2nd ERCIM DIS Workshop
On Large scale and federated information spaces
Amsterdam, May 20, 2010
Position Paper

Collaborative information spaces: Research directions at IMIS

George Papastefanatos[1], Yannis Stavrakas[1], Christos Papatheodorou[2], Theodore Dalamagas[1], Dieter Pfoser[1], Timos Sellis[1]

"Athena", Research Centre, Institute for the Management of Information Systems, Athens, Greece
[1]{gpapas, yannis, dalamag, pfoser, timos}@imis.athena-innovation.gr
[2]papatheodor@ionio.gr

1. Introduction

The Institute for the Management of Information Systems (IMIS) is a research institute within the Research and Innovation Center in Information, Communication and Knowledge Technologies "Athena". The research at IMIS has a strong collaborative aspect, and ranges from basic to applied research. The collaborative aspect is expressed in that research is conducted with national and international partners from industry as well as academia, often also in the context of novel and innovative projects.

Data management in the context of information spaces is one of the primary research areas at IMIS. Data and information spaces is a growing research area aiming at exploiting collections of heterogeneous, highly distributed data which may vary in structure, origin and semantics. The amount and complexity of data increase in a variety of applications, such as enterprise data management, personal data management, large-scale scientific databanks, sensor networks, multimedia collections, and most importantly an increasingly structured and semantically rich social Web. Traditional data integration techniques cannot handle this diversity in data and meaningfully provide uniform access to heterogeneous data sources. Data and information spaces approaches provide methods, tools and platforms for modeling, understanding, semantically enriching, querying and furthermore integrating this wealth of data into a unified platform. Multiple research challenges are regarded in this context, including data and schema integration over heterogeneous sources, semantic mapping and matching, indexing and data mining techniques for dataspaces, data extraction and querying, etc.

Our research activities at IMIS cover a wide range of the above challenges, focusing mainly on the semantic enrichment, integration and retrieval of web and spatiotemporal data via user participation and collaboration. Thus, we aim at developing methods and tools that enable the collaborative management and integration of information spaces for a wide range of application areas. In short, this position paper will present the main research directions of IMIS with respect to information spaces. We enroll our efforts into the following topics:

- Advanced Retrieval using MindMap Models.
- Semantic annotation and search.
- Management of spatiotemporal data in information spaces.
- Personalized data management systems.
- Preservation and provenance in evolving interrelated information spaces.

- Modeling digital curation processes.
- Domain knowledge representation in the form of ontologies for ontology-driven search and fact discovery.

In the following sections we present in details each of the above directions.

2. Advanced Retrieval using MindMap Models

Web search engines are widely used for searching information on the Web. Their increased popularity is due to the simple but intuitive search model offered (i.e., keyword-based search). However, there are use cases where the information need is complex, and is part of a so-called creativity cycle. Consider for example a researcher that needs to set up her research agenda and generate innovative ideas. She often has the "big picture" for her search plan, that is an abstraction with thoughts, ideas and concepts that actually describes the domain to explore. Based on this initial abstraction, she starts gathering information from several data sources. She studies and generates hypothesis. Then, she refines her abstractions, searches for new information etc, starting a new cycle. Such cycles (creativity cycles) actually enable discovery and innovation.

New search models and techniques are necessary to promote creativity and innovation. Web applications for Bio-sciences is one of the use-cases adopted. A critical objective is to support creativity cycles, and also to provide effective presentation and visualization capabilities for the lists of retrieved resources that will guide users in their search and exploration during a creativity cycle:

[Search plan abstraction - Information harvest and retrieval - User personalization - Organizing information]

The scientific and techical challenges are the following [MindMap]:

- *Knowledge representation and abstraction*. Knowledge abstractions is a popular feature for mindmapping tools. Mind mapping refers to graphical representations of elements such as concepts, ideas, notes, tasks, or other items related to a topic of study. Mind mapping elements are organized in branches or groups according to the semantic interpretation given by the user. We study methods to exploit mind maps for gathering, organizing and exchanging information on the Web.
- *Web information harvest and retrieval*. In a metasearch paradigm, user queries are propagated to several search engines. Relevant resources are retrieved, merged, ranked and presented to user. For effective harvest and retrieval, intelligent services are needed to orchestrate the metasearch. Depending on the type of resources, certain search engines will be favored against other engines in order to answer the query. We are interested in (a) automated web scraping methods, (b) retrieval methods assuming data evolution and change, (c) dynamic facet extraction from result lists, and (d) using the semantics of mindmaps to improve the precision of the retrieval and harvest tasks. We put emphasis on searching for scientific material (e.g., papers, technical articles, etc).
- *User personalization*. Users with different backgrounds or viewpoints may interpret the same data in a different way. To avoid such ambiguous situations, searching may exploit the user context under which information becomes relevant in order to adapt the presentation of the retrieved resources to user needs (i.e., personalized ranking of the results). We are interested in adapting ranking lists to user interests, taking into consideration (a) users past search behaviour and (b) the semantics of mindmaps created by users in the past.

