a selected set of secondary bibliography on such topics, as well as links to scholar websites discussing similar issues.

3. Through the history ontology and the appropriate mappings to the news ontology, the student will also have access to news agency archival material and will be able to use articles and news items as

primary sources for her research. The public controversy between creationism and evolutionary theory can best be studied by such kind of material, since this can provide information about possible research questions, such as how this controversy was reflected in news media, how it was perceived by the general public and in what sense did it affect political choices and decisions. Restricting the time period, the student would get a number of articles negotiating the topic of his/ her interest. Through the reading of some of the articles, the student may come across to highlighted words, such as "education", "textbooks", etc., that can be further used as keywords in order to broad or refine his search in the primary source material and help her orient her essay on specific topics. By using such phrases she would discover, for example, that the creationism-evolutionary theory public controversy affected political decisions of the American government in respect to the educational system and the curricula of secondary schools. The student would have reached thus the point of answering specific questions for her historical research by employing primary source material with help from the history ontology of Papyrus.

#### 3.3.2.2 Students of history and social sciences departments

We have already mentioned the need for a tool such as Papyrus for the education of History for future historians. However, the use of Papyrus could allow engineering and science students to actually compare their secondary literature readings with primary research to an archive. To follow our example, a student asked to write a term paper on the history of the artificial intelligence could use Papyrus to undertake research in a digitized news agency archive in a manner that would be both efficient and engaging. For example, a student asked to write a term paper on the history of artificial intelligence could use Papyrus in order to undertake research in digitized news agency archives as following:

1. The student would first go to the history ontology of Papyrus and type the key phrase "artificial intelligence". Papyrus will include an in-built essay on the history of computing that will take into account changes in the history of the public image of artificial intelligence. This essay will inform the student that what is now called "artificial intelligence" was expressed in the past by terms such as "electronic brain" or "mechanical brain". The history ontology of Papyrus will also offer relevant secondary bibliography and links to professional websites, for those students who would like to go further on this topic.
2. After reading the essays in history ontology of Papyrus, the student will go to the news ontology in order to search for primary source material, i.e. articles related to his/ her research topic from the news agency archive. There, instead of typing the phrase "artificial intelligence", which would return hundreds of results that would probably be invaluable in order to understand past expectations of computing technology, the student will type "electronic brain" and "mechanical brain", a useful information gained by the reading of the history ontology of Papyrus. In this way he/ she will get more relevant results to the history of artificial intelligence avoiding dealing with a bulk of articles related to the current state of this technological domain that would probably be useless for the writing of his/ her term paper.
3. Papyrus can further be useful to the student, since it might propose some relevant terms to electronic and mechanical brain, such as "progress", "robot", etc., which can orient student's term paper. Student may go back to the news ontology and type phrases such as "electronic brain and progress", in order to find articles on this topic.
4. The student could eventually decide to read some of the return results in order to construct a term paper essay that expands on the essay that he/she started with. In this way, history and news ontologies of Papyrus can be proved very helpful and informative for the engineering student to engage him/ her in historical research by using primary source material and to write his/ her term paper essay without getting lost in the archives and spending valuable time.

### *3.3.3. Journalist User Scenario*

Papyrus can also be used by professional journalists, for example by *Marie-Alice Fleuette*. She covers rapidly evolving scientific fields for *La Recherché,* a Science and Technology magazine, addressed to the wider public. She covers a whole set of fields, periodically returning to them for publishing new stories.

Today *Marie* is working on an article that refers to stem-cell policy, a field involving technological and scientific events in the key area of biomedical technology and science. She is the first time she is faced with this topic and decides to use Papyrus to firstly have a brief overview of this area of biotechnology. With a simple query on “stem-cells” she gets the ontology entities that concern stem-cells and all related issues like policies, bioethics, cloning, etc. She spends some time studying definitions and following the history of terminologies and when she feels confident that she has a clear idea of the subject, she moves to related news items.

There she may find interesting photographs as well as more detailed articles on the progress of policies related to the issue.

When she feels she has collected enough material she may request a summary text of selected results with links to the actual secondary and primary content, as well as save the collected material in order to have it available when she is compiling her article.

## 3.4. Papyrus System Requirements

The requirements for the Papyrus system can be divided into functional and non-functional ones. They come from the perspective of the potential groups of users and they also refer to specific technical aspects of the system itself.

The functional requirements can directly be extracted from the use cases of the prototype. In order to identify the appropriate non-functional requirements, the main aspects of the Papyrus system, from a technical point of view, should be considered.

