# Modeling and Mapping Multilingual and Historically Diverse Content

Nadzeya Kiyavitskaya[1], Akrivi Katifori[2], Yannis Velegrakis[1]
Chrisa Tsinaraki[1], Siarhei Bykau[1], Eirini Savaidou[3], Aristotle Tympas[3],
Yannis Ioannidis[2] and Manolis Koubarakis[2]

[1] Department of Information Engineering and Computer Science, University of Trento, via Sommarive 14, 38100 Trento, Italy
{nadzeya, velgias, chrisa, bykau}@disi.unitn.eu
[2] Department of Informatics and Telecommunications, National and Kapodistrian Univeristy of Athens, Panepistimioupolis, Ilissia, 157 84, Athens, Greece
{vivi, yannis, koubarak}@di.uoa.gr
[3] Department of Philosophy and History of Science, National and Kapodistrian Univeristy of Athens, Panepistimioupolis, Ilissia, 157 84, Athens, Greece
{tympas, savaidou}@phs.uoa.gr

**Abstract.** Recent digitization efforts made archival content more available and even searchable through the Web. History researchers use this content for studying past events in relation to general historiographical issues, which may involve politics, society, ethics, and others. However, for locating relevant content the researchers need to know the terminology used for the topic of interest in the past. This problem is crucial for history researchers, because it affects the time and quality of their work. Another problem of great importance when dealing with the archival content is multilingualism. Simple translation is not enough to identify a relevant term in other language, because a term may undergo different changes in different social or cultural contexts. In order to address these challenges, the EU-funded project Papyrus aims to develop tool support for cross-disciplinary information retrieval of news content for historical research. To model both disciplines, i.e. history and news, we developed two ontologies. The News ontology reflects the perspective of news professionals on digital archives using the NewsML-G2 standard. Whereas the History ontology models the history perspective on the events and topics covered by the news. The History ontology is based on the CIDOC Concept Reference Model that embraces several standards of modeling information in the cultural heritage domain. To provide a means of communication between these two disciplines, we use mappings that establish correspondences between the News and History ontologies. This work discusses the major challenges in modeling and mapping terms and concepts describing the archival content that is multilingual and historically diverse.

**Keywords:** history ontology, conceptual modeling, digital libraries.

## 1 Introduction

The last decade has brought more and more libraries and archives towards the full digitization of their textual and in some cases also multimedia content. This on-going effort opens new opportunities for searching, accessing and retrieving library and archival content. The main need is currently not as much for content digitization, but for new research methods and software tools to access digitized textual, audio and video materials. This type of tools is sometimes referred to as “second generation” digital information retrieval tools, as opposed to the simple and in most cases not so effective tools used currently for searching digitized content.

To accommodate this need, the Papyrus research project of the European Union brings together historians of technology and science who specialize in using media archives, journalists, and computer scientists in order to investigate issues related to the use of media for historical research and to propose ways to support this research. Our main objective is to provide a dynamic digital library that will accept queries in terms relevant to the history researcher and then help this researcher to look for media content relevant to the query. In order to realize this objective, we have been collecting and formalizing the requirements of end users in the Papyrus History ontology, which attempts to provide a model for storing and retrieving historical information and it is based on the CIDOC CRM standard. The metadata important for describing the original archival content have been represented in the so called Papyrus News ontology.  Modeling these two different domains presents a number of research challenges.

This paper presents the Papyrus approach to model and map terms and concepts describing archival content that is multilingual and historically diverse, with specific focus on the CIDOC CRM – based History ontology and the historians’ perspective while interacting with the Papyrus prototype.

The remainder of this paper is structured as follows. Section 2 provides a general introduction of the Papyrus project and its main objectives. Section 3 describes the History ontology we developed in the project, while Section 4 proceeds with the News ontology. Section 5 presents the way of linking the two Papyrus ontologies through mappings. Finally, Section 6 shows how all the modeling solutions are used in the end user system, and conclusions are drawn in Section 7.

## 2 The Papyrus Project

The Papyrus project provides a methodology and a set of semantic web tools in order to approach the important issue of information retrieval within this diverse and large digital library content. Papyrus is an EU funded research project that started in March 2008. It intends to provide a dynamic digital library which will understand user queries in the context of a specific discipline, look for content in a domain alien to that discipline and return the results presented in a way useful and comprehensive to the user. To be able to achieve this, the source content has to be ‘understood’, which means analysed and modelled according to a domain ontology. The user query also has to be ‘understood’ and analysed following a model of this different discipline.

Correspondences will then have to be found between the model of the source content and the realm of the user knowledge. Finally, the results have to be presented to the users in a useful and comprehensive manner according to their own ‘model of understanding’.

To realize this vision, Papyrus has applied and extended existing Semantic Web technologies. The Papyrus platform is designed with two ontologies at its core [10], the History and News one, which model the History and News domains respectively. The two ontologies have been created as extensions of existing standards with the cooperation of the corresponding domain experts, journalists and historians. The platform, shown in Figure 1, showcases its approach with the use case of historical research in news archives. In the context of the project, these are the news archives of Agence France Press (AFP) [24] and Deutsche Welle [25]. These archives are represented in XML and stored in a relational database.

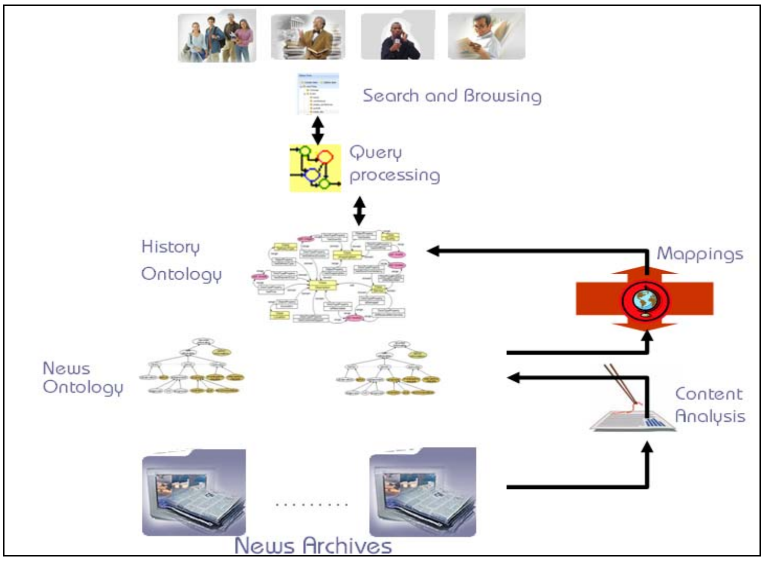


**Figure 1. The Papyrus Platform**

Moreover, the platform offers a specialized web-based ontology browser [16] which, together with the keyword search and the mapping mechanisms of the platform, enables users to navigate from History ontology entities to News ontology entities and achieve effective access to the archival material. Several Web tools were also developed to support distributed multi-user ontology editing, creation of mappings between the two domain ontologies, and management of news content and content analysis results.

