ESWC'06 Industry Forum

Business applications of Semantic Web challenge Research
Budva, Montenegro, 12 June, 2006
http://www.eswc2006.org

The goal of the Industry forum is to present to and to debate on the current and foreseen needs in Business, Industry and Services with the top leading research community and technological solution providers in Semantic Web technology.

Preface

This proceeding contains the presentations presented at the Industry Forum during the 3nd European Semantic Web Conference (ESWC 2006) held in Budva, Montenegor, 12th May 2006.

The goal of the Industry forum is to present the current and foreseen needs in Business, Industry and Services to the top leading research community and technological solution providers in Semantic Web technology. It constitutes a unique occasion to exchange and confront on concrete facts the promises of the Semantic Web with the leading stakeholders. Moreover associated to the ESWC conference it gives the opportunity to keep up with the latest research agenda developments.

The forum was designed to maximize cross-fertilization between research and industry. It shows concrete business cases needs from industry and case study presentations from developers, implementers and clients of the technology. The forum includes 10 invited presentations from key industrial leaders in diverse economical sectors, it also feature 2 keynote presentations from Research and Industry. The program of the forum includes presentations on the following topics:

- Market needs, Technology / Service challenge and vision
- Semantics needs in Enterprise and Internet
- Existing products and Semantic Web services envisioned solutions
- Key business benefits expected
- Business Use Cases
- Current practices and bottlenecks
- Business taxonomies and Vocabularies
- Current problems in content and unstructured data in legacy systems
- Current practices and difficulties in introducing knowledge technology
- Business rules

The ESWC 2006 Industry Forum has been strongly support by the FP6 NoEs Knowledge Web and REWERSE.

At the time for this proceeding to go in press, we would like to thank the International Program Committee for their effort in the reviewing process and guidance for the speakers presentation.

Our gratitude also goes to the entire conference organization committee and especially to Chriten Ensor and Ilona Zarembra for holding this together.

May 2006

Alain Léger, Andrea Kulas
Lyndon Nixon, Robert Meersman

Industry Forum Organization

Alain Léger, France Telecom, France
Andrea Kulas, webXcerpt, Germany
Lyndon Nixon, FU Berlin, Germany
Robert Meersman, VUB, Brussels, Belgium
Ruediger Klein, Daimler Chrysler, Germany
Massimo Paolucci, NTT DoCoMo Lab, Germany
Christine Golbreich, LIM-Hospital Rennes, France
Christian de Sainte Marie, ILOG, France
Ora Lassila, Nokia, Finland
Vagan Terziyan, University of Jyvaskyla, Finland
Amit Sheth,University of Georgia, and Semagix Inc.
Richard Goodwin, IBM, TJ Watson Research, USA

Program Committee

Chris van Aart, Y'all, NL
Richard Benjamin, ISOCO, Spain
Bertrand Braunschweig, Institut Franais du Ptrole, France
Elmar Dorner, SAP AG, Germany
Christian Fillies, Semtation Gmbh, Germany
Tim Geisler, webXcerpt Software GmbH, Germany
Jens Hagendorff, worldwidejobs, Germany
Mustafa Jarrar (VUB)
Yiannis Kompatsiaris, ITI-CERTH, GR
Ruben Lara, Tecnologia, Informacion and Finanzas, Spain
Sara Carro-Martnez, Telefonica, Spain
Pramilla Mullan, France Telecom R&D, USA
Andreas Persidis, Biovista, Greece
Alexander Polonsky, Cognium Systems SA, France
Jean Rohmer, Thales group, Franc
Jan Ross, Merrall-Ross International Ltd, UK (tbc)
Hans-Peter Schnurr, Ontoprise & Customer, Germany
Susie Stephens, Oracle Corporation, USA
Helen Kaykova, University of Jyvaskyla, Finland (tbc)
Luk Vervenne, Synergetics NV, Belgium
Paul Warren, British Telecom, UK
Ronald Wertlen, Neofonie Gmbh, Germany

Presentation Schedule

9:00 – 9:45	***Keynote 1*** **"Issues and strategies for digitization and preservation of the digital content"** Natasa Milic-Frayling, Microsoft
9:45 – 10:15	**"Is Semantic Web technology ready for Healthcare?"** Chris Wroe, BT Labs Health care, UK
10:15 – 10:45	**"Knowledge Management in the Petroleum Industry"** David Norheim, Computas, Norway
11:00 – 11:30	**"Contextual Intelligence for Mobile Services through Semantic Web Technology"** Massimo Paolucci, NTT DoCoMo, Germany
11:30 – 12:00	**"Semantic Laboratory Notebook: Managing biomedical research notes and experimental data"** Alexander Polonsky, Cognium sa, France
12:00 – 12:30	**"Semantic Web Technology Roadmap"** Roberta Cuel, University of Trento
14:00 – 14:45	***Key Note 2*** **"Semantic Business Automation"** Christian Drumm, SAP, Germany
14:45 – 15:15	**"News Agency needs : XML News"** Stéphane Guérillot, AFP, France
15:15 – 15:45	**"Training Management System for Aircraft Engineering :Indexing and Retrieval of Corporate Learning Objects"** Joanna Guss, EADS Airbus, Europe
16:30 – 17:00	**"Use of Ontology for production of access systems on Legislation, Jurisprudence and Comments"** Jean Delahousse, Mondeca, France
17:00 – 17:30	**"Data Integration using Semantic Technology : a Use Case"** Jurgen Angele, Michael Gesman, Software AG, Germany
17:30 – 18:00	**"Integrated Access to Biological Data. A use case"** Ainhoa Llorente, Robotiker, Spain

Preserving Information for Posterity: Is ‘Going Digital’ the Answer?

Natasa Milic-Frayling

Microsoft Research Ltd
Roger Needham Building, 7 J J Thomson Avenue, Cambridge, UK
natasamf@microsoft.com

Abstract. Digitization has been adopted as a strategy for preserving content of deteriorating physical artefacts. At the same time, by removing the boundaries of physical containment, it provides new opportunities for sharing and exploiting information. Generally, the use of digital format has dramatically changed how we create, manage, and communicate information. However, ensuring long term preservation of digital media is a non-trivial matter. Failing to find an adequate solution threatens the survival of valuable information created in the digital era. In this paper we reflect on the many issues associated with digitization and preservation of the digital content. We describe the economic climate that sets the context for increased activities in this area.

Keywords: digital library, digital archive, metadata, standards, XML, preservation.

1 Introduction

Proliferation of information and communication technologies has resulted in a dramatic shift from recording information in paper manuscripts, photographs, and audio-visual tapes to creating content in the digital format. Our ability to produce, publish, and communicate digital content with ease and in abundance, has transformed the way we view and manage information. We recognize the benefits of having relevant information at the right time and continue to develop technologies that enable us to exploit information optimally.

Through various digitization techniques, we transform the traditional documents into the digital format and exploit it alongside the ‘born digital’ content. In this paper we reflect on the increased value of information that stems from the highly agile nature of the digital format and the risk of loosing digital information unless we take appropriate steps to ensure its long term preservation. We discuss the current business climate that drives digitization and possible economic models that can make digitization a self-sustainable effort.

2 From Physical Artefacts to Digital Media

Through generations, information has been recorded to pass on the acquired knowledge and experience. Nowadays, information is mostly created in the digital form, even when it is disseminated as paper books, newspapers, and similar. The digital content has the advantage of being amenable to automatic analysis and aggregation. We can more easily create new knowledge with the aid of software analysis tools and disseminate information using online communicate channels. Thus, it is not surprising that the content of books, sound recordings, and video material have all been converted to the digital form. Digitization techniques essentially *liberate the content from its containment within physical objects*. The digital format significantly increases *information agility*, which in turns has significant social, economical, political, and legal implications on societies.

Furthermore, digitization has been adopted as a strategy for preventing the loss of information from deteriorating paper artefacts, film repositories, and sound recordings. Technological requirements for this process have instigated research in optical character recognition, speech recognition, video and image processing, and related areas. At the same time, the infrastructure requirements for managing the digitized data are pushing the limits of typical IT architectures and dramatically changing the cost structure.

The cost of digitization is significant. It may cost up to £1 to digitize and apply OCR to a single page of a newspaper. The national, regional, and international newspaper collection at the British Library alone contains approximately 750 million pages, some dated back to 18th Century. In order to make this effort economically feasible, we need to identify economic forces that can drive digitization of deteriorating information resources. Assuming that the public sector cannot absorb all the cost of this effort, what are the incentives for businesses to get involved in content digitization? While current or recent information is probably most valuable to businesses, most of it is copyright protected and thus needs to be handled appropriately. It is clear that the success of the effort depends on our ability to address many intricate issues.

In the following section we reflect on the recent activities in content digitization that have been influenced by a highly competitive on-line search market and the associated advertising business.