### *3.4.1. Papyrus System Functional Requirements*

The Papyrus functional requirements can be clustered into the following categories:

- Content management requirements
- Query and retrieval requirements
- User Interface requirements
- Security requirements
- Requirements for providing miscellaneous functionality

Based on these categories, the rest of this section presents the functional requirements of the Papyrus platform, which can be associated with the Use Cases of the Papyrus prototype. Thus, for every category of requirements, a table is provided, which contains the following:

- The incremental Requirement Identification (RID);
- The respective requirement description;
- The association with the Use Cases defined in Section 3.1.2.

It should be noted that in some cases, the association with use cases is not applicable, since the respective functional requirements may stem from the general concept of the goals of Papyrus. The following tables present the requirements for each category

**Table 3-16 Content management requirements**

| RID | Description | Associated use case |
|---|---|---|
| 1 | User creation | 1 |
| 2 | User deletion | 1 |
| 3 | User editing | 1 |
| 4 | Edit user rights | 1 |
| 5 | Upload content | 3 |
| 6 | Organize multimedia content in categories | 3 |
| 7 | Download content | 3, 12 |
| 8 | Edit content metadata | 3 |
| 9 | Create ontology class | 4, 5, 13 |
| 10 | Create ontology instance | 4, 5, 13 |
| 11 | Delete ontology class | 4, 5, 13 |
| 12 | Delete ontology instance | 4, 5, 13 |
| 13 | Edit ontology class | 4, 5, 13 |

| | | |
|---|---|---|
| 14 | Edit ontology instance | 4, 5, 13 |
| 14 | Add class property | 4, 5 |
| 16 | Edit class property | 4, 5 |
| 17 | Delete class property | 4, 5 |
| 18 | Edit instance property value | 4, 5, 13 |
| 19 | Export multimedia and ontology content | 12 |
| 20 | Download exported multimedia and ontology content | 12 |
| 21 | Analyze multimedia content | 6 |
| 22 | Add multimedia content and news ontology links | 6 |
| 23 | Edit multimedia content and news ontology links | 6 |
| 24 | Approve multimedia content and news ontology links | 6 |
| 25 | Remove multimedia content and news ontology links | 6 |
| 26 | Search for mappings between history and news ontology | 7 |
| 27 | Add history and news ontology mappings | 7 |
| 28 | Edit history and news ontology mappings | 7 |
| 29 | Approve history and news ontology mappings | 7 |
| 30 | Remove history and news ontology mappings | 7 |
| 31 | Create new ontology version | 8 |
| 32 | Save ontology version | 8 |
| 33 | Activate ontology version | 8 |
| 34 | Submit new history ontology material | 13 |
| 35 | Authorize new ontology material | 13 |
| 36 | Edit new history ontology material | 13 |
| 37 | Define ontology parts for which end users may submit material | 4, 13 |

**Table 3-17 Browsing and querying requirements**

| RID | Description | Associated use case |
|---|---|---|
| 38 | Browse the content categories | 3 |
| 39 | Browse the ontology | 4, 5, 9 |
| 40 | View the search results categorized by metadata | 11 |
| 41 | Navigate from the history ontology to the news ontology and backwards | 9 |
| 42 | Navigate from the news ontology to multimedia content and backwards | 9 |
| 43 | Filter the ontology | 4, 5, 9, 11 |

| | | |
|---|---|---|
| 44 | Use keyword search to query the history ontology | 10 |
| 45 | Use advanced keyword search to query the history ontology | 10 |
| 46 | Use complex queries to query the ontology | 10 |
| 47 | Create summaries of the content | 11 |

**Table 3-18 Miscellaneous functionality requirements**

| RID | Description | Associated use case |
|---|---|---|
| 48 | Login | 2 |
| 49 | View login authorization (login) results | 2 |
| 50 | Register | 14 |

### *3.4.2. Papyrus System Non-Functional Requirements*

The non-functional requirements that should be considered during the development of the Papyrus platform prototype are presented in

**Table 3-19 Papyrus non-functional requirements**

| RID | Requirement | Description |
|---|---|---|
| 1 | Storage | The system contains a repository for the multimedia content (image, audio, video, text) as well as for the history and news ontologies and their instances |
| 2 | Back-up | The system offers back-up options for the content |
| 3 | Security | Single authentication process is required for accessing the functionality of the system. |
| 4 | Scalability/Expandability | The system should be able to handle the increasing size of the ontologies as well as the multimedia content. |
| 5 | Availability | Ensure that authorized users have always access to data and associated assets 24/7 with 99.9% reliability. This requirement entails stability in the presence of failure. |
| 6 | Usability | Easy to use. User documentation should not be necessary for ordinary tasks. |
| 7 | User Interface | Should give access to all system functionalities providing easy navigation through all features. |


# 4. Conclusions

This document presented the process followed by the Papyrus project to capture, analyse and record the requirements for the Papyrus prototype platform, as expressed by all user groups that perform historical research. The tools used to achieve these objectives have been described in detail along with the results, which will be later employed in the system design process. Although this document describes a baseline of requirements, it is expected that subsequent refinements will be applied through the design and implementation phases according to a feedback loop connecting the users testing and evaluating the prototypes with the implementation team.