The current working prototype is already available [26] and its preliminary evaluation was concluded.

# 3 The History Ontology

The main Papyrus end user group for our selected use case includes all users that may be interested in performing historical research by using news archive content. These include historians, journalists and social scientists as well as other professionals or amateurs who may have a passing interest in history.

In order to properly model the user needs and requirements of the Papyrus target users, we fulfilled a comprehensive requirements analysis involving interviews and user questionnaires among representatives of all the user groups, but mostly historians [17]. This section briefly presents the main user needs derived from this study.

## 3.1 Requirements

An important step to understanding the user needs to be supported within Papyrus has been the study of representative topics and questions for historical research. An example maybe the following: “I am interested in information on the changes in biotechnology from the beginning of the 20th century until 1970”. History researchers proceed in specific steps when attempting to gather the material needed to investigate a specific topic like the aforementioned one. These steps include (in any order):

- *Collecting relevant secondary material*, which includes essays of other history researchers on related subjects. This material typically comes with a set of common vocabularies used by historians to refer to the topics covered by particular essays. These include historiographical issues, like “Controversies and Disputes” or “Discipline formation” or “Change in science” as well as general concepts like that of “Religion” or “Politics”.
- *Collecting primary material*, i.e., news archive content related to the research subject. This material usually comes with a different set of vocabularies, the one prominent during the time of the creation of the archive documents.

A very relevant issue to our project has been the way that historians use to search and explore archival content. Either with printed material or with digitized one, their preferred methods seem to be keyword searching and exploration. More specifically, the usual way to proceed when searching for relevant material is to break down the research topic into keywords and then try to find material related to these keywords. Through our study it was evident that historians feel comfortable with keyword search and it is their main method for retrieving content from an archive. However, most of them pointed out the deficiencies of existing keyword search tools for archives, related both to precision and recall. As a result, it is important for them to be able to have an effective keyword search tool to support archival research. Another important requirement is the one for providing efficient ways to browse vocabularies and catalogues related to their research.

As our history partners explained, a very important issue is the change in concepts with the passage of time, which may include changes in their name or subtle changes in their definition.

An example is the modeling of the history of the term “biotechnology” which has changed in meaning and names many times within the 20th century. Biotechnology as a concept and scientific discipline has progressed from food technology and fermentation to genetics and biomedical engineering [2]. Time is an important factor as the assignment of time periods, in some cases not having exact limits, is essential for describing this evolution of concepts.

The issue of multilingualism in the context of a digital repository providing access to archival content of different countries and in different languages is particularly important for historical research. One dimension of the problem is related to the fact that a term may have been introduced in different time points in different languages. For example, “biotechnology” has undergone different development paths in German-speaking and English-speaking countries [2]. A second, more complicated dimension of the problem is related to the fact that the same term, during the same time period, could mean different things in one language than in another. If we take for example the development of biotechnology in the German-speaking countries, there were two terms with different connotations used to refer to this one term in English, i.e., “biotechnik” (biology-based technology) and “biotechnologie” (microbiology and fermentation).

In general, different terms can be used to describe the same concept under different contexts, essentially different working environment conditions, specified by the parameter values of time, place, language, dialect, domain, historiographical issues (i.e. social, cultural etc.), formality and diatype (i.e. a language variation, determined by its social purpose).

### 3.2 Using CIDOC CRM as a Basis

To model the historian world, we started from the CIDOC conceptual reference model (CRM) [5]. The CIDOC CRM is an international standard, ISO 21127:2006, providing definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. This ontology is the best available in the area being the result of 10 years of development work by an interdisciplinary team of experts, coming from fields such as computer science, archaeology, museum curation, art history, natural history, library science, physics and philosophy.

The reason for reusing this model was that CIDOC CRM includes a model not only for cultural heritage objects but also for their history. Moreover, given that the model is aligned to the upper-level ontology DOLCE, this allows the representation of general domain-independent concepts as well as its extension with some domain-specific knowledge under the same basic structure. For example, CIDOC CRM introduces concepts like Actor, Period, Place, and Time-Span, all useful to model History. Furthermore, in line with the user needs for the History Ontology, this model distinguishes between Conceptual Objects and Physical Things and offers an elaborate structure where a more History-oriented ontology could be built upon. These similarities in modeling perspectives allowed us to adopt the CIDOC CRM as an initial building structure to flesh out the historical knowledge.

However, while Papyrus aims to model history in general, the CIDOC CRM focuses on the attributes and activities related to varied museum artifacts. For example, the model describes the way of an item's production, the transfer of its physical custody, and the assignment of attributes. In order to preserve the CIDOC CRM consistency, we decided to keep all the CRM concepts and modify the ontology by means of augmenting the model with additional knowledge that is important for the needs of historians. The version we started with is 4.2.1 [5].

### 3.2 Extending CIDOC CRM with Abstract Domains and Issues

In our effort to "formalize" the History of Science and Technology domain in an ontology, our historian partners suggested two very important societies in the field. The one is the History of Science Society with its journal "ISIS" [27] and the other is the Society for the History of Technology with its journal "Technology and Culture" [28].

These were used as sources with the following objectives:

- Periodization. In order to represent time properly within the ontology, we needed to have an in-depth understanding of how historians use time periods and chronological divisions.
- Classification. In order to create a rich as well as structured ontology, we needed to study the formal classification used in the subject index of the journals of the two selected societies.

After collecting this information, our historian partners proceeded with organizing it in a list of time periods of interest.

Furthermore, the historians combined the two subject classifications, selected a set of inclusive subjects, and clustered them in the following six sets:

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

We call such subject clusters, systematically used for historic research in the area of science and technology, *historiographical issues*. Table 1 includes the full list of the subjects selected and the subject clusters that they contain.

