The Papyrus Digital Library: Discovering History in the News

A. Katifori[1], C. Nikolaou[1], M. Platakis[1], Y. Ioannidis[1], A. Tympas[1], M. Koubarakis[1], N. Sarris[2], V. Tountopoulos[2], E. Tzoannos[2], S. Bykau[3], N. Kiyavitskaya[3], C. Tsinaraki[3], and Y. Velegrakis[3]

[1] University of Athens, Greece
[2] Athens Technology Center S.A., Greece
[3] University of Trento, Italy
vivi@di.uoa.gr

Abstract. Digital archives comprise a valuable asset for effective information retrieval. In many cases, however, the special vocabulary of the archive restricts its access only to experts in the domain of the material it contains and, as a result, researchers of other disciplines or the general public cannot take full advantage of the wealth of information it offers. To this end, the Papyrus research project has worked towards a solution which makes cross-discipline search possible in digital libraries. The developed prototype showcases this approach demonstrating how we can discover history in news archives. In this demo we focus on demonstrating two of the end user tools available in the prototype, the cross-discipline search and the Papyrus browser.

Keywords: cross-discipline digital library, ontologies, keyword search, ontology browsing, multilingualism.

1 Introduction

In the last few years digital libraries have emerged providing electronic access for many user communities to information of their discipline. However, in many cases experts of one discipline turn to archives created by another discipline in the context of their research. An example of this need is the historical science, which takes advantage of archives, either cultural, scientific, press or personal, to discover information that will provide a better understanding of past events. The main problem in this process is the possible difference in the vocabulary of the historical researcher to that of the domain of the archive. This problem is related with specific challenging issues relevant to several vital research areas: coping with differences in terminology and its temporal aspects, developing techniques for semantic annotation and mappings, elaboration of query and presentation techniques for contemporary end users consuming archive information, and mitigating scalability issues. Vast amounts of digital content are available and could be incredibly useful to many user communities if it could be presented in a comprehensive to them way. The Papyrus

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 465–468, 2011.

project[1] approaches this need by introducing the concept of a Cross-Discipline Digital Library Engine. It intends to build 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 in this case means analyzed and modeled according to a domain ontology. The user query also has to be 'understood' and analyzed 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'. Papyrus showcases this approach by using two domain ontologies, the history ontology as the user one and the news ontology as the content one. News archives are a major source for primary material for history researchers of different topics, ranging from political history to the history of science. This demonstration will focus on two of the tools that Papyrus offers for the end user, the Papyrus browser and the Cross-discipline search functionality, as well as on the Papyrus ontologies.

2 The Papyrus Digital Library

The conceptual flow of the Papyrus DL is depicted in Fig. 1. ***Multimedia Analysis*** includes all components that operate on the content in order to semantically annotate it with concepts of the content (news) domain ontology [5]. ***Ontology Editing and Mapping*** groups the modules which provide all the operations for building the two domain ontologies, for defining the semantic correspondences between them and for

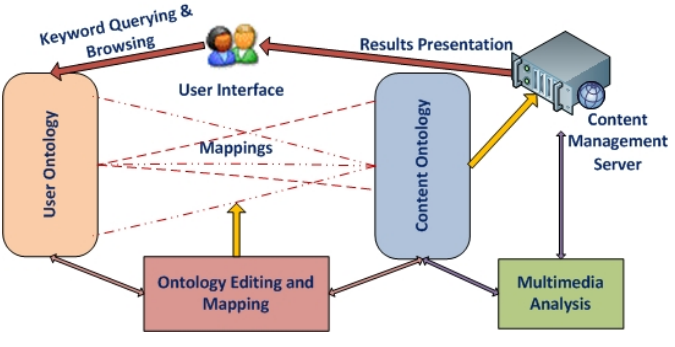

Fig. 1. The Papyrus Digital Library Engine conceptual flow

[1] FP7-ICT-215874 Papyrus Project: Cultural and historical digital libraries dynamically mined from news archives, www.ict-papyrus.eu, May 2007. The Papyrus platform was partly funded by the European Commission under the 7th Framework Programme.

semantically interpreting the user queries according to the user (history) domain ontology [4]. The ***User and Content ontologies*** [2, 3] are correspondingly history and news ontologies and the mappings provide the correspondences between them. The two ontologies have been modeled based on existing standards (CIDOC-CRM[2] and the IPTC[3] respectively) in collaboration with experts of the respective disciplines. **The *Results presentation* layer** provides the means for interfacing with the end user and accessing the underlying functionalities. ***Keyword querying and browsing*** is responsible for retrieving the information the user requests either by exploring visually the ontologies with the Papyrus browser, or by keyword search. This demonstration focuses on the two functionalities that take advantage of the history ontology to retrieve news items along with historical information: the Papyrus browser and cross-discipline search.

3 Keyword Querying and Browsing

The Papyrus end user tools to be presented in this demonstration are the Papyrus browser, a visual exploration tool that provides unified access to the two ontologies and the news content, and the Cross-discipline search functionality, which implements an ontology keyword querying technique through a visual interface.