3. Semantic annotation and search

Semantic annotation and search tools are at the core of Semantic Web Technology. Annotations involve tagging of data with concepts (i.e., ontology classes) so that data becomes meaningful. Annotating data can help in providing better search facilities, since it helps users to search for information not only based on the traditional keyword-based search, but also using well-defined general concepts that describe the domain of their information need. A great number of approaches on semantic annotation have been proposed in the literature. Most of them are focused on annotating web resources such as html pages or plain text. As far as popular document formats are concerned, there are approaches that differ in the annotation and search facilities they offer.

We have developed a semantic annotation and retrieval tool, called *GoNTogle* [GBDS10]. GoNTogle supports manual and automatic annotation of several types of documents (doc, pdf, rtf, txt, odt, sxw) using ontology classes, in a fully collaborative environment. It also provides searching facilities beyond the traditional keyword-based search, using a flexible combination of keyword and semantic-based search. In contrast with other works, our aim was to implement an easy-to-use document annotation and search tool, that would fully support (a) viewing and annotating popular document types while maintaining their initial format, (b) sharing those annotations and (c) searching for documents combining keyword and semantic-based search [GoNTogle]. The key features of our tool are the following:

- It allows users to open and view widely used document formats such as .doc and .pdf , maintaining their original format.
- It provides an easy and intuitive way of annotating documents (or document parts) using OWL and RDF/S ontologies.
- It provides an automatic annotation mechanism based on models trained from user annotation history, so that annotation suggestions are tailored to user behavior.
- It is based on a server-based architecture, where document annotations are stored in a central repository. Thus, we offer a collaborative environment where users can annotate and search documents.
- It combines keyword and semantic search, providing advanced search facilities for both types of search.

4. Data provenance and preservation

Recently a lot of attention is drawn to the provenance and preservation of information. Provenance deals with tracing the origins and transformations of information in order to be able to assess its quality. Preservation deals with ensuring that data will be maintained and be always available, despite modifications in the data itself or evolution in storage technology. From an information system point of view provenance and preservation are closely related, as they both deal with change: provenance can be used to interpret data, an element which is essential in the preservation of knowledge.

The objective of our research is to support data management in evolving interrelated information spaces, in such way that it is always possible to step back (support for schema & data evolution) and examine how and why changes took place (provenance support). This is especially useful for spaces published on the Web, which are copied or referenced by other spaces, thus forming a web of interrelated and evolving systems. Such spaces may contain, for example, scientific information (such as the biological databanks UniProt, Rfam, IUPHAR), and are very important since they constitute a record of the evolution of a scientific field.

In order to incorporate provenance and preservation in such systems we need data models that can tolerate and record change, query languages that can reveal the data trails, and tools that can be adapted to and enhance existing operational systems.

5. Management of spatiotemporal data in information spaces

5.1 Geocoding of persistent Web content

Information and specifically Web pages may be organized, indexed, searched, and navigated using various metadata aspects such as keywords, categories (themes), and also space. While categories and keywords are up for interpretation, space represents an unambiguous aspect to structure information. The solution to the basic problem of providing spatial references to content is solved by geocoding; a task that relates identifiers in texts to geographic co-ordinates.

A methodology was developed for the semi-automatic geocoding of persistent Web pages, i.e., relating identifiers in texts to geographic co-ordinates using a combined automatic and human-centered approach [ALEP08]. Specifically we will focus on Greek Web pages and related geo information. The methods however are universally applicable. Automatic geoparsing and geocoding algorithms are successfully applied to identify phone numbers and addresses, however when more generic geo identifiers are involved, automatic algorithms produce a significant number of false positives (Venizelos as a person) and false negatives (Venizelos as the name of Athens international airport). This work advocates human intervention to improve on automatic geocoding results and develops therefore a Web browser extension that (i) allows for the manual geocoding of text portions and (ii) the updating, including deletion of automatically generated results. This proposed approach is especially helpful for persistent Web pages such as Wikipedia, i.e., pages that have a certain value to the community, are well cared for and change rather slowly. Here, geocoding can become a regular part of Web page authoring!