# 5. References


[1] Alison Jones, “The many uses of newspapers”, Technical report for IMLS project- “The Richmond Daily Dispatch". 30/06/2008, Available at http://www.perseus.tufts.edu/hopper/publications

[2] Vincent Price, Public Opinion - Communication Concepts, Sage Publications, Inc, 1992.

[3] David Stoker, “Should newspaper preservation be a lottery?” Journal of Librarianship and Information Science, 31, no. 3, (1999), p. 132

[4] Massimiano Bucchi and Renato G. Mazzolini, “Big Science, Little News: Science Coverage in the Italian Daily Press, 1964-1997”, Public Understanding of Science 12 (2003), pp. 7-24

[5] Martin W. Bauer, Kristina Petkova, and Pepka Boyadjieva, “Long-term trends in the public representation of science across the 'Iron Curtain’: 1945-1995”, Social Studies of Science, 36 (2006), pp. 99-131

[6] Eirini Mergoupi-Savaidou, Faidra Papanelopoulou and Spyros Tzokas, “The Public Image(s) of Science and Technology in the Greek Daily Press, 1908-1910”

[7] Matiana González-Silva and Néstor Herran, “Ideology, Elitism and Social Commitment: Alternative Images of Science in Two fin de siècle Barcelona Newspapers”

[8] Casper Andersen and Hans Henrik Hjermitslev, “Direct Public Interest: Danish Newspaper Science 1900-1903”. Prepared for a collective volume in Centaurus (submitted for peer review)

[9] Eirini Mergoupi-Savaidou, Faidra Papanelopoulou and Spyros Tzokas, “Methodological and Historiographical Reflections on the Use of Newspapers in the History of Science”, 6th STEP Meeting, Istanbul, June 18-22, 2008. (oral presentation, the manuscript under preparation for publication is available from the authors)

[10] Cohen, Daniel J. & Rosenzweig, Roy, Digital History: A Guide to Gathering, Preserving, And Presenting the Past on the Web, (University of Pennsylvania Press, 2005).

[11] Janet Gertz, “Selection for Preservation in the Digital Age”, Library Resources & Technical Services, 44(2) (2000): 97-104.

[12] Alison Jones, “The many uses of newspapers”, Technical report for IMLS project-“The Richmond Daily Dispatch”, http://www.perseus.tufts.edu/hopper/publications [accessed on 20/04/2008].

[13] “Summit on Digital Tools for the Humanities: Report on Summit Accomplishments”.

# 6. Annex I – Questionnaire


# QUESTIONNAIRE

***Papyrus is a project that intends to showcase the use of a cross-discipline digital library for the recovery of history from news digital content. It will attempt to understand user queries in the context of the History discipline, look for content in the News domain and return the results in a way useful to the respective users.***


**This questionnaire is purposed**

**to identify User Needs for potential Services and Methods,**

**which may in the future be available through the Papyrus platform.**

**Through this User Survey we aim to collect information,**

**which we will analyse and finally utilize in building a Papyrus platform**

**in the way that it will be most useful to:**

**professional historians, students and journalists**

**as well as to the amateur historian.**

## Respondent's details

*Name:

*Organization:

Profession:

Country:

* Optional fields

# 1. Personal Information

**To help us evaluate your answers, please indicate:**

**1.1. Your age is between**

☐ 18-25
☐ 26-35
☐ 36-50
☐ 51-

**1.2. Your gender**

☐ Male
☐ Female

**1.3. Your educational level**

☐ High school
☐ BA/ College degree
☐ MA/ MSc degree
☐ PhD student
☐ PhD holder
☐ Other - please specify:

**1.4. Which languages do you feel comfortable to use in your research?**

☐ English
☐ German
☐ French
☐ Italian
☐ Spanish
☐ Other - please indicate:

## 2. ICT (Information and Communication Technology) general skills

**2.1. Your computer experience:**

☐ 1 year or less

☐ 2 to 4 years

☐ 5 years or more

**2.2. How comfortable do you feel using computers, in general?**

☐ Very comfortable

☐ Comfortable – I can do most things I want to do

☐ Neither comfortable nor uncomfortable

☐ Uncomfortable

☐ Very uncomfortable

**2.3. How much time do you spend in front of a computer?**

☐ Under 1 hour per day

☐ 1-3 hours a day

☐ 4-6 hours a day

☐ More than 6 hours per day

**2.4. How often do you use the internet?**

☐ Under 1 hour per day

☐ 1-3 hours a day

☐ 4-6 hours a day

☐ More than 6 hours per day

# 3. Researcher's Profile

### 3.1. What is your field?

☐ Historian

☐ Journalist

☐ Cultural studies worker

☐ Other - please specify:

### 3.2. What is your level of engagement in your field?

☐ Professional

☐ Amateur

☐ Other - please specify:

### 3.3. Why do you usually undertake historical/cultural research?

| | OFTEN | RARELY | NEVER |
|---|---|---|---|
| Write a monograph/dissertation | ☐ | ☐ | ☐ |
| Write an academic paper | ☐ | ☐ | ☐ |
| Write a journalistic article | ☐ | ☐ | ☐ |
| Prepare a TV programme | ☐ | ☐ | ☐ |
| Prepare a radio programme | ☐ | ☐ | ☐ |
| Professional work (other than historical or journalistic) | ☐ | ☐ | ☐ |
| Entertainment/curiosity | ☐ | ☐ | ☐ |
| Other | ☐ | ☐ | ☐ |

Please specify ...

### 3.4. What type of sources do you employ in your research?

| | ALWAYS | OFTEN | NEVER |
|---|---|---|---|
| Primary[10] | ☐ | ☐ | ☐ |

[10] Primary source is an original source of information that has been created by an authoritative individual with direct knowledge of the fact he/ she describes or impresses. On the contrary, secondary source discusses information originally presented elsewhere (primary sources or other

| | | | |
|---|---|---|---|
| Secondary | ☐ | ☐ | ☐ |

**3.5. What are the places you visit in order to find your material?**

| | OFTEN | RARELY | NEVER |
|---|---|---|---|
| Historical archives (public/private) | ☐ | ☐ | ☐ |
| Newspaper archives | ☐ | ☐ | ☐ |
| Radio archives | ☐ | ☐ | ☐ |
| TV archives | ☐ | ☐ | ☐ |
| Libraries | ☐ | ☐ | ☐ |
| Museums | ☐ | ☐ | ☐ |
| Other institutions | ☐ | ☐ | ☐ |

Please specify ...

**3.6. What methods do you use to save your research material in your personal archive?**

| | OFTEN | RARELY | NEVER |
|---|---|---|---|
| Notes | ☐ | ☐ | ☐ |
| Photocopies | ☐ | ☐ | ☐ |
| Digital camera | ☐ | ☐ | ☐ |
| Other | ☐ | ☐ | ☐ |

Please specify ...

**3.7. How often do you use you research material in physical form?**

| | OFTEN | RARELY | NEVER |
|---|---|---|---|
| Text | ☐ | ☐ | ☐ |
| Images | ☐ | ☐ | ☐ |
| Audio | ☐ | ☐ | ☐ |
| Video | ☐ | ☐ | ☐ |

secondary sources). For example, an article in the history of stem cells in the journal History of Biology is a secondary source for a historian who researches the history of stem cells by using as primary material references on stem cells found in news agencies items

**3.8. How often do you use you research material in digital form?**

| | OFTEN | RARELY | NEVER |
|---|---|---|---|
| Text | ☐ | ☐ | ☐ |
| Images | ☐ | ☐ | ☐ |
| Audio | ☐ | ☐ | ☐ |
| Video | ☐ | ☐ | ☐ |

**3.9. You use printed/manuscript material for your research available in the form of:**

| | OFTEN | RARELY | NEVER |
|---|---|---|---|
| Books | ☐ | ☐ | ☐ |
| Journals | ☐ | ☐ | ☐ |
| Newspapers | ☐ | ☐ | ☐ |
| Encyclopaedias/ dictionaries | ☐ | ☐ | ☐ |
| Other documents | ☐ | ☐ | ☐ |

Please specify ...

**3.10. You use digital textual material for your research available in:**

| | OFTEN | RARELY | NEVER |
|---|---|---|---|
| Books | ☐ | ☐ | ☐ |
| Journals | ☐ | ☐ | ☐ |
| Newspapers | ☐ | ☐ | ☐ |
| Encyclopaedias/ dictionaries | ☐ | ☐ | ☐ |
| Websites | ☐ | ☐ | ☐ |
| Other documents | ☐ | ☐ | ☐ |

Please specify ...