**Table 1. Historiographical Issues: General History Ontology Subjects**

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

These clusters were then further analyzed to produce an extended list of concept candidates which were arranged into a set of concepts, instances and relationships to be inserted in the ontology.

### 3.3 Extending CIDOC CRM with Time and Evolution Constructs

In the historical context, temporal references cannot often be expressed by exact time notation. Consider, for example, such references as “the beginning of the 80s'”, “Atomic Age” or “after the industrial revolution”. Therefore, in the Papyrus History ontology we need to provide a proper mechanism for dealing with such fuzzy time specifications [19]. Although CIDOC CRM with its modeling of time using the TimeSpan concept partly fulfills this requirement, in the context of the History ontology it is needed to be able to perform more complicated operations taking into account time notation in order to place events on a timeline.

For this purpose, we proposed a new data model for temporal information supported by the Papyrus ontology editor called TrenDS [22]. In this model, time is represented in terms of fuzzy intervals. An interval is a continuous period in time represented by two time points, its beginning and its end, which in turn are represented as intervals:

```
[[bb,eb],[be,ee]],
```

where the four values are specific time points in form of 8 digits for an interval of the beginning of some event [bb, eb] and for the end of this event [be, ee].

Thus, any entity or attribute in the ontology may be assigned an interval of its validity, by specifying the values for four time properties: “time:bb”, “time:eb”, “time:be”, “time:ee”. In the trivial cases, the values of bb=eb, be=ee, for instance when one needs to specify an interval between two exact dates [22/06/1942, 09/05/1945]. Another possibility is when all four values are equal, if the whole validity interval is exactly one day like [1/1/1908].

This notation allows expressing complex intervals like “20 century” and supports universal dates like, “now” and “always”, where “Now” is encoded as an interval close to the maximum positive integer number, and “Always” is encoded as the interval spanning from the maximum negative to the maximum positive integer.

For instance, the time interval shown below corresponds to the date [1908 year, Now], e.g. the interval of its beginning spans from 01/01/1908 to 31/12/1908 and its end is some point in time after that.

```
<time:bb rdf:datatype="http://www.w3.org/2001/XMLSchema#string">19080101</time:bb>
<time:eb rdf:datatype="http://www.w3.org/2001/XMLSchema#string">19081231</time:eb>
<time:be rdf:datatype="http://www.w3.org/2001/XMLSchema#string">2147483644</time:be>
<time:ee rdf:datatype="http://www.w3.org/2001/XMLSchema#string">2147483647</time:ee>
```

“Always” is encoded in the ontology as an interval between the two extremes of integer:

```
<time:bb rdf:datatype="http://www.w3.org/2001/XMLSchema#string">-2147483648</time:bb>
<time:eb rdf:datatype="http://www.w3.org/2001/XMLSchema#string">-2147483648</time:eb>
<time:be rdf:datatype="http://www.w3.org/2001/XMLSchema#string">2147483647</time:be>
<time:ee rdf:datatype="http://www.w3.org/2001/XMLSchema#string">2147483647</time:ee>
```

To represent the evolution relationship between entities [4], the ontology instantiates these properties by means of five evolution properties: “join”, “merge”, “detach”, “evolve”, “retract” and a “partOf” property.

The *evolve* operator is specified between two entities that have consequent intervals i.e. e1.eb = e2.be. In this case, the system inserts the “evolve” relationship between e1 and e2 and assigns the lifespan [e1.eb,e2.be] to it.

In case of entity merging, the first entity’s lifespan must be the same with the lifespan of the second one. The system creates a relation *merge* between those entities for the time point e1.ee (when the first entity ceased to exist).

The *detach* operators requires the second entity’s lifespan to be within the lifespan of the first one and the system creates a “detach” relation for the time point e2.bb (when the detached entity appeared).

To create a *join* between two entities it is necessary to have an overlap of the entities’ life spans. In that case, the system inserts two attributes: a “join” relation between these two entities for the time point of join (specified by the user) and a “partOf” relation with the lifespan from the time of join creation to the end of the shortest entity lifespan (among these two given).

In case of *retract*, the system checks whether there has been already created a join, of so than it just breaks the lifespan of the existing “partOf” (till the moment of retract creation) and moreover inserts a “retract” relation with the time of its creation (specified by the user). If there is no existing “partOf” for these two entities, the system notifies about an error.

### 3.4 Extending the CIDOC CRM with Language Constructs

One of the key requirements for the Papyrus system was the implementation of a multilingualism model, given that the historians pose queries containing terms in their mother tongues. These terms are used for the description of real-world entities under specific conditions (i.e. time, place, language, dialect, domain, historiographical issues, formality and diatype). The evolution of the term semantics is different under different sets of conditions.

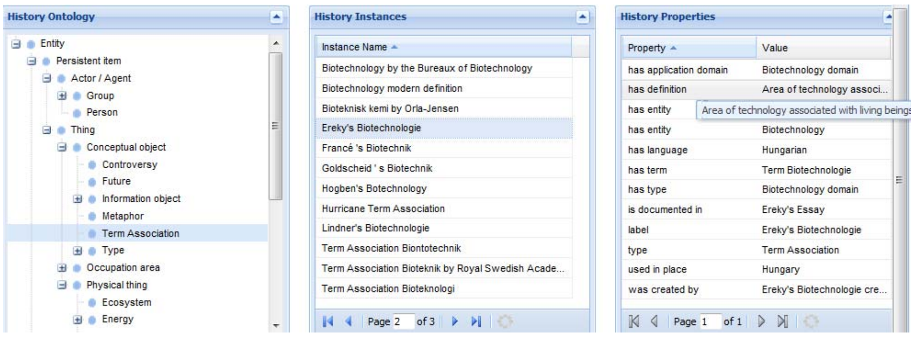
To this end, we developed a multilingualism model [22] that allows associating terms with entities under specific contexts (i.e. sets of conditions). This model includes: (a) The "MultilingualTerm" class (subclass of "E73.Information_Object"), which specifies a term name, and (b) The "TermAssociation" class (subclass of "E28.Conceptual_Object"), which associates the term with an ontology entity and carries the details on the context of the association.