The geocoding of a Web page is stored in a central repository, i.e., a Web page is stored in terms of its URL, the geocoded text portions in terms of their position on the Web page and the respective co-ordinates, and the date of the geocoding, i.e., the version of the Web page. Further, the geocoding of a Web page is displayed as highlighted text and by means of a map, i.e., clicking on a text portions shows the respective position on a map. In our case, Google Maps was used for this task. The respective functionality is accessible through a Firefox browser extension. Currently, the system is deployed in cooperation with a Greek travel blog.

5.2 User-Contributed Content - Geoblogging Platform "GEOCROWD" - Augmented Reality

One cannot deny that space and time are important to us. We perceive our world with respect to where and when we do things. We advocate *geoblogging as a tool to capture such experiences by means of collecting and organizing images, audio, video and text in relation to space and time*. This application showcases a Web application that allows for a simple upload of content, geocoding, and map-based authoring of geoblogs. Export capabilities free the created content from a specific application context and allow for sharing and use of geoblogs in other applications, publications or social networking services. When coming home from a memorable journey, wouldn't it be great to create a digital replica of the trip, i.e., quickly organize collected images, videos, etc. and have a simple means of adding some thoughts? With our application, termed "*Geocrowd*", we propose geoblogging as a means for *spatiotemporal storytelling*, more specifically the story of a journey, be it an afternoon walk in your neighbor- hood, a chase for a great coffeeshop, your mountainbike trip, or hiking adventure [GEOCR].

The scope is to provide a simple to use application that allows one to tell the story of the trip based on the content collected during the trip. The role of content is to support the story. In our application, the essential aspects are a *map*, a *storyboard* and a *timeline*.

We see geoblogging as a means to harness the ability humans have to massively collect and share knowledge (i.e., consider conventional blogging and other Web 2.0 phenomena) for the spatiotemporal domain. The ultimate goal will be to *digitize the world using such user contributed content***.** As early maps were traces of people's movements in the world, i.e., view representations of people's experiences, digitizing the world in this context relates to collecting pieces of knowledge gained by a human individual tied not only to space and time, but also to her context, personal cognition, and experience.

6. Personalized Data Management Systems

Recently, a lot of research has focused on developing methods for *personalized search, organization and management of data*. Numerous applications and systems (such as search engines, social networks, targeted advertising, etc.) seek to provide personalized services to their users, using information on their demographics, interests, etc. Furthermore, the presence of devices (e.g., mobile phones, low cost sensors) and services (e.g., Web 2.0) make it possible to collect and record data and preferences for the current state of users.

The subject of our research is to design and develop advanced models, algorithms and techniques that will allow the personalization of systems that manage information. The techniques developed are examined in accordance with the following directions. (a) *Contextual information*, i.e., data concerning the current state of a user, such as environment, time, interests and demographic characteristics, the means of access to the data, etc. (b) User *preferences* that have been indirectly extracted from previous queries or from other users with similar interests. (c) The *tasks* and desires of users. These technologies are used as the basis for developing ranking algorithms that respect the current situation as well as the preferences and the tasks of users.

Specifically, with respect to task computing, our goal is to develop and evaluate a complete framework that will enable the task-aware provision of content to mobile users. Task computing will be used to structure content and to provide meaningful information on mobile computing devices. Here, we bridge the gap between traditional LBS and general Web content in that task computing will provide purposeful, rich, geo-enabled Web content to mobile devices[TALOS].

7. Digital Curation

Digital curation services are envisaged as a main mechanism of ensuring the preservation of the evidential function of records as the vocabularies, methods and frames of reference in science and scholarship evolve. Besides digital archiving and preservation, activities also cover developing and supporting standards and methodologies, providing semantically enhanced information and communication services based on curated collections, and mentoring and "incubating" best practice among organizations [Co++09].

The Digital Curation Unit of IMIS is addressing the needs of a wide constituency of organisations and communities, in fields as diverse as e-government, organisational records management, research repositories in the sciences and humanities, and digital heritage. Among these, the DCU has developed MOPSEUS, a powerful repository management and digital preservation platform that support information providers in managing their digital assets. Mopseus, is built on top of Fedora-commons middleware

and is designed to facilitate medium and small institutions to develop and preserve their own repositories. In comparison to the Fedora-commons platform, Mopseus provides a ready-to-use repository system, without the need of customization and the programming workload that Fedora-commons involves [ACGP09]. In Mopseus indexing is efficiently performed using a RDMS. Additionally, a major characteristic of Mopseus is its ease of installation and development of front and back ends.