The **Papyrus browser** [1] allows exploring news content through its association with the News ontology concepts and the corresponding mappings of these concepts to the History ontology. Besides its ability to be used as a simple Web-based ontology browser, it is a specialized tool combining two different domain ontologies and the content they describe. We will show how we can firstly select one or more historiographical issues and concepts and then retrieve news ontology concepts and related content using the mappings (Fig. 2).

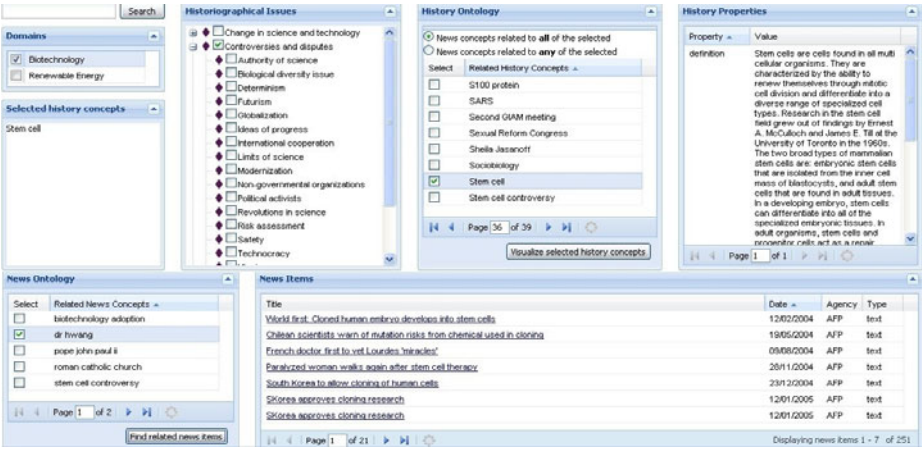

Fig. 2. Papyrus Browser– "Controversies on Stem-cells"

[2] http://www.cidoc-crm.org/
[3] http://www.iptc.org/

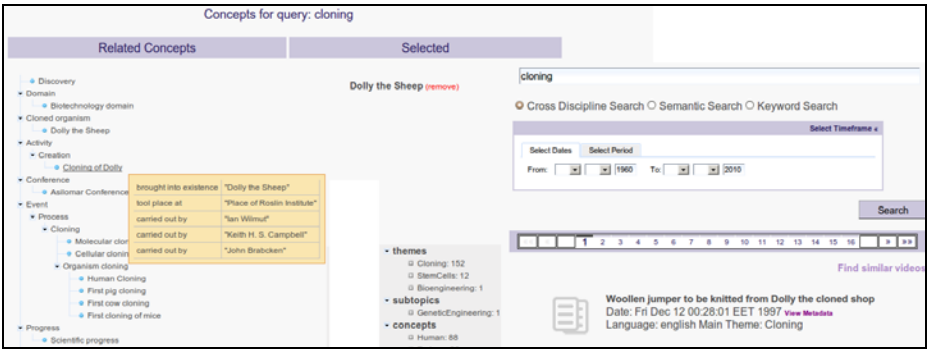

Fig. 3. Cross-discipline search - "*cloning 1960-2010*"

The **Cross-discipline search,** like the Browser, allows the user to query the History ontology, study returned History ontology entities providing the context, i.e., the secondary information related to her query, and then retrieve related news items for the selected entities. To do this, we employ an appropriate keyword search algorithm over the history ontology and the mappings between the ontologies. The query can be restricted to different time periods (Fig. 3).

4 Conclusions

The Papyrus Digital Library Engine is an integrated platform for cross-discipline search in digital archives made possible through state-of-the-art technologies. Papyrus bridges the gap between different knowledge domains and assists users in discovering information targeted to other audiences. Through the deployment of the system in the domains of history and news, Papyrus illustrates a practical example which may serve as a potential exploitable application on its own. Papyrus proves that it is possible to bridge different worlds and allow cross-discipline search through a careful indexing and mapping across their respective domains.

References

1. Platakis, M., Nikolaou, C., Katifori, A., Koubarakis, M., Ioannidis, Y.: Browsing News Archives from the Perspective of History: The Papyrus Browser Historiographical Issues View. In: WIAMIS, Desenzano del Garda, Italy (2010)
2. Kiyavitskaya, N., Katifori, A., Velegrakis, Y., Tsinaraki, C., Bykau, S., Savaidou, E., Tympas, A., Ioannidis, Y., Koubarakis, M.: Modeling and Mapping Multilingual and Historically Diverse Content. In: CIDOC, Shanghai, China (2010)
3. Kiyavitskaya, N., Katifori, Paci, G., Pedrazzi, G., Turra, R: The Papyrus News Ontology – A Semantic Web Approach to Large News Archives Metadata. In: VLDL, Glasgow, UK (2010)
4. Bykau, S., Kiyavitskaya, N., Tsinaraki, C., Velegrakis, Y.: Bridging the Gap Across Heterogeneous and Semantically Diverse Content of Different Disciplines. In: FlexDBIST, Bilbao, Spain (2010)
5. Paci, G., Pedrazzi, G., Turra, R.: Wikipedia based semantic metadata annotation of audio transcripts. In: WIAMIS, Desenzano del Garda, Italy (2010)