The idea is that every term value is associated to some real world entity, which is generic and independent of any language or any other factor. The term, however, is associated to the entity only under certain conditions that are determined through a set of parameters representing a context. The context dimensions that the ontology developer can specify are represented as properties of the TermAssociation class (see an example in Figure 2) named as follows:

- *has_language* of range "E56.Language", which specifies the language of an association,
- *used_in_place* of range "E53.Place", which specifies a location where the term is used, since even within the same language the terminology can develop in a different way depending on the location (e.g., the word "truck" in American English refers to the same thing as the word "lorry" in British English),
- *has_historiographical_issue* of range "Historiographical_issue", which can be used to relate the term with a specific historiographical issue that specifies cultural or social conditions,
- *has_time* of range "E4.Period", which contains the validity interval for a term, for instance what was called in the middle of the 20th century an "electronic brain" in English is now referred to as "computer".
- *has_application_domain* of range "Domain" (subclass of E55.Type), for instance, philosophy, databases, or biotechnology,

- *has_dialect* of range “Dialect” (subclass of “E56.Language”) ,
- *has_formality* of range “Formality” (subclass of “E73.Information_Object”), which can range from Very_formal to Very_informal.
- *has_diatype* of range “Diatype” (subclass of “E56.Language”), which is the language variation, determined by its social purpose, under which the association is valid. It mainly relates to the channel of communication, such as spoken, written or signed.
- *has_confidence* of range float is a confidence score from 0 to 1. Confidence scores may be used in situations where a real-world entity can be associated with different terms even within the same context.
- *has_entity* of range “rdf:Class”, which defines a relationship between the term association and one or more entities of the ontology.
- *has_term* of range “MultilingualTerm”, which defines a relationship between the term association and a multilingual term.

This model allows for locating the corresponding entities in the history ontology by looking for a term either in a specific context or in a context-independent way. The mappings between the history ontology and the news ontology will be then used to locate the news ontology terms. Based on the news ontology terms, the appropriate news items are retrieved and returned to the historians.



| Instance Name |
|---|
| Biotechnology by the Bureaux of Biotechnology |
| Biotechnology modern definition |
| Bioteknisk kemi by Orla-Jensen |
| Ereky's Biotechnologie |
| Francé 's Biotechnik |
| Goldscheid ' s Biotechnik |
| Hogben's Botechnology |
| Hurricane Term Association |
| Lindner's Biotechnologie |
| Term Association Biontotechnik |
| Term Association Bioteknik by Royal Swedish Acade... |
| Term Association Bioteknologi |

| Property | Value |
|---|---|
| has application domain | Biotechnology domain |
| has definition | Area of technology associ... |
| has entity | Area of technology associated with living beings |
| has entity | Biotechnology |
| has language | Hungarian |
| has term | Term Biotechnologie |
| has type | Biotechnology domain |
| is documented in | Ereky's Essay |
| label | Ereky's Biotechnologie |
| type | Term Association |
| used in place | Hungary |
| was created by | Ereky's Biotechnologie cre... |

**Figure 2. Term association example**

### 3.5 Introducing New Concepts and Instances

In order to model the History domains used by Papyrus as a case study, the History ontology has been extended with concepts and instances, both general and domain-specific. Our interest in Papyrus was to represent some of the vital and urgent technology issues with focus on biotechnology and renewable energy.

Biotechnology embraces such crucial subjects as genome, DNA, genetically modified organisms, and cloning; while renewable energy includes the subjects related to alternative energy sources like wind energy, climate change, and environmental protection.

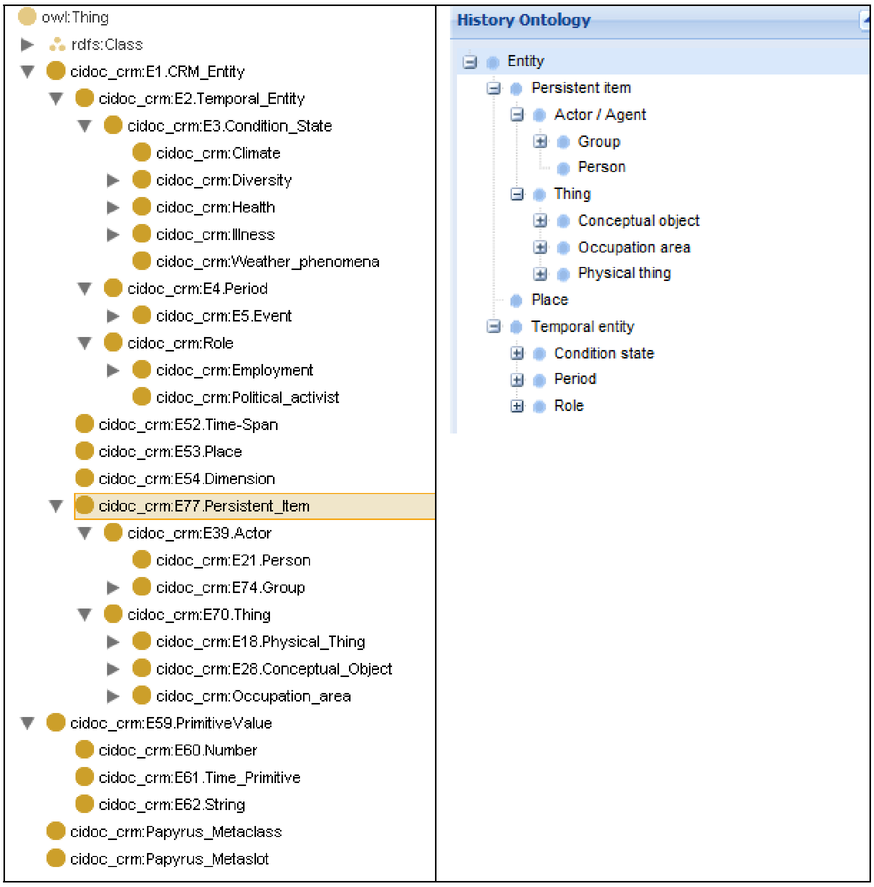
In order to identify concepts important in the two selected domains, we first undertook research in journals that host articles on the way these technological and scientific areas have been covered by the media. Tens of books and articles were selected for each area, the most relevant being those suggested to us by journals like Science Communication, Public Understanding of Science, Science, Technology and Human Values and Social Studies of Sciences. An example of the most relevant literature that we were able to identify is that of Miltos Liakopoulos, on the metaphors used in the media coverage of biotechnology [13]. Based on articles like this, we extracted the clusters of concepts-keywords (concepts if seen from the perspective of History Ontology; keywords if seen from the perspective of the news archives content) for each area-domain of technology and science under consideration.