2.1 Business Climate and Digitization

Almost all major libraries have been engaging in digitization projects over the past decade. Such efforts are typically sponsored by national funding agencies or private donations. It is interesting that the recent boost to their digitization efforts came from investments by businesses involved in Web search.

Online information services gain revenue from advertisements related to search results and clicks on ads placed on Web pages. Typically, the 80-20 rule applies by which a large portion of search service revenue (80%) is generated by a small percentage of queries (20%) that correspond to most popular topics and Web content. Thus, the main business objective is to gain the market share of on-line queries. That is typically achieved by strategies to increase the users' loyalty though an improved on-line experience. In addition to the user interface enhancements and branding through the browser extensions, online services are expanding the spectrum and increasing the quality of the content they include in their search results. Instead of collecting and indexing only freely available Web data they are partnering with publishers to provide access to recently published premium content. With libraries and archives, they are exploring ways to digitize and bring on-line valuable content from more distant past.

2.1.1 Investment in Content Digitization

Increased competition in the Web search marketplace has caused a flurry of activities in content digitization. In summer 2005, Google reached an agreement with three leading university libraries in the US, university of Stanford, Harvard, and Michigan, with the Public Library in New York City, and the Oxford University Library in the UK to scan and index selected material. The scanning involves creation of two copies, insuring that the material is fully indexed and searchable through the Google book search services (http://books.google.com/googlprint/library.html). For legal considerations, the selection of the material is restricted to the works that are out of the copyright. The digitization is carried out in the dedicated scanning centers established locally.

On October 3, 2005, the Internet Archive, Yahoo! Inc., Adobe Systems Inc., the European Archive, HP Labs, the National Archives (UK), O'Reilly Media Inc., Prelinger Archives, the University of California, and the University of Toronto formed the Open Content Alliance (OCA) (http://opencontentalliance.org), a global consortium focused on providing open access to content while respecting the rights of copyright holders. The main remit of the OCA is to provide infrastructure and services to enable permanent storage and free downloads of material, including cultural, historical and technological digitized print and multimedia content from libraries, archives, and publishers. In October 2006, MSN joined the Open Content Alliance and reached an agreement with the British Library in the UK to digitize 100,000 selected books in partnership with the Open Content Alliance.

2.1.2 Coordination Effort by the European Union

Following the early signs of business initiatives to engage with academic libraries and invest in digitization, on April 28, 2005 six Member States from the EU put forward a request for an organized effort by the EU Commission to harmonize and coordinate national digitization efforts across Europe. On September 30, 2005 the EU Commission responded in favour by releasing a Communication document and a call for online public consultation "i2010: Digital Libraries", inviting feedback on important issues around preservation of the national heritage through digitization [2], [3]. The scope of the challenge faced by the EU member states in preserving the cultural and national heritage of European nations is best illustrated by the statistics quoted in the Communication document [2]:

> *"The total number of books and bound periodicals (volumes) in European libraries (EU 25) was 2,533,893,879 in 2001"*. (Ibid)

The concern about the deteriorating material particularly applies to the audiovisual documents since the analogue formats deteriorate with time and cause a loss of content:

> *"A survey of ten major broadcasting archives found 1 million hours of film, 1.6 million hours of video recordings, and 2 million hours of audio recordings. Total European holdings of broadcast material are probably 50 times larger. Most of the material is original and analogue. 70% of the material is at risk …"* (Survey by the IST Presto project, Oct 2002, http://presto.joanneum.ac.at/index.asp).

The request for consultation referred to specific questions on digitization and online accessibility of digitized content as well as the long term preservation of digital media (Table 1). In March 2006 the EU Commission published a report on the responses received from 225 contributors [4] and on March 27, 2006 formed the High Level Expert group to assist with defining the strategy for the EU Digital Library effort.

Table 1. 'i2010 digital libraries' Questions for online consultation, by the EU Commission. September 30, 2005 [3]

Digitisation and online accessibility

1) *What additional measures could be taken at national and European level to encourage digitisation and online accessibility of material in all European languages?*

2) *What measures could be taken to promote private investments and new business models such as public-private partnerships for digitising and making historical collections accessible?*

3) *What measures of a legislative, technical, organisational or other nature, could facilitate the digitisation and subsequent accessibility of copyrighted material, while respecting the legitimate interests of authors?*

4) *Is the issue of orphan material economically important and relevant in practice? If yes, what technical, organisational and legal mechanisms could be used to facilitate wider use of this material?*

5) *How could public domain material and other material available for general use (voluntary sharing) be made more transparent and widely known in order to facilitate its online availability for subsequent use?*

Preservation of digital content

6) *What priority measures – in particular of an organisational and legal nature-– should be taken at national and European level to optimise the preservation of digital content with the limited resources available?*

7) *Is there a risk that national legal deposit schemes lead to a multiplication of requirements on internationally active companies? Would European legislation help avoiding this?*

8) *How could research contribute to progress on the preservation front? Which axes of work should be addressed in priority by the forthcoming Specific Research Programmes as part of the 7th Framework Programme?*

The consultation responses cover a wide range of suggestions from various stakeholders, reflecting different interests and perspectives on the digitization and preservation issues. For details we refer the reader to the full report. Here we outline the issues that we believe are in the very core of the content digitization challenge.

The shear scope of the digitization effort calls for a long term commitment and systematic approach to the key issues of storage, access, and preservation. We expect that some coordination of the effort can help, such as establishment of reporting and information services that hold information about all the content that has been already digitized, or at least about the content that incurs high digitization cost and thus duplication of effort should be avoided as far as possible. However, we believe that it is most important to establish a rich ecosystem around information services industry that will drive the digitization process in an economically viable and self sustainable manner. Let's reflect on a couple of key issues.

Guiding Principles for Content Selection

Defining a principled way of prioritizing material for digitization is essential. That is rather difficult considering the number and inter-dependencies of factors that need to be taken into consideration. Such are the condition and deterioration rate of physical artefacts, the cost, the relevance of the digitized content, and the legal, social, and technological constraints on the access to the digital content.

The value of information is not absolute; it depends on the context, in particular, its relation to the current events and needs. *Thus, the value of past information is maximized when optimally aggregated with the contemporary information.* Starting with this premise, we realize that prioritizing material solely on how relevant we expect it to be in the future is difficult – it is equivalent to predicting the future itself. A correct inference is feasible, however, for materials tied to recurring events. Second, it implies that the selection strategy should be tied to the *content exploitation models*, in fact, the model that identifies the demand for a particular type of information and adds value through integration with a related archived content.

One example is the aggregation of historical data with educational material. Augmentation of text books with digitized content of related archived documents provides a clear add-on value that can be captured through the supplementary cost of educational material and re-invested into digitization. The key is a clear connection with and full integration with the educational curriculum. Furthermore, at the national level, teaching history, language, geography, and literature is primarily done in the native language and focused on the national aspects of the shared history. Thus, education scenarios are particularly amenable to boosting digitization of materials in the native language.

Preserving Value Distribution through Copyright and Digital Rights Management (DRM)
In order to ensure that publishers and authors can recover the value of the published work, it is absolutely necessary to have two pieces of technology in place: a DRM and a micro-payment technology. Once the information stakeholders can control the revenue, they will be open to providing information online. A successful service will respect legal requirements and ensure that the interests of publishers of copyrighted information are protected and the protection of authors' rights is enforceable.

2.2 Metadata and Search

It is interesting to draw an analogy between online search and catalogue based retrieval in libraries. On-line search engines do not deliver the content of the live pages but rather provide a list of Universal Reference Locators (URLs) pointing to the Web servers that host the content of the result pages. The results are obtained on the basis of content features that the search engine automatically extracts from the crawled pages and hyperlink structure. Although the term metadata is typically defined as 'the structured data about the data', in a broad sense, we can say that the search engine extracts various types of metadata and uses it for indexing and search

Traditional libraries use carefully structured metadata for referencing physical objects and the quality of library catalogues is absolutely essential for accessing archived material. The first step towards 'digitization' of library services involved creation of electronic versions of catalogues and search over bibliographic data and abstracts. Searching for an item results in a bibliographic record with an indicator of the location where the physical item can be found in the library. The British Library currently stores 26 million catalogue entries in their Integrated Library System, comprising of the subject, title, and author information, with another 25 million still to be entered. Back in 1970's, search over library metadata, initiated a flurry of new activities and brought to existence an exciting area of research - online information retrieval.

We also note the emergence of Web archives. Organizations such as the San Francisco based Internet Archive (www.archive.org) collect and store Web data for future reuse. Brewster Kahle, Digital Librarian, Director, and Co-Founder of the Internet Archive has set an ambitious goal:

> *"The Internet Archive is building a digital library of Internet sites and other cultural artefacts in digital form. Like a paper library, we provide free access to researchers, historians, scholars, and the general public."*

Through its simple search facility, Wayback Machine, the Internet Archive provides access to 55 billion pages stored since 1996. The user can type in the web address and browse through the stored pages by date, with an easy access to other pages that have been collected around the same time. Keyword searching is not currently supported.