The core of Mopseus is implemented as a set of Java services. Furthermore, an API, developed in PHP, allows for rapid development of back and front-end functionalities. The content of a Mopseus repository is stored as digital objects, consisting of datastreams, which can be text/xml, text/rdf or binary. Thus datastreams can be correlated to form digital objects that are structures of data and metadata. A digital object may be correlated to one or more containers using various types of relations. A container may include digital objects or other containers as well. The main Mopseus achitectural components are [GPCA10]:

- *Dynamic definition of XML schemas*. Mopseus provides a service for the definition of metadata schemas. The service supports the development of an XML schema, which defines the syntax of the metadata elements, their functionality (mandatory/optional elements) and presentation.
- *RDBMS Synchronization*. A mechanism was developed to dynamically synchronize all the elements of the hosted XML schemas with an external RDBMS database (currently MySQL). This process drastically improves the efficiency and flexibility of the indexing of any kind of XML or RDF document stored in datastreams.
- *Mapping between XML schemas*. This mechanism allows the mapping between metadata schemas.
- *Workflow engine*. The workflow engine allows for easy automation of simple tasks such as ingestion, revision, etc. Each workflow is encoded as an XML document, while a graphical interface guides the user to complete the task.

Regarding preservation, Mopseus is inspired by the OAIS model principles, in the sense that (a) the digital objects carry meaningful information about their binary content and relationships and (b) this representation information constitutes itself a digital object. Each digital object, as well as its relationships with other objects, is described in Mopseus by a set of datastreams, each of them being versionable. Furthermore, for every submission, a new version is created and all the previous versions are stored in Fedora's FOXML format. Moreover PREMIS metadata are automatically generated each time a digital object is generated / modified. Mopseus supports ingestion, access, storage, data management, administration and preservation planning OAIS functionalities. The ingestion/modification workflows are described by XML documents. Concerning preservation planning, Mopseus provides a migration process from existing repositories, facilitated through the use of a desktop tool implemented in Java. Currently it supports migration from DSpace repositories.

Mopseus is the repository management platform of CARARE (http://www.carare.eu), an EU funded Best Practice Network and aggregation service aiming to enrich Europeana digital content from the archaeology and architectural heritage domain. CARARE involves more than 30 heritage organisations from 20 countries, that will provide to Europeana information for 2 milion items concerning unique archaeological monuments, historic buildings and town centres. CARARE will work with the EDL Foundation to establish efficient processes to address practical issues relating to the harvesting of 3D/VR formats, the rich semantics of archaeology and architecture content and the handling of geographic information.

References

[ACGP09] S. Angelis, P. Constantopoulos, D. Gavrilis, C. Papatheodorou, "A Digital Library Service for the Small", Proceedings of the 2nd Digital Curation Curriculum Symposium. DigCCurr 2009: Digital Curation Practice, Promise and Prospects", Chapell Hill, NC, USA, April 2009.

[ALEP08] A. Angel, C. Lontou, A. Efentakis, D. Pfoser. *Qualitative Geocoding of Persistent Web Pages*. 16th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, 2008.

[Co++09] P. Constantopoulos, C. Dallas, I. Androutsopoulos, S. Angelis, A. Deligiannakis , D. Gavrilis, Y. Kotidis, C. Papatheodorou, "*DCC&U: An Extended Digital Curation Lifecycle Model*", International Journal of Digital Curation, Vol. 4(1), pp. 34-45, 2009.

[GBDS10] G. Giannopoulos, N. Bikakis, T. Dalamagas, T. Sellis. *GoNTogle: a Tool for Semantic Annotation and Search*. Extended Semantic Web Conference (ESWC'10), 30 May - 3 Jun, Heraklion, Greece (System Demo).

[GEOCR] C. Lontou, D. Pfoser. *Geoblogging – Storytelling beyond texts and images*. Web page: http://www.geocrowd.org.

[GoNTogle] *GoNTogle: A Semantic Annotation and Search Tool*. http://web.imis.athena-innovation.gr/~dalamag/gontogle/

[GPCA10] D. Gavrilis, C. Papatheodorou, P. Constantopoulos, S. Angelis, "Mopseus – A digital library management system focused on preservation", accepted poster ECDL 2010.

[MindMap] *Scrap and search with Freemind: an approach for creativity support.* http://web.imis.athena-innovation.gr/~dalamag/mm/.

[TALOS] *TALOS – Task aware location based services for mobile environments*. Project Web page: http://www.talos.cti.gr.