In order to properly represent the identified concepts, we have been considering existing knowledge bases as Wikipedia [23] and Open CYC [15]. Most of the domain knowledge was represented under the concept tree of E77.Persistent_Item. We also added relationships between domain-dependent concepts where possible.

An example is E40.Legal_Body, which has been sub-classed with a set of organizations, institutions, e.g., “Company”.

### 3.6 Labeling Concepts and Instances

An important step for our CIDOC CRM – based History ontology has been the addition of labels to all concepts. This was decided in order to hide from the end users the CIDOC-CRM concept coding (a concept’s number before its name in the identifier, e.g., *E28.*Conceptual_Object) as well as the underscore (“_”) symbols used in the names and simplify the concept names. To assign such user friendly labels, the rdfs:label attribute was used. Accordingly, one can see concept labels visualized in the Papyrus ontology browser vs. Protégé in Figure 3.



- owl:Thing
- rdfs:Class
- cidoc_crm:E1.CRM_Entity
  - cidoc_crm:E2.Temporal_Entity
    - cidoc_crm:E3.Condition_State
      - cidoc_crm:Climate
      - cidoc_crm:Diversity
      - cidoc_crm:Health
      - cidoc_crm:Illness
      - cidoc_crm:Weather_phenomena
    - cidoc_crm:E4.Period
      - cidoc_crm:E5.Event
    - cidoc_crm:Role
      - cidoc_crm:Employment
      - cidoc_crm:Political_activist
  - cidoc_crm:E52.Time-Span
  - cidoc_crm:E53.Place
  - cidoc_crm:E54.Dimension
  - cidoc_crm:E77.Persistent_Item
    - cidoc_crm:E39.Actor
      - cidoc_crm:E21.Person
      - cidoc_crm:E74.Group
    - cidoc_crm:E70.Thing
      - cidoc_crm:E18.Physical_Thing
      - cidoc_crm:E28.Conceptual_Object
      - cidoc_crm:Occupation_area
- cidoc_crm:E59.PrimitiveValue
  - cidoc_crm:E60.Number
  - cidoc_crm:E61.Time_Primitive
  - cidoc_crm:E62.String
- cidoc_crm:Papyrus_Metaclass
- cidoc_crm:Papyrus_Metaslot

History Ontology

- Entity
  - Persistent item
    - Actor / Agent
      - Group
      - Person
    - Thing
      - Conceptual object
      - Occupation area
      - Physical thing
  - Place
  - Temporal entity
    - Condition state
    - Period
    - Role

**Figure 3. History ontology main tree based on CIDOC CRM in (a) Protégé and (b) Papyrus web browser**

Apart from the simple cases like the concept E12.Production, to which the label "Production" has been added, in other cases the label has slight differences. In the case of the concept E22.Man-Made_Object, for example, the label set is "Human-made object" in order to make the concept label less gender-specific.

## 4 The News Ontology

The News ontology [10, 11] was developed within Papyrus in close cooperation with news professionals working in AFP and is intended to describe the structure and the semantics of the news content. The ontology was constructed based on the NewsML-G2 XML standard [13] designed by the IPTC for conveying and annotating news content [8]. The purpose of this standard is to provide a model for the description of news items, and their related topics and keywords. It is used by major news providers like EBU [7] and Reuters Media [17].

For the needs of the Papyrus project, we integrated two different parts in the ontology: (a) the modeling of the format in which news items are produced by the main news agencies, i.e., the constructs adopted from NewsML-G2 (the presentation of this part of the ontology is omitted in the present paper; more information can be found in [10], [21]), and (b) the modeling of concepts present in the news items and relevant to the application domains, i.e., Biotechnology and Renewable Energy. These include named entities, concepts to accommodate domain-specific concepts, and instances. We further discuss the basic structure of the Papyrus extension of the ontology.

In the extended model, shown in Figure 4, each news item is identified by its URI and can have a list of related topics that may contain: *themes* – IPTC categories to be respected by the news agencies when annotating their news content [9], as well as domain-dependent – and *terms*, such as *named entities* (like Person, Organization, Location), *concepts* (other entity types), or *slugs*, i.e., terms defined as relevant to the IPTC subjects. In turn, each term can be defined by a set of keywords. Thus, a news item has a rich set of metadata, for instance a theme "Cloning", a location "Seoul", an event "press-conference", a person "Hwang Woo-suk", etc.

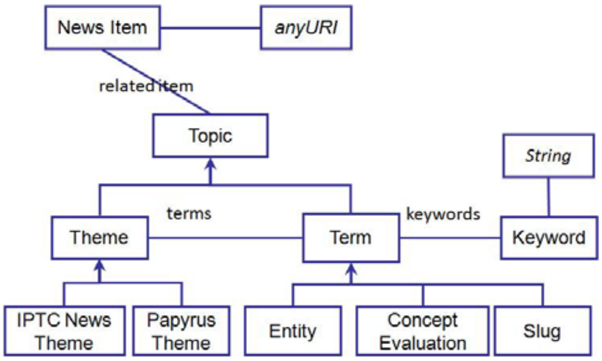


**Figure 4. News ontology model for annotating news items. Arrows represent *is-a* relations and named arrows role ones**

In more detail, *Topic*, *Theme* and *Term* are abstract concepts in this model and their underlying concepts are:

- *IPTCNewsTheme*. In the News ontology we adopted those IPTC categories that can be important for two application domains of Papyrus: biotechnology and renewable energy. To do so, the AFP experts manually selected a subset of IPTC topics that may contain information pertaining to either Renewable Energy or Biotechnology areas. As a result, 280 instances of this class have been included in the News ontology.
- *PapyrusTheme*. In order to represent more specific domain knowledge that is not represented by the IPTC categories, we created a new class, PapyrusTheme. All domain-related topics have been represented as instances of this class, starting from the two main topics of interest: Renewable energy and Biotechnology, and then their subtopics, such as Cloning, Stem cells,

Hydrogen energy, Environmental protection, and others. In order to support is-a relations between the instances, we exploited *skos:broader* and *skos:narrower* properties. Where relevant, we also linked IPTC news topics to one of Papyrus domains using *skos:sameas* property. So far the News ontology contains 32 Papyrus Themes.