2.2.1 Evolving Issues

Many benefits of digital archives and libraries stem from the services that aggregate information from distributed repositories. For that reason, they have been investigating frameworks for interoperability, faithfully using metadata standards, and collaborating on various metadata initiatives.

As new types of information services were introduced via the Web, the community tried to comprehend the implications on their metadata creation practices. The paper by Duval et al. [6] provides good insights in the concerns and the breath of issues that have drawn their attention. Through a joint effort of the Dublin Core Metadata Initiative (DCMI) and the Institute for Electrical and Electronics Engineers (IEEE) Learning Object Metadata (LOM) Working Group they derived a set of 'principles and practicalities' for building useful and sustainable metadata systems. Among 'principles' they outlined the metadata modularity, extensibility, refinement, and multilingualism as essential for defining effective metadata systems. Under 'practicalities' they provide advices on practical decisions that one is faced with when implementing metadata schemas for a particular domain.

However, it became clear that the issues around practical use of metadata reached far beyond the specification guidelines. For illustration, we point out important points raised by McClelland et al. [9]. If one decides to merge metadata from a different resource to an existing digital library and finds inconsistencies, such as missing elements in the imported metadata, or incompatible field sizes, under what circumstances can the imported data be altered? What should be the formal mechanism to communicate that alterations have occurred? How should copyright information of the particular object be distinguished from the copyright for the associated metadata? Thus, besides the standard concerns, it is the copyright of the metadata itself that requires attention.

Digitization of the physical artefacts introduces further questions and complexities. If the artefacts, like newspapers, journals, manuscripts, and other publications are scanned and processed using OCR software, the representation of the content may require a range of components, from a simple scanned image and OCR text to multiple scanned images at different resolution levels, with corresponding descriptors of the content layout and content analysis. The quality of data representation and the captured metadata will determine the types of applications and services that can utilize such data.

3 Preserving Digital Media

The immediate use and management of digital information often overshadows the importance of systematic planning that is required for a long term preservation of digital content. Without an appropriate strategy for preservation, today's digital information, unlike paper documents, will not be accessible in 50, 20, even 10 years from now. Ironically, the digital content is in danger of becoming a victim of its own success. Continuous technological advances that facilitate authoring and use of digital content introduce new formats and new storage media. The use of old formats and applications quickly fades away. *Unless we use adequate technologies and best practices to ensure that the past digital content is compatible with new information environments, we will loose access to the material created in the digital era.* This is a challenging issue with serious implications on the collective memory of our civilization. If not addressed, it also undermines the very strategy we chose to preserve information from physical artefact, i.e., the digitization of paper and media documents.

3.1 Problems and Initiatives

According to Bergman [1] and Lyman and Varian [8], the estimated value of digital documents that are produced in the EU and in danger of digital obsolescence is in excess of 3 billion per year. This is a tremendous cost to businesses and governments. Furthermore, from the timely intervention to save the content of the important and visionary Domesday Project [5] lead by the BBC back in the 1986, we know that such an effort can be quite costly. It is important to incorporate plans for long term use of the content early in its lifecycle, ideally right at the authoring time.

Organizations with legal responsibilities for safeguarding digital information, such as national archives and libraries, have been active in educating about the current state of the art on preservation issues and encouraging innovation in building tools and designing procedures. However, meeting the challenges of preservation goes beyond the capabilities of any single institution. For that reasons the EU Commission has committed resources within the Sixth Framework Programme to address issues of access and preservation of cultural and scientific resources. The objective is to promote collaboration among the libraries, archives, and research institutions who can tackle the problem from different perspectives and with complementary skills.

Similarly, in 2000 the US Library of Congress established National Digital Information Infrastructure and Preservation programs. This effort was further strengthen through partnership with the National Science Foundations which in 2004 launched research grant programs to address digital repository models, tools, technologies and processes, and organizational, economic, and policy issues of digital content preservation.

3.2 Industry Involvement

Preservation issues equally concern businesses, public sectors, and individuals. Joint efforts between libraries and industry have resulted in new insights and innovative approaches to addressing the problem of digital content preservation.

For example, in 2000 the National Library of the Netherlands (KB, Koninklijke Bibliotheek) and IBM started building an electronic deposit system, the Digital Information Archiving System (DIAS). . In order to address the problem of durable and large volume storage with long-term preservation requirements, they initiated a Long-term Preservation Study (LTP Study). The study resulted in 6 reports on important aspects, including the content preservation approach based on the concept of the Universal Virtual Computer (UVC) [7]. The method comprises storing the data and a specifically designed program that decodes and provides a logical view of the data. The logical data view can be used in an emulated UVC environment, running on a real machine of the given time.

An alternative approach to data preservation involves conversion of proprietary document formats into a widely adopted standard. In order to ensure long term preservation of Office documents, Microsoft (MS) Corporation produced a specification of the Office Open XML format [10] that defines an XML schema and its semantics for the MS Office applications. It retains high-level information suitable for editing documents or undergoing transformations using XSLT and other XML-based languages or tools. The Open Office XML format is in the process of approval for industry standard.

4 Summary

Digital media has opened new opportunities for creating, publishing, and communicating information. It has connected the contemporary information with valuable archived content from physical artefacts. It provides new ways of dissemination knowledge beyond traditional boundaries and thus has an unprecedented impact on all aspects of our lives. However, it brings with itself a challenge that must be addressed – the volatility of the digital formats and computing environments in which it can be used. Identifying methods and strategies for ensuring the long-term preservation of the digital format is of utmost importance for the survival of data and information created in the digital era.

References

1. Bergman, M.K.: Untapped Assets: The $3 Trillion Value of U.S. Enterprise Documents. BrightPlanet Corporation White Paper, July (2005) http://www.brightplanet.com/pdf/DocumentsValue.pdf
2. Commission of the European Communities: Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. i2010: Digital Libraries, Brussels, 30. Sept. (2005) http://europa.eu.int/information_society/activities/digital_libraries/doc/communication/en_comm_digital_libraries.pdf
3. Commission of the European Communities: Communication Staff Working Document. Annex to the Communication from the Commission. i2010: Digital Libraries. Questions for online consultation. Brussels, 30. Sept. (2005) http://europa.eu.int/information_society/activities/digital_libraries/doc/communication/annex2_en.pdf
4. Commission of the European Communities: Results online consultation 'i2010: digital libraries', Brussels, 2. March (2006) http://europa.eu.int/information_society/activities/digital_libraries/doc/communication/results_of_online_consultation_en.pdf
5. Darlington, J., Finney, A., Pearce, A.: Domesday Redux: The rescue of the BBC Doomsday Project videodiscs. Ariadne, Issue 36, July (2003) http://www.ariadne.ac.uk/issue36/tna/
6. Duval, E., Hodgins, W., Sutton, S., Weibel, S.L.: Metadata Principles and Practicalities. D-Lib Magazine, Vol. 8, Num. 4, April (2002)
7. Lorie, R., van Diessen, R.: UVC: A Universal Computer for Long-Term Preservation of Digital Information. RJ 10338, IBM Almaden Research Center, San Jose, CA (2005)
8. Lyman, P., Varian, H. R.: How Much Information. The Journal of Electronic Publishing. December, 2000 Vol. 6, Issue 2, December (2003) http://www.press.umich.edu/jep/06-02/lyman.html
9. McClelland, M., McArthur, D., Giersch, S., Geisler, G. Challenges for Service Providers when Importing Metadata in Digital Libraries. D-Lib Magazine, Vol. 8, Num. 4, April (2002)
10. Open Office Specification 1.0. Committee Draft 1, 22. March (2004) http://xml.coverpages.org/OpenOfficeSpecificationV10-CD.pdf

Is Semantic Web technology ready for Healthcare?

Chris Wroe

BT Global Services, St Giles House, 1 Drury Lane, London, WC2B 5RS, UK
chris.wroe@bt.com

Abstract. Healthcare IT systems must manipulate semantically rich and highly structured clinical data in a distributed environment. To address this, the healthcare sector has developed standards for medical vocabulary (SNOMED-CT) and message information models (HL7 Version 3) that carry many of the features present in Semantic Web standards such as the Web Ontology Language (OWL). In this paper we examine this correspondence, and specifically present our experience in implementing SNOMED-CT. We go on to describe the fledgling use of Semantic Web technology in BT Global Services health projects and examine the obstacles to adoption of more of the Semantic Web repertoire in healthcare IT solutions.