**3.11. Who is your research actually for?**

☐ University

☐ Media

Specify ☐ public or ☐ private

Specify kind of media:

☐ Private Company

Specify sector:

☐ Other public/ private institutions

Specify:

☐ Myself

## 4. Use of Digital Tools

**4.1. Do you use the web for acquiring information regarding your research/work/studies, for example through a search engine, like Google?**

☐ YES

NO, because:

☐ I prefer printed material /sources

☐ I do not trust the Internet/Web

☐ I am not aware of any useful online services

☐ I find using the Internet search engines complicated /troublesome

Comments:

**4.2. Do you do multilingual research on the web?**

☐ Yes

☐ No

Comments:

**4.3. Have you ever visited a digital library or archive[11] through the web?**

☐ Yes

☐ No

Comments:

**4.4. If yes, please indicate which one(s):**

**4.5. Do you read the instructions when using digital libraries and archives?**

☐ Always

[11] For example, http://echo.mpiwg-berlin.mpg.de/home

☐ Rarely

☐ Never

**4.6. In what order do you employ the following search methods in digital libraries and archives and how often?**

| select order(1-3) | | OFTEN | RARELY | NEVER |
|---|---|---|---|---|
| 1 | Index | ☐ | ☐ | ☐ |
| 1 | Keywords | ☐ | ☐ | ☐ |
| 1 | Phrases | ☐ | ☐ | ☐ |

**4.7. When using keywords, what kind of keyword search do you use and in what order?**

| select order(1-3) | | OFTEN | RARELY | NEVER |
|---|---|---|---|---|
| 1 | Simple search | ☐ | ☐ | ☐ |
| 1 | Advanced search | ☐ | ☐ | ☐ |
| 1 | Search within results | ☐ | ☐ | ☐ |

**4.8. What are the problems you encounter when you search in digital libraries and archives?**

☐ Difficulty in browsing

☐ Inconsistent/inadequate material categorization

☐ Too many irrelevant results

☐ Not enough digitized material

☐ Limited access to digitized material

☐ Inability to accommodate alternative spellings of keywords

☐ Inability to accommodate research in multilingual material

## 5. Historical Informatics

**5.1. Do you read reviews of web sites in history journals?**

☐ Systematically

☐ Often

☐ Accidentally

☐ Never

**5.2. Do you read articles/essays on the impact of ICT (Information and Communication Technology) on the humanities, the social sciences, and cultural studies more generally?**

☐ Systematically

☐ Often

☐ Accidentally

☐ Never

**5.3 Have you attended events relevant to on the impact of ICT on the humanities, the social sciences, and cultural studies more generally?**

☐ Yes

☐ No

**If yes, where?**

☐ Conferences

☐ Lectures

☐ Other – please specify:

**5.4. Do you discuss with colleagues the impact of ICT on the humanities, the social sciences, and cultural studies in general?**

☐ Systematically

☐ Often

☐ Accidentally

☐ Never

**5.5. Are you familiar with areas like Digital Humanities, Humanities Computing, Historical Informatics, History and Computing, or other similar areas (please specify: )?**

☐ Very much

☐ Barely

☐ No

**5.6. Have you participated in initiatives/projects in any of the above areas?**

☐ Yes – please specify:

☐ No

**5.7. Are you familiar with institutions specializing in these areas?**

☐ Yes – please specify:

☐ No

**5.8. Have you ever used special software for historical research or other similar tools from the aforementioned (Digital Humanities, Humanities Computing, Historical Informatics, History and Computing) or other similar areas?**

☐ Yes – please specify:

☐ No

**5.9. If yes, have you done so for research in the history of technology and science?**

☐ Yes – please specify:

☐ No

## 6. Would you find useful…

**6.1. Would you find useful a tool that provides access to primary source historical material through related secondary material?**

☐ Yes

☐ Maybe

☐ No

☐ I don't know

Please Comment…

**6.2. Would you find useful a tool that presents secondary source historical information through a categorization of concepts (persons, events, etc)?**

☐ Yes

☐ Maybe

☐ No

☐ I don't know

Please Comment…

**6.3. Would you find useful a tool that combines primary source material into a coherent text?**

☐ Yes

☐ Maybe

☐ No

☐ I don't know

Please Comment…

**6.4. Would you find useful a search tool on audio and video primary source material?**

☐ Yes

☐ Maybe

☐ No

☐ I don't know

Please Comment…

**Thank you for your time!**