- *Entity*. The News ontology was largely populated with varied types of named entities. The taxonomy of named entities extends the usual three classes – Organization, Location, Person (2,820 instances). The Location class of entities is represented by "GeoArea" concept, while Person and Organization are grouped under a more general concept called "Party". Apart from these common types of entities, we added the concepts of "Event", "Landscape", and "POI" (Point of interest) that includes, for instance, monuments.
- *ConceptEvaluation*. Instances of this concept are used to group several single keywords under one entity (e.g., "rotor blades", "rotor blade", "blades"). At the moment the ontology contains 6,930 ConceptEvaluation instances.
- *Slug*. This construct is inherited from the IPTC categorization, where each IPTCNewsTheme can be assigned one or more slugs, i.e., relevant terms. In total, 205 slug instances were selected given the two Papyrus domains.
- *Keyword*. Finally, the Keyword concept stores natural language expressions related to varied Term types. The total number of instances is around 30,000.

Thus, a (Papyrus or IPTC) theme instance can be related to a set of Entities, ConceptEvaluations or Slugs by means of "terms" relationship, where these are defined by sets of Keywords.

## 5 Mappings

The mappings are the tools for bridging information across the two domains. The Papyrus mapping framework adopts an entity-based data model, which is based on a dataspace data model like the one in [6] making the entity the primitive mapped unit. This facilitates the formulation of the information modeling and of the mappings, since it is conceptually closer to the way humans are thinking. In terms of the mapping language, Papyrus adopts an entity-based language that is similar, in spirit, to logical languages like Datalog. The semantics of the language are, however, fundamentally different from Datalog in that it allows the definition of mappings even in the absence of any schema information. More information on the mapping framework may be found in [3].

An example of a simple mapping is the following:

```
'history:Cloning'(),'history:Ethics'() --> , 'news:Concept_Cloning_00085'()
```

The domain expert, in this case, has defined that when the user is interested in the historiographical issue "Ethics" in relation to "Cloning", one of the related news ontology concepts to be retrieved will be "Bioethics committee" (Concept_Cloning_00085). To support the construction of complex expressions as mappings, Papyrus offers a graphical mapping interface in which the ontology information is presented as a set of entities [3].

Mapping the News ontology entities to History ontology tasks has also been a difficult issue. These mappings are so far performed mostly manually, which is a very time-consuming task. We have been working towards tools that may propose simple mappings to the user, which she may validate or reject. Advanced and intelligent mapping tools are needed to achieve greater automation of this process.

## 6 End User Perspective

Having modeled our two domains, History and News, and defined the way to bridge them, the next step was to create tools that would allow the historians to access and explore this rich material.

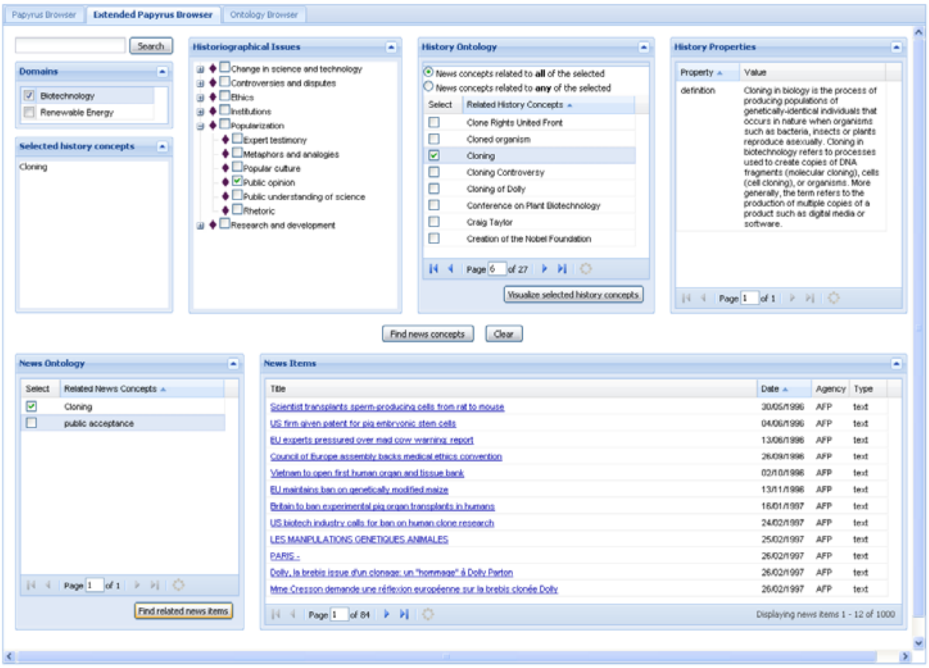
Taking in account the need for both search and browsing of the information offered through the system, we designed and implemented two different ways to access the same material: a browsing tool, called the *Papyrus Browser*, and an advanced search mechanism, *Cross-Discipline Search*. These are presented in the following sections.

### 6.1 The Papyrus Browser

The Papyrus Browser [16] is a tool that allows the exploration of news content through its association with the News ontology metadata and the corresponding mappings of this metadata to the History ontology. Besides its ability to be used as a simple Web-based ontology browser, it is a specialized tool that allows us to browse two different domain ontologies in parallel, as well as the content they describe. Combined with the keyword search functionality over the History ontology, the tool enables historians to research effortlessly both primary (News ontology and content) and secondary (History ontology) material. The Browser aims to provide the following views to the two ontologies, available in different tabs.

**Papyrus Browser View** (or Historiographical Issues View). In this view the user may filter the History ontology entities by selecting a domain and historiographical issue. Then, by selecting a History ontology entity, related News ontology entities will be displayed and then related news items.

**Extended Papyrus Browser View.** The extended view offers similar functionality to the Papyrus Browser View, extended with some distinctive features. In this case historiographical issues do not filter History ontology entities. Instead, the user may select one or more historiographical issues and one or more concepts, and then view related News ontology entities through mappings. Furthermore, it allows multiple selections of News ontology entities for refining news items retrieval. An example screen of this view is shown in Figure 5.