Introduction

BT has been investigating the use of Semantic Web technology for several years in its Next Generation Web Research Group. The group led by John Davies is the coordinator of the Semantically Enabled Knowledge Technologies (SEKT) project[1], and is a core partner in the Data, Information and Process integration with Semantic Web Services (DIP) project[2]. BT's aim is to translate the use of these technologies in a research setting into benefits for our customers in divisions such as BT Global Services.

BT Global Services provides networked IT solutions for multi-site organizations. Global Services is playing a prominent role in the £6.2 billion UK National Health Service Connecting for Health (NHS CFH) National Programme for IT (NPfIT) and has won contracts worth more than £2.1 billion. One of the key services to be delivered by the National Programme is the NHS Care Records Service (NHS CRS). The NHS CRS will provide "a live, interactive patient record service accessible 24 hours a day, seven days a week, by health professionals whether they work in hospital, primary care or community services" [1]. The core of the NHS CRS is provided by the Spine, which is the name given to the national database of key information about patients' health and care. In addition more detailed patient information will be held at a local level where care is delivered. This will include records of medical conditions, medication, operations, tests, X-rays scans and other results. The scale of the National Programme has led to the local implementation of

[1] http://www.sekt-project.com/
[2] http://dip.semanticweb.org/

services being coordinated by five regional clusters. Local service providers are responsible for supporting more detailed care record information and other services at a local level within a cluster. Different activities such as radiology or hospital pharmacy will often be supported by different healthcare applications. In order to provide a care record system, these local applications will need to exchange messages containing semantically rich clinical information, and in turn summaries of this information will need to be fed to the national Spine database. BT Global Services is the national application service provider of the NHS Care Records Service providing the Spine, and the London local service provider.

In summary NHS CFH envisages a distributed system in which many diverse applications need to interoperate at a semantic level. To provide a cohesive summary of relevant clinical details for a patient, it will be necessary to aggregate information from multiple sources. These are specifically the requirements that Semantic Web technologies are being developed to address [2]. In this paper we look at the approach that is being taken to support interoperability within the National Programme, specifically a common vocabulary (SNOMED-CT), and a common information model for messaging (Health Level 7 – HL7). SNOMED-CT is a leading international medical terminology and within the National Programme, where appropriate, structured clinical information will be entered using medical terms drawn from SNOMED-CT[3]. HL7 provides standards for the definition of messages between clinical applications[4]. The National Programme is making use of the latest version 3 of HL7, which also provides an underlying information model, and specifications of how information represented using SNOMED-CT is to be conveyed within a message.

We will describe the many similarities between SNOMED-CT representation and the World Wide Web Consortium's Web Ontology Langauge (OWL)[5]. Given these similarities we also relay our experience in implementing SNOMED-CT in the hope that it informs projects implementing large scale OWL ontologies. We go on to describe the beginnings of our work that draws upon Semantic Web technology to support SNOMED-CT. Finally, we examine the barriers for more widespread adoption of Semantic Web technology within a care records system.

SNOMED-CT

SNOMED-CT is the Systemized Nomenclature of Medicine - Clinical Terms [6]. It was formed by the merger of a US medical terminology *SNOMED* with the United Kingdom medical terminology *Clinical Terms*. It aims to support the recording of clinical information using a controlled vocabulary that then enables machine interpretation whether simply for information exchange, or for decision support, aggregation and analysis. Its ongoing development is overseen by an editorial board with representatives from the College of American Pathologists and the UK National

[3] http://www.connectingforhealth.nhs.uk/technical/standards/snomed

[4] http://ww.hl7.org

[5] http://www.w3.org/2004/OWL/

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

Health Service. NHS CFH has specified the use of SNOMED-CT in the Care Record Service at both a national and local level. There are several features of note.

SNOMED-CT is large with over one million terms, associated with over 400,000 concepts. SNOMED-CT is much larger than most available OWL ontologies and so poses scalability issues for OWL software tools that are only beginning to be addressed.

SNOMED-CT is concept based, in which a concept can be represented by more than one term. For example the terms ‘pancreatoduodenectomy’ and ‘Whipples procedure’ represent the same medical concept. Also some term strings can represent more than one concept. For example, the term ‘cold’ can refer to a cold sensation concept or a common cold. It is possible to represent SNOMED-CT concepts in the OWL language as OWL classes, and different terms used to denote those classes can be represented using RDF Schema labels if required. There is nothing in SNOMED-CT equivalent to OWL instances.

Each concept is placed in a pure subsumption hierarchy within SNOMED-CT. That is, if a concept has an ‘is-a’ relationship with a more general concept (asthma is-a respiratory disorder), all data annotated with the more specific concept (asthma) will imply an annotation with the more general concept (respiratory disorder). The semantics of the ‘is-a’ relationship are equivalent to that of the OWL subclassOf axiom. For example in OWL abstract syntax:

```
SubClassOf(Asthma Respiratory_disorder)
```

Each concept may also have non taxonomic relationships with other concepts that provide more information about that concept, and may actually fully define a concept. For example, ‘appendicectomy’ has an method relationship with ‘excision’, a procedure site relationship with ‘appendix structure’, and is fully defined. This enables applications to infer that any procedure that includes these two relationships must be an ‘appendicectomy’. The majority of these non taxonomic relationships can be regarded as existential restrictions in an OWL ontology. For example in OWL abstract syntax:

```
Class(Appendicectomy defined intersectionOf(
  Surgical_procedure
  restriction(method someValuesFrom Excision
  restriction(procedure_site someValuesFrom Appendix_structure))
```

SNOMED-CT is underpinned by a description logic (DL) based on Ontylog[3] supplied by Apelon Inc[7]. A description logic reasoner is used to check the consistency of concept definitions and classify concepts in the subsumption hierarchy. In the same way the OWL language is also underpinned by description logic . However the expressivity of the logic differs from that of Ontylog. One of the differences is the use of role grouping in SNOMED-CT [3]. Another difference is the use in SNOMED-CT of an equivalent construct to the property chain inclusion axioms planned to be a feature of OWL 1.1 [4].

SNOMED-CT is extensible at the point of data entry through the use of what is called ‘post coordination’. For example, no pre-existing term exists in SNOMED-CT for ‘left kidney excision’, commonly referred to in medical practice. Instead, the

[7] http://www.apelon.com

terms for ‘kidney excision’ and ‘left’ exist, together with rules that specify how it is appropriate to combine them together. In an OWL ontology these would correspond with anonymous class expressions defined in terms of a number of parent classes and existential restrictions. For example in OWL abstract syntax:

```
intersectionOf(Excision
  restriction(procedure-site someValuesFrom
             intersectionOf(kidney
                       restriction(laterality someValuesFrom left))))
```

It can be seen that the move to more machine interpretable semantics with SNOMED-CT is broadly aligned with the semantics of the OWL. The examples shown in this section have been simplified for illustrative purposes and do not show the steps needed to deal with the different constructs used in the two description logics. However, those involved with the development of SNOMED-CT have made available a script to translate SNOMED-CT into OWL[8].

Health Level 7

Health Level 7 (HL7) is a standards organisation which develops message specifications to enable consistent exchange of information between healthcare applications [5]. NHS CFH have specified the use of HL7 version 3 for the messaging in the National Programme for IT. There are several features of note:

- A common reference information model (RIM) upon which all messages are based.
- For the National Programme for IT, the use of SNOMED-CT to convey the machine interpretable semantics of clinical information.
- An HL7 specific representation for the specification of the information model (not UML or OWL). Version 3 messages are commonly *implemented* using XML.

Complexity arises when clinical information can be structured using either the entities and relationships within the information model of HL7 or the concepts and relationships of SNOMED-CT. A group has been formed to work through issues in the interface between these two standards: TermInfo[9]. The ability of OWL to represent both the conceptual model of SNOMED-CT and the information model of HL7 offers an opportunity to simplify the interaction between these two standards. Rector and Marley have begun to demonstrate the utility of this approach [6].

[8] Much of the development of the mapping between SNOMED-CT and OWL and the subsequent script is the work of Kent Spackman – Scientific Director for SNOMED International.

[9] http://www.hl7.org/Special/committees/terminfo

Experience of implementing SNOMED-CT: a large ontology based terminology

As a local service provider within London, the role of BT Global Services is to integrate a collection of healthcare applications and host them. A core objective is therefore to ensure the consistent use of SNOMED-CT and HL7 by each healthcare application developer, and in some cases provide common services to be used by all applications. A key example is our development of terminology services which provide a common implementation of and access to the SNOMED-CT terminology. The initial focus is on deploying efficient term selection and browsing services so that healthcare applications can provide users with effective means of entering structured clinical information using SNOMED-CT terms. With the increased use of large ontology based terminologies to enter structured data in many domains, we expect the issues relayed here in the context of healthcare applications will be relevant to other areas.

Issues in delivering a large terminology to users

Term search: Experience using previous medical terminologies has shown that a clinician may need to enter 2-15 terms per 10 minute consultation with a patient. Considering they may have over a million terms to choose from and that these are often long, difficult to spell phrases, we must ensure the term selection process is as effective as possible. An application providing a drop down list would always be too long to use easily, but often too short to include the required term from the million available. A search box is the most straightforward solution as exemplified by leading Web search engines, but in this case we are providing search over small phrases rather than complete Web documents, and so common search strategies are not generally applicable. When performing a term search we must aim to ensure that users find all terms relevant to their search (a sensitive search strategy) and only terms relevant to their search (a specific search strategy). Increasing sensitivity decreases specificity and so we therefore have to find a balance between the two. If we search for complete phrases, relevant results are missed because of different word order. For example a search for '`strawberry allergy`' will not find a term but '`allergy to strawberries`' will. If we ask users to enter complete words to search for, it will take too long and be prone to misspelling. For example we can't expect the user to have to type in '`pancreatoduodenectomy`'. If we allow users to enter text that could appear anywhere in a word, the application often returns unexpected results. For example a user searching for '`straw all`' will intend to find '`allergy to strawberry`' but the search service will return the unexpected result '`strawberry gallbladder`'. A search for words in any order *starting with* the search strings has proven the best balance.

Focusing selection and browsing of terms to the context (subsets): SNOMED-CT has over 1 million terms as a result of its goal of being a comprehensive reference medical terminology. That is healthcare applications in many contexts designed for many purposes can all use SNOMED-CT as a common point of reference when using medical vocabulary. However, for any one context only a fraction of the terms are relevant. To address this, SNOMED-CT has a subset mechanism. Subsets are lists of concepts or terms specified as relevant for that specific situation. For example, in an operating theatre system, a subset may be developed specific to surgical procedures. Our search services must support constrained searches within these subsets.

Hierarchies calculated using description logic reasoners (such as those in SNOMED-CT), whilst logically complete are often difficult to navigate by users. SNOMED-CT therefore also has a navigation subset mechanism in which concepts can be grouped by navigation relationships that sit outside the logical definition of those concepts. Our terminology services must therefore support applications in presenting hierarchies that follow these simpler more familiar views on the terminology.

Delivering terminology reasoning at the point of use: As already mentioned, extensibility is a central feature of SNOMED-CT through the use of post coordination (equivalent to OWL anonymous class expressions). Allowing this extensibility at the point of data entry however raises issues at every stage of the lifecycle of clinical data.

Data entry: Applications must provide an effective user interface to allow clinicians to build these expressions. The key is to present only what is sensible to construct in any one clinical context in order to reduce screen clutter with spurious options.

Data storage: Many applications expect to store a fixed length identifier for a term. Expressions can be of arbitrary length.

Data presentation: It must be possible to render the expression back into text that is familiar to clinicians. To prevent overly verbose text, this requires non trivial language generation techniques.

Data analysis: Much data analysis relies on linking specific concepts in individual patient data with more general concepts used to describe decision support rules or statistical categories. Figure 1 illustrates the architecture necessary to support post coordinated expressions. If a clinician enters a novel post coordinated expression, links must be made to the more general concepts referenced elsewhere. Therefore the application must submit these expressions to the terminology service, which in turn uses a description logic reasoner to make the subsumption links. These inferred links can then be made available back to healthcare applications and used in the execution of queries.

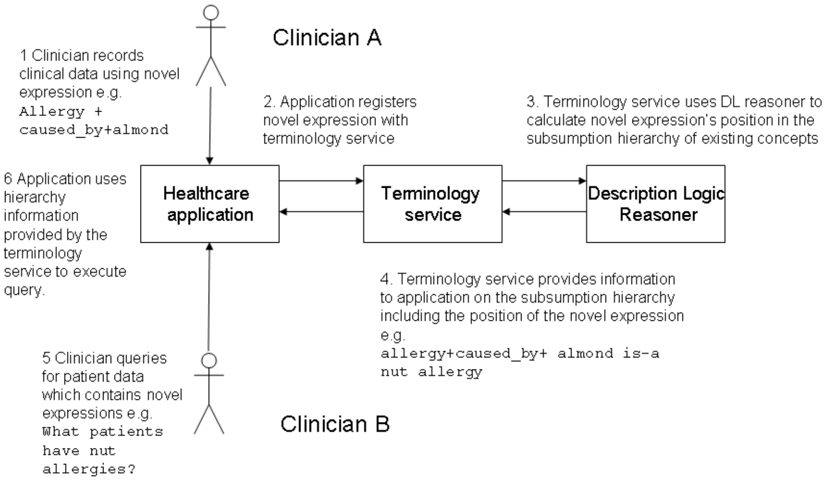

Clinician A
1 Clinician records clinical data using novel expression e.g. Allergy + caused_by+almond
2. Application registers novel expression with terminology service
3. Terminology service uses DL reasoner to calculate novel expression's position in the subsumption hierarchy of existing concepts
6 Application uses hierarchy information provided by the terminology service to execute query.
Healthcare application
Terminology service
Description Logic Reasoner
4. Terminology service provides information to application on the subsumption hierarchy including the position of the novel expression e.g. allergy+caused_by+ almond is-a nut allergy
5 Clinician queries for patient data which contains novel expressions e.g. What patients have nut allergies?
Clinician B

Fig. 1. The interaction between healthcare applications, a terminology service and description logic reasoner necessary to support post coordination.

Experience of using Semantic Web technology to support SNOMED-CT

We are just at the beginning of addressing the challenges of post coordination and are turning to existing Semantic Web technology to do so. The first step has been to investigate the feasibility of using description logic reasoners in the live Care Record System environment. Within BT Global Services we have developed a proof of concept system using the description logic reasoner FaCT++ from the University of Manchester[10]. This system pairs the existing terminology services used to support term selection and navigation, with the DL reasoner. As mentioned previously, SNOMED-CT developers have released a script to translate SNOMED-CT release files to OWL. This has been used as a specification to provide a more direct translation between a SNOMED-CT database and language used to interact with many DL reasoners, DIG[11]. A load process takes the SNOMED-CT release from the terminology server database, translates each statement to DIG and submits these to reasoner. To simulate the live environment, anonymous class expressions are then submitted to the reasoner as part of a subsumption query, for example `intersectionOf(allergy restriction( causative_agent someValuesFrom almond)`. That is the reasoner is asked what concepts subsume this anonymous class expression. This reflects the linking task needed between patient data and general concepts in decision support rules or statistical

[10] http://owl.man.ac.uk/factplusplus/
[11] http://dl.kr.org/dig/interface.html

categories. The result is assessed both for speed of response and validity. In this case we must ensure at least one of the results is `nut allergy`.

Initial qualitative results show that the ontology of preexisting concepts defined in SNOMED-CT can be loaded and reasoned over in an acceptable time (4 hrs on a Sunfire V210 4GB memory). Submission of test expressions as part of a subsumption query returns appropriate results in an acceptable period (<10ms for the example above). Further work is needed to assess response times with expressions of increasing complexity and also benchmark commercial alternatives to FaCT++.

Obstacles to further adoption of Semantic Web technology

Although we have found the use of OWL and associated description logic reasoners to be promising, their adoption in the eventual solution is not certain at this point. Also OWL forms only one part of Semantic Web technology. The following section describes the obstacles to wider adoption.

Lack of harmonisation between Health Informatics and Semantic Web standards: As described earlier although both OWL ontologies and SNOMED-CT are underpinned by description logic, the expressivity of the two logics is slightly different. Work needs to be done to compare the results of the two reasoning processes on the same statements to ensure the conversion from SNOMED-CT to OWL and subsequent use of OWL DL reasoners does not produce different inferences than those used in the original creation of SNOMED-CT. Only when we and our customers are confident this is the case can we use Semantic Web based DL reasoners in the Care Record System.

Obstacles in using Semantic Web technology for data representation: So far in this paper we have concentrated on Semantic Web technology purely to specify the vocabulary used to represent clinical information. With the Resource Description Framework (RDF), the Semantic Web provides a flexible graph based model to represent structured data itself with several advantages over alternative approaches including:

- a standard mechanism for the identification of resources (Universal Resource Identifier)
- a mechanism for the aggregation of data from distributed sources.
- reification which allows statements about statements. This echo's standard patient record architectures in which all entries are statements attributed to a clinical author.
- a link to the well defined semantics of OWL.

Despite this promise, the use of RDF remains at the research level within UK health informatics and is not yet being considered for implementation by suppliers in the NHS. Reasons for this include:

- **Novelty:** exposure to RDF is limited in this community

- **Alternative technology:** A relational data model is used by the majority of health care applications and clinical data warehouses. The ubiquity of this model has ensured that the tools and expertise are available with which to straightforwardly build an application. The same is not yet true for building an RDF based application or data warehouse.
- **Scalability and performance:** The flexibility of RDF comes with the downside of reduced performance. Although examples are appearing of RDF repositories containing millions of RDF statements, more evidence of performance and scalability will be needed to ensure its adoption.