Domains

☑ Biotechnology
□ Renewable Energy

Selected history concepts

Cloning

Historiographical Issues

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

History Ontology

● News concepts related to **all** of the selected
○ News concepts related to **any** of the selected

| Select | Related History Concepts |
|---|---|
| □ | Clone Rights United Front |
| □ | Cloned organism |
| ☑ | Cloning |
| □ | Cloning Controversy |
| □ | Cloning of Dolly |
| □ | Conference on Plant Biotechnology |
| □ | Craig Taylor |
| □ | Creation of the Nobel Foundation |

Page 6 of 27

History Properties

| Property | Value |
|---|---|
| definition | Cloning in biology is the process of producing populations of genetically-identical individuals that occurs in nature when organisms such as bacteria, insects or plants reproduce asexually. Cloning in biotechnology refers to processes used to create copies of DNA fragments (molecular cloning), cells (cell cloning), or organisms. More generally, the term refers to the production of multiple copies of a product such as digital media or software. |

Page 1 of 1

News Ontology

| Select | Related News Concepts |
|---|---|
| ☑ | Cloning |
| □ | public acceptance |

Page 1 of 1

News Items

| Title | Date | Agency | Type |
|---|---|---|---|
| Scientist transplants sperm-producing cells from rat to mouse | 30/05/1996 | AFP | text |
| US firm given patent for pig embryonic stem cells | 04/06/1996 | AFP | text |
| EU experts pressured over mad cow warning report | 13/06/1996 | AFP | text |
| Council of Europe assembly backs medical ethics convention | 26/09/1996 | AFP | text |
| Vietnam to open first human organ and tissue bank | 02/10/1996 | AFP | text |
| EU maintains ban on genetically modified maize | 13/11/1996 | AFP | text |
| Britain to ban experimental pig organ transplants in humans | 16/01/1997 | AFP | text |
| US biotech industry calls for ban on human clone research | 24/02/1997 | AFP | text |
| LES MANIPULATIONS GENETIQUES ANIMALES | 25/02/1997 | AFP | text |
| PARIS - | 26/02/1997 | AFP | text |
| Dolly, la brebis issue d'un clonage: un "hommage" à Dolly Parton | 26/02/1997 | AFP | text |
| Mme Cresson demande une réflexion européenne sur la brebis clonée Dolly | 26/02/1997 | AFP | text |

Page 1 of 84

Displaying news items 1 - 12 of 1000

**Figure 5. Extended Papyrus Browser view – "Public opinion related to Cloning"**

**Ontology Browser View**. This view is addressed to the Papyrus ontology designer and administrator. It allows the user to browse the hierarchy of both ontologies, select entities, and edit them in the Papyrus ontology editing tool, or use them in a mapping.

Figure 5 illustrates how the user may get news items on the public opinion related to Cloning by selecting the appropriate historiographical issue ("Public opinion") and concept ("Cloning").

News ontology entities like "public acceptance" are retrieved through the mappings and the user may select to see one or more related news items.

## 6.2 The Papyrus Keyword Search

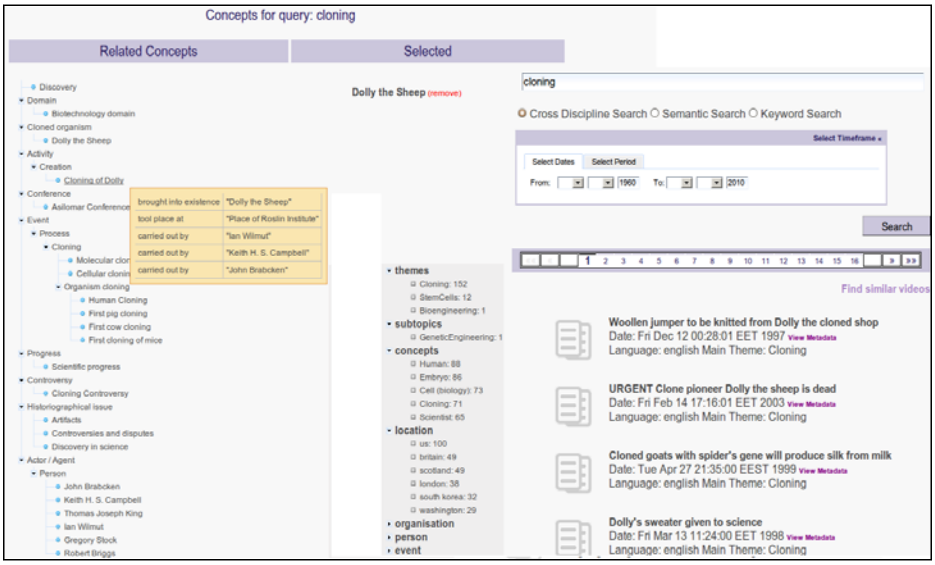
The keyword search tool implements one of the main functionalities of the platform for facilitating historical research. It allows the user to query the History ontology, study the returned History ontology entities, which provide the context, i.e., the secondary information, related to her research need, and then, retrieve related news items for the selected entities.

To accomplish this, Papyrus makes the following contributions:

- Extends the approach of [19] on keyword querying over RDF by adding a temporal dimension, adding support for concept evolution, and implementing several optimizations [12].

- Adds an extra step, where the user may select some of the results and, through the mappings, retrieve related news content.
- Provides a simple and intuitive user interface tailored to the needs of the non-computer expert users.

As an example, let us consider the keyword query “*cloning 1960-2010*”, which might be posed by a historian when she researches issues related to Cloning during the past fifty years. The semantics of an answer to such a query is to retrieve related information concerning the keywords, which is valid on the specified time. The temporal dimension in the query can be specified in terms of a date (e.g., “1960/05/15”), a time interval (e.g., “1960 – 2010”), a specific historical period (e.g., medieval), and in relation to all the aforementioned time elements (e.g., “*after* 1960/05/15”). To this end, all Allen’s temporal relations [1] have been implemented.
1.