Barriers in using Semantic Web technology for Web Services: At present the number of Web Services available within the National Programme has not necessitated the need for service registries or orchestration of those services. As and when the number of Web Services increases, the need for machine interpretable semantic descriptions of service functionality may then being to appear.

Conclusion

The healthcare IT sector as exemplified by the National Programme for IT appears a viable target for the adoption of Semantic Web technology. Interest in this area internationally can be gauged by the significant activity in the W3C Semantic Web for Health Care and Life Sciences Interest Group (http://www.w3.org/2001/sw/hcls/). Within BT Global Services we are beginning to explore the adoption of Semantic Web technologies specifically around the implementation of medical terminologies. However, the use of Semantic Web technology for the representation of the data, data schema and services remain the focus of research activity.

References

1. NHS Connecting for Health Implementation Guidance Team. National Programme Implementation guide v4.0, Section 3 What is the programme? (March 2006). Available at: http://www.connectingforhealth.nhs.uk/implementation

2. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic Web. Scientific American 284:55 (May 2001) 28-37.

3. Spackman, K.A., Dionne, R., Mays, E., Weis, J:. Role grouping as an extension to the description logic of Ontylog, motivated by concept modeling in SNOMED. Proc AMIA Symp. (2002) 712-6.

4. Patel-Schneider, P.F.: The OWL 1.1 Extension to the W3C OWL Web Ontology Language. Editor's Draft of 19 December 2005. http://www-db.research.bell-labs.com/user/pfps/owl/overview.html/

5. Jones, T. M., Mead, C. N.: The Architecture of Sharing. An HL7 Version 3 framework offers semantically interoperable healthcare information. Healthcare Informatics, (November 2005). Available at: http://www.healthcare-informatics.com/ issues/ 2005/11_05/jones.htm

6. Marley, T., Rector, A. L.: Use of an OWL meta-model to aid message development. Current Perspectives in Healthcare Computing (2006), Conference Proceedings**,** Harrogate, UK, March 2006, In Press.

AKSIO – Active knowledge management in the petroleum industry

David Norheim, UniK and Computas AS, Norway, dn@computas.com

Roar Fjellheim, Computas AS, Norway, raf@computas.no

Abstract. The AKSIO project is developing a process-enabled knowledge management system to support operations of offshore oilfields. The system will provide timely and contextual knowledge for work processes. Experiences will be processed and annotated by experts and linked to various resources and specialist knowledge networks. AKSIO will allow discovery of experiences through the support of a domain ontology. Core functionality of the AKSIO system is provided by careful application of Semantic Web technology, including ontology-based annotation and contextual ontology driven retrieval of content.

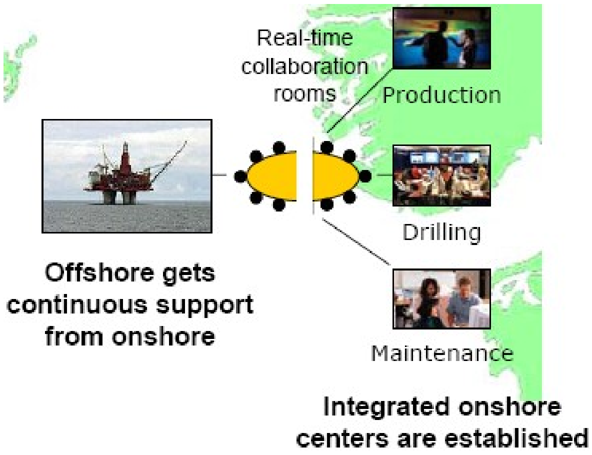
Introduction

As the third largest exporter of crude oil (ca. 3 million barrels/day or 4% of the world's oil production), the oil and gas industry is of major importance to Norway. The petroleum resources are located in the North Sea, a challenging environment for oil and gas production. Oil companies have deployed advanced technology to increase output and reduce cost. Increased output from the fields is possible due to new use of technology and methods e.g. injection of Natural Gas, Water and CO_2 in the oil wells.

To various degrees all major operators and service companies in the North Sea are implementing the concept of Integrated Operations (IO). In IO, the offshore oil platforms are connected by high-speed data links to on-shore control centers, where multidisciplinary teams collaborate to optimize operations and rapidly solve any problems. The ultimate goal of IO is to maximize value created from petroleum resources, which can only be ensured by a continuous stream of right decisions made at the right time.

The largest operator in the Norwegian Continental Shelf, Statoil, has about 20 ongoing drilling projects in the North Sea, Venezuela, and the Middle East. Through massive use of IT for on-line monitoring, analysis, and decision tasks, knowledge transfer within projects are satisfactory, but still sometimes failing between projects, resulting in costly repeated failures. In the drilling process, downtime costs around $0.5 - $1 mill/day, which calls for systems that focus on knowledge-enabled work processes, and use of semantic technologies to facilitate reuse of knowledge from the drilling process. Appropriate and timely knowledge has to be presented to decision-making processes both in planning and subsequent drilling operations.

Figure 1 – Integrated Operations



Active Knowledge Support in Integrated Operations (AKSIO) is an active socio-technical system for knowledge transfer between drilling projects, through documented experiences, best practices, and expert references. Our hypothesis is that an active knowledge system needs to be completely embedded in main work processes and be part of daily work.

The general idea is to provide decision makers with the best available knowledge in a task-relevant, timely, and contextual manner and provide feedback loops to capture new or delete obsolete knowledge.

Use Cases

Experience transfer is a social process, involving knowledge creation and knowledge reuse, where there is a win-win for everyone involved. The system is created around these two use cases.

1) Capture and qualify knowledge gained in drilling operations, and

2) Supply relevant and timely knowledge to planning of new wells

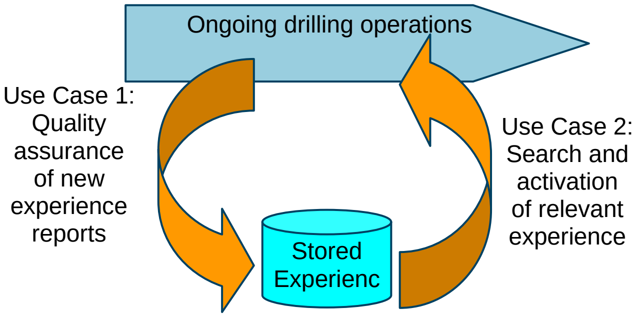

Ongoing drilling operations
Use Case 1: Quality assurance of new experience reports
Stored Experienc
Use Case 2: Search and activation of relevant experience

Figure 2 – AKSIO use cases

Business challenges with current approach

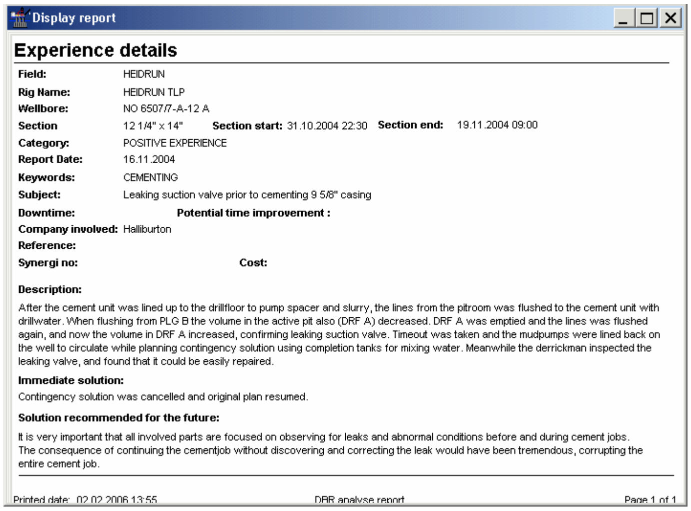
The individual drilling projects are responsible for recording positive and negative experiences encountered during the drilling operations in a database-system called Daily Drilling Report (DBR). These reports are recorded mostly in free text, with minimal metadata. Based on experiences they keep an up-to-date local best practice.

The main challenge with the current approach seems to be that the knowledge processes are not well connected to existing core working processes.

For instance the recording of the experiences shows great variation, and there is no feedback loop to the originator of the experience telling him that the experience was useful or not. “Quality in the first step”, is being used as a slogan to ensure the quality of the original experiences; reality however shows that experiences are quite project specific and some are just random notes.

The annotation process relates the experience to best practices, people and actions. The challenge here is to make this part of the daily work performed by the discipline advisors, and not seen just as “extra work”.

On the consumer side, especially during operation, users rarely search for experiences outside their own projects. Interviewed users claim that they use their own networks to find out what is happening in other projects, and not the existing knowledge base. AKSIO needs to take a non-invasive process driven search where information is pushed to the user in the context he is in. This process driven search will require detailed contextual data about the process, user, and well in question.



Experience details

Field: HEIDRUN
Rig Name: HEIDRUN TLP
Wellbore: NO 6507/7-A-12 A
Section 12 1/4" x 14" **Section start:** 31.10.2004 22:30 **Section end:** 19.11.2004 09:00
Category: POSITIVE EXPERIENCE
Report Date: 16.11.2004
Keywords: CEMENTING
Subject: Leaking suction valve prior to cementing 9 5/8" casing
Downtime: **Potential time improvement :**
Company involved: Halliburton
Reference:
Synergi no: **Cost:**

Description:

After the cement unit was lined up to the drillfloor to pump spacer and slurry, the lines from the pitroom was flushed to the cement unit with drillwater. When flushing from PLG B the volume in the active pit also (DRF A) decreased. DRF A was emptied and the lines was flushed again, and now the volume in DRF A increased, confirming leaking suction valve. Timeout was taken and the mudpumps were lined back on the well to circulate while planning contingency solution using completion tanks for mixing water. Meanwhile the derrickman inspected the leaking valve, and found that it could be easily repaired.

Immediate solution:

Contingency solution was cancelled and original plan resumed.

Solution recommended for the future:

It is very important that all involved parts are focused on observing for leaks and abnormal conditions before and during cement jobs.
The consequence of continuing the cementjob without discovering and correcting the leak would have been tremendous, corrupting the entire cement job.

Printed date: 02.02.2006 13:55 DBR analyse report Page 1 of 1

Figure 3 – Sample experience report

The current DBR Experience system has a simple approach to categorizing experiences. The editor of the experience can select a single pre-existing keyword from a list of terms containing both activities and equipment, and in various details. The approach is seen as inadequate for aiding in retrieval, and calls for a more advanced and complete taxonomy driven approach.

Producer use case

The first use case involves creating a cross-project quality assurance process involving discipline advisors, experts in various technical areas required for drilling projects. Discipline advisors have self-interest in keeping best practices for their discipline, and to keep a network of experts on the given discipline.

In practice the use case involves cleaning the knowledge base of experiences not relevant for cross-project reuse, adding annotations from discipline advisors, classifying the information, and linking to experts, best practices and actions.

Consumer use case

The consumer use case is driven by drilling projects, either in the planning stage or ongoing operations. The objective is to discover relevant experiences that could affect their current or planned operation. This can for instance be that some particular equipment planned to be used has shown failure under certain circumstances, or that a certain procedure could save time and money.

AKSIO provides a search engine utilizing a shared ontology for discovering relevant experiences, and embeds this in the existing work processes. The information will be presented in such a way that it shows the relevance to the project and references to best practices, experts and who made the experience.

Drilling ontologies

The oil industry has for many years worked to establish dictionaries and taxonomies for the industry.

- The Posc Caesar organization [1] promotes the development of openly available specifications to be used as standards for interoperability for the oil sector using ISO 15926, “Integration of life-cycle data for process plants including oil and gas production facilities".
- IIP [2] aims to create an information platform for the industry by integrating ontologies from several industrial data and technical standards and also by adding new ontologies. The project integrates data and information for subsea seismic, equipment, drilling, production, operation and maintenance. IIP includes information from Posc Caesar and currently has some 60.000 classes described in ISO 15926-7, using OWL, most of which concerns equipment.
- There are also other schemas and taxonomies focused on reporting to the government – ongoing work. Because of the nature of reporting, the engineers find these inadequate for annotation of experiences.

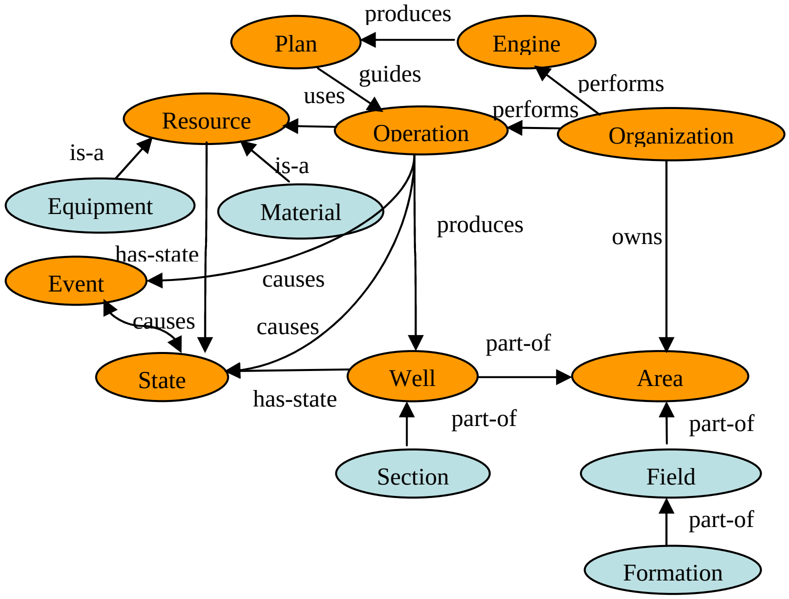


Figure 4 - Top-level terms in the drilling ontology

Though it would be interesting to use the IIP ontology in a stage where it is more complete, AKSIO has a demand for establishing the relationships between various equipment, disciplines, and activities, which is currently not supported by this ontology. AKSIO is to use the ontology to annotate experiences, and to utilize it in retrieval; it seems reasonable to only use a subset of IIP as the ontology may become quite large.

Based on this AKSIO has established its own drilling ontology for this purpose. The ontology is scoped around a few main top-level classes: Equipment, Material, Operation, Plan, Engineering, Organization, Area, Well, State, and Event and relations between these terms.

Process support

Statoil has rolled out Microsoft SharePoint as their portal framework. Each drilling well has a Web site (SharePoint) supported by an underlying document management system that tags all documents with the wells metadata.

The user interface of the system is implemented using Microsoft SharePoint Custom Web parts. This gives access to exchange services like tasks and email and metadata about the well.

Use of semantic technologies

AKSIO leverages the Semantic Web standards RDF and OWL for semantic annotation of experiences and to link information and experts to build a social and intelligent knowledge support for operations.

Structured information from relational databases and LDAP directories creates virtual RDF graphs to be queried by federated SPARQL queries.

Annotations and links to resources (information and people) are made in RDF and stored in a common Knowledge Resource Map using Jena

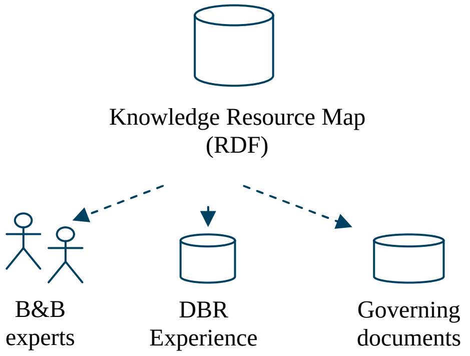


Figure 5 – AKSIO knowledge base

Common metadata vocabularies like Dublin Core, Friend-of-a-friend (FOAF) are used to represent common concepts, a.k.a microformats, mixed with more complex structures. This allows for easy rendering of common features as events, business cards etc. in the user interface.

The drilling ontology shown above is implemented in OWL and serves as the basis for annotation.

A flat folksonomy tagging approach is also included to facilitate creation of new ontology concepts.

The drilling ontology and relations between terms within it serves as a reasoning engine for query expansion during retrieval. Various tactics, such as specialization, part-of etc. are used for ranking the results.Metadata is attached to the individual steps of the process, and in addition to login role-details the metadata is used as contextual data to automatic search execution (a.k.a pull) as well as to narrow the query results.

Results from the engine are presented with links to other resources by additional investigation of the ontology.

Socio-technical challenges

The socio-technical approach to knowledge management taken by AKSIO includes finding approaches to answer among others the following questions:

- Whose job is it to improve the quality of experiences, and how can we make this a more social process including the originating project, discipline advisors and end-users?
- Why is an ontology approach better than using mere keywords? And how can we include the stakeholders in establishing the required ontologies?
- How can we leverage contextual information, and existing metadata initiatives to increase search precision and improve recall?

Lessons Learned

The production use-case was run over a period of a month end of 2005 [3]. Results show that some 60% of the experiences added were marked as suppressed due to either incomplete information or only local relevance. Of the remaining experience reports, the AKSIO process involving specialist annotation resulted in significantly improved explanation, analysis, and recommendation for future practice.

The next pilot aims to show that a process driven approach in where the search is conducted by the context will discover useful experiences otherwise undetected. A drilling ontology will serve as the backbone for the search.

The problem described in this paper could possibly be solved with other technologies. However, we believe that using semantic web technologies will enable improved search as well as scalability for bridging the structured database world with the unstructured world of text and metadata.