Concepts for query: cloning

cloning

○ Cross Discipline Search ○ Semantic Search ○ Keyword Search

Select Timeframe: From: 1960 To: 2010

Selected: Dolly the Sheep (remove)

| | |
|---|---|
| brought into existence | "Dolly the Sheep" |
| tool place at | "Place of Roslin Institute" |
| carried out by | "Ian Wilmut" |
| carried out by | "Keith H. S. Campbell" |
| carried out by | "John Brabcken" |

**Woollen jumper to be knitted from Dolly the cloned shop**
Date: Fri Dec 12 00:28:01 EET 1997 View Metadata
Language: english Main Theme: Cloning

**URGENT Clone pioneer Dolly the sheep is dead**
Date: Fri Feb 14 17:16:01 EET 2003 View Metadata
Language: english Main Theme: Cloning

**Cloned goats with spider's gene will produce silk from milk**
Date: Tue Apr 27 21:35:00 EEST 1999 View Metadata
Language: english Main Theme: Cloning

**Dolly's sweater given to science**
Date: Fri Mar 13 11:24:00 EET 1998 View Metadata
Language: english Main Theme: Cloning

**Figure 6. Querying History ontology with keywords “cloning 1960-2010” (upper-left and upper right), and retrieving results according to the selected entities (lower-right)**

Going back to the example, our history researcher would like to proceed with searching information about the history of Cloning. She firstly focuses on the beginning of the 20th century and attempts to retrieve any information relevant to Cloning research, firstly until the 1960s. Among the history entities returned, she discovers that the first cloned animal actually was a frog, and this took place in 1952.
Extending the period to cover the whole century, she discovers information on the first cloned sheep, Dolly, as well as other Cloning events that followed, e.g., that of the first Cloning of mice in 1998 (Figure 6). Scientists involved in these events, e.g., Ian Wilmut, also appear in the results. By selecting one or more of the result entities, for example, Dolly the Sheep, she may retrieve news items that refer to it.

## 7 Conclusions and Future Work

This paper has presented the Papyrus approach in providing tools to support historical research in news archives. Papyrus allows the history researcher to explore both primary and secondary sources which have been structured and unified through their respective domain ontologies, the History and the News one. Bridging these two domains addresses a very important user need, that of bringing together two different sources. The benefits of this approach, which incorporates Semantic Web technologies have already been recognized by the history researchers.

One of the main challenges has been the modeling of the History ontology by appropriately extending the CIDOC-CRM standard and adding time and evolution support. The multilingual aspects of both the domains have also been studied and modeled in both the ontologies.

Our future work includes a large-scale evaluation of the Papyrus platform in its current state to record not only its advantages and shortcomings, but also study how users interact with this novel way to use archival material in supporting historical research.

**Acknowledgements:** This work was partially funded by the EU FP7 ICT framework.

## References

1. Allen, J.F.: Maintaining knowledge about temporal intervals. ACM, NY, USA. 1983
2. Bud, R.: Biotechnology in the Twentieth Century, Social Studies of Science, Vol. 21, No. 3 (Aug., 1991), pp. 415-457
3. Bykau, S., Kiyavitskaya, N., Tsinaraki, C., Velegrakis, Y.: Bridging the Gap between Heterogeneous and Semantically Diverse Content of Different Disciplines. In FlexDBIST-2010 (to appear).
4. Bykau, S., Velegrakis, Y., Kiyavitskaya, N.: Annex to D3.3: Theoretical foundations for time and evolution in the Editor/Mapper. 5 March 2010
5. Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M. (editors): Definition of the CIDOC Conceptual Reference Model, October 2006. The version 4.2.1 of the reference document
6. Dalvi, N. N., Kumar, R., Pang, B., Ramakrishnan, R., Tomkins, A., Bohannon, P., Keerthi, S., Merugu, S.: A web of concepts. PODS, pp. 1-12, 2009.
7. European Broadcasting Union: www.ebu.ch
8. Fernandez-Garcia, N., Sanchez-Fernandez, L.: Building an Ontology for NEWS Applications. In Poster Session of the 3rd Int. Semantic Web Conf. (ISWC 2004), Hiroshima, Japan, November 7-11, 2004.
9. IPTC categorization: http://www.iptc.org/NewsCodes/index.php
10. Kiyavitskaya, N.: Documentation on Papyrus ontologies, Technical report available at http://www.ict-papyrus.eu/files/Documentation%20on%20Papyrus%20Ontologies.pdf
11. Kiyavitskaya, N., Katifori, A., Paci, G., Pedrazzi, G., Turra, R.: The Papyrus News Ontology – A Semantic Web Approach to Large News Archives Metadata. To be published in proceedings of VLDL 2010, Glasgow, UK, September 10, 2010.
12. Lei, Y., Uren, V. S., Motta, E.: Semsearch: A search engine for the semantic web. In Steffen Staab, Vojtech Svátek, editors, EKAW, LNCS 4248, pp. 238-245. Springer, 2006.
13. Liakopoulos, M.: Pandora's Box or panacea? Using metaphors to create the public representations of biotechnology. Public Understanding of Science, Vol. 11, 5–32, 2002.

14. NewsML-G2 official website: http://www.iptc.org/cms/site/index.html?channel=CH0111
15. OpenCyc: www.opencyc.org
16. Platakis, M., Nikolaou, C., Katifori, A., Koubarakis, M., Ioannidis, Y.: Browsing news archives from the perspective of history: the papyrus browser historiographical issues view. WIAMIS 2010, Desenzano del Garda, Italy.
17. Papyrus Deliverable D2.2 – User requirements specification, http://www.ict-papyrus.eu/files/Papyrus-D2.2-v02.1.pdf
18. Reuters Media: www.reuters.com
19. Rizzolo, F., Velegrakis, Y., Mylopoulos, J., Bykau, S.: Modeling Concept Evolution: A Historical Perspective. In Proc. of ER 2009, pp. 331-345.
20. Tran, T., Wang, H., Rudolph, S., Cimiano, P.: Top-k exploration of query candidates for efficient keyword search on graph-shaped (rdf) data. In ICDE, pp. 405-416. IEEE, 2009.
21. Troncy, R.: Bringing the IPTC News Architecture into the Semantic Web. In Proc. of 7th Int. Conf. on The Semantic Web (ISWC'2008), LNCS 5318, pp. 483-498, Karlsruhe, Germany, 2008.
22. Tsinaraki C., Velegrakis Y., Kiyavitskaya N., Mylopoulos J.: A Context-based Model for the Interpretation of Polysemous Terms. In ODBASE 2010 (to appear).
23. Velegrakis, Y., Kiyavitskaya, N., Bykau, S., Tsinaraki, C.: TrenDSTMap: The TrenDS Mapping Tool. Ontology Mapper/Editor Manual. 10 September 2009
24. Wikipedia: www.wikipedia.org
25. Agence France Press, http://www.afp.com
26. Deutsche Welle, http://www.dw-world.de
27. Papyrus platform prototype: http://iris.atc.gr/CMS_Papyrus_1_1/
28. ISIS journal, http://www.journals.uchicago.edu/toc/isis/current
29. Technology and Culture Journal, http://etc.technologyandculture.net/