Future work

The AKSIO project continues until 2007, and over the next year we will inclusion of other relevant structured and unstructured sources by federating the queries and query rewrite., look more into information extraction tools for annotation, and using Oracle 10g as the backend RDF store

We will also look more into the reuse of domain ontologies, documenting approaches for ontology development and scoping of these, and improved integration into work processes.

In a continuation prospect we will also look into a collaborative decision making process for integrated operations.

Conclusions

The AKSIO project is developing a process supported knowledge management system to support operations of offshore oilfields. The core functionality of the AKSIO system is provided by careful application of Semantic Web technology, including ontology-based annotation and contextual ontology driven retrieval of content.

Preliminary results show that by eliminating noise and precise annotation we have been able to improve the quality of the experiences applicable for cross-project reuse. Subsequent tests will show if we are also able to improve the use of search and the precision and recall for the searches, and thereby the ability to avoid million dollar mistakes in the future.

Acknowledgements

The AKSIO project is carried out by a team consisting of researchers from Norwegian organizations Statoil ASA, Hydro ASA, Computas AS, Institute of Energy Technology, Det Norske Veritas, The Norwegian University of Science and Technology, and UniK (University Graduate Studies at Kjeller). The project is supported by The Norwegian Research Council through grant PETROMAKS 163365 (2004-2007).

References

[1] Posc-Caesar Assosiation, http://www.posccaesar.org/

[2] Integrated Information Platform for Reservoir and Subsea Production Systems Project (IIP), http://www.posccaesar.org/downloads/POSCCaesar_IIP.pdf

[3] Semantic Support for Knowledge Recall in Integrated Operations, R. Fjellheim and D. Norheim, Semantic Technology Conference 2006, San Jose, CA, March 2006

Contextual Intelligence for Mobile Services through Semantic Web Technology

Matthias Wagner, Massimo Paolucci, Marko Luther, Sebastian Boehm
John Hamard, Bertrand Souville

Future Networking Lab
DoCoMo Communications Laboratories Europe GmbH
D-80687 Munich, Germany
<lastname>@docomolab-euro.com

Abstract. The success of future mobile services will largely depend on their ability to maximize their value in varying context. Contextual intelligence in devices, mobile applications and service platforms is needed to manage different mobile terminals, personalize content and services or to narrow down possibly very large sets of applicable services in a given situation. With this vision of Contextual Intelligence in mind, we are exploiting the use of OWL and related technologies with a particular focus on providing access to the Semantic Web for mobile applications. The aim of this paper is to provide a brief and concise overview of our respective project activities at DoCoMo Euro-Labs.

1. Introduction

For future mobile services and applications to succeed, a substantial amount of contextual intelligence in devices, mobile applications and service platforms will be required. By “contextual intelligence” we mean added capabilities of mobile services and applications that allow flexibly adapting to and maximizing their value in varying contexts. To this end, we are exploiting Semantic Web technology as an enabler for this added intelligence in context-aware mobile services. Our vision is to overlay the Semantic Web on ubiquitous computing environments making it possible to represent and interlink content and services as well as users, devices, their capabilities and the functionality they offer.

Our research in the Evolutionary Systems Project at DoCoMo Euro-Labs focuses on user guidance and contextual intelligence in living systems to keep up with a continuously changing environment. We identified flexible services and service platforms that adapt in response to changes in their context utilizing knowledge-based augmentation as key targets. Here, knowledge representation techniques such as ontologies are of great importance. Services will be modeled and tailored to preferences and needs of individual customers to better meet their requirements, matching the capabilities of devices and platforms available. While changes to our environment happen continuously and implicitly, our technology has to be kept current and in sync explicitly. Even more effort is required to meet steadily increasing expectations of customers of mobile communication systems. We believe that semantically well-founded personalization concepts with a universal serviceability to be an important step into this direction. The aim of this paper is to provide a brief and concise overview of the group's related project activities at DoCoMo Euro-Labs.

2. Project Overview

Figure 1 sketches the range of projects related to the Semantic Web that we currently pursue in our laboratories. On a research map, the ongoing activities can be aligned according to their level of concern with representation fundamentals, in our case mostly through OWL, as well as reasoning support.

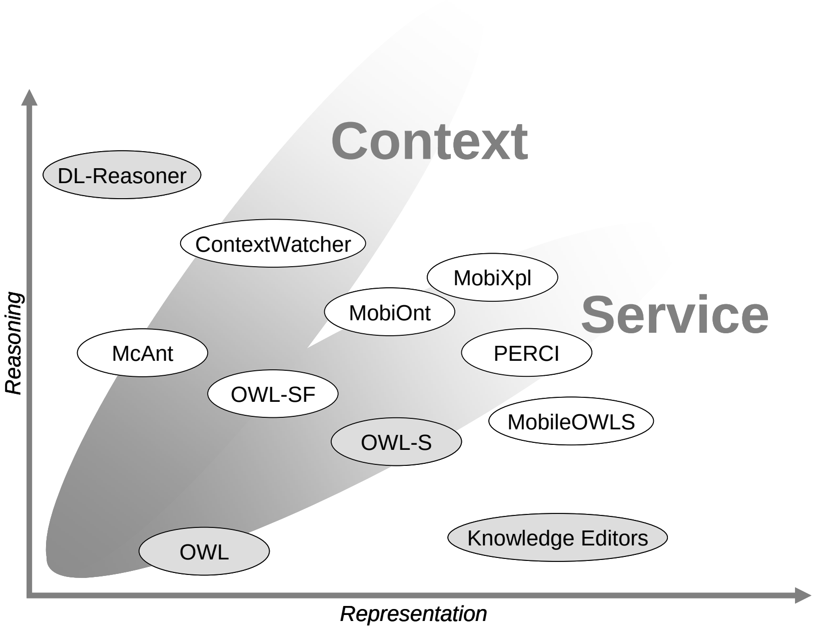


Figure 1: Project activities exploiting Semantic Web technology.

The projects are also projected to an additional application dimension representing activities that target at semantic-based mobile applications in terms of support for context representation and management, support for semantic mobile services or both of these aspects. The single activities are described in further details in the following.

2.1 OWL-SF

Major challenges in designing ubiquitous context-aware systems include the distributed nature of context information and the heterogeneity of devices that provide services and deliver context. We have approached these challenges within the project OWL-SF [1], a distributed semantic-based service framework. In the OWL-SF prototype, OWL is used to capture high-level context elements in semantically well founded ways. Devices, sensors and other environmental entities are encapsulated and connected to the upper context ontology using OMG's Super Distributed Objects technology and communicate using the Representational State Transfer model. An early use case of OWL-SF studies enhanced presence control to realize intelligent call forwarding [2].

2.2 ContextWatcher

Within the IST project MobiLife[1] [3] we have implemented ContextWatcher, an early prototype for semantic-based monitoring of mobile users. The project aims at frameworks for context-aware services that support users in their daily life. OWL upper context ontologies define the basic contextual categories and the relations among them. Such high-level structuring of context information enables its integration and consolidation on a semantic basis. Furthermore, the axiomatic descriptions of context elements such as personal situations (i.e., *Working*, *At_home*, etc.) can directly be used by logical inference engines to realize reasoning about the user's presence and virtual location [4]. Context Watcher acts as a client within the MobiLife Context Management Framework which defines a general framework for gathering contextual information and is itself based on Web service technology.

2.3 McAnt

With our prototype McAnt we try to explore possibilities and core technologies towards leveraging the Semantic Web for desktop application enhancements [5]. The idea is to deploy qualitative reasoning on the user's personal desktop environment to enable a more refined support for personal information organization. As in our related activities, OWL and DL-based reasoning are explored as enabling technologies for semantic enrichment. McAnt currently extends Apple's desktop environment in terms of the Apple Address Book and the iCal calendar tool. The project introduces a set of core ontologies that describe novel OWL-based smart groups that build on Apple's smart groups and folders.

2.4 MobiONT and MobiXPL

The discovery of adequate services will become a more and more demanding problem especially for the mobile user who has to cope with changing context and limitations of mobile terminals. We have implemented MobiOnt and MobiXpl – a semantic matchmaker for service discovery and a personal mobile client – to explore mobile

[1] MobiLife is an integrated project within the 6th Framework Program of the European Commission, Project-No. IST-2004-511607(IP).

user-centered services on the Semantic Web [6]. Our vision is to take full advantage of future complex service offerings on limited client devices and to handle the need for personalized service discovery in mobile environments. Main contributions are in support for browsing service ontologies, the cooperative discovery of services as well as an intuitive preference model that can be easily managed on restricted clients [7].

2.5 MobiOWLS

MobiOWLS is a new project in which we attempt to extend the OWL-S [8] upper ontology for services to describe services for mobile and ubiquitous computing. Our initial investigation concentrated on extending the OWL-S Profile to include crucial quality-of-service information and contextual information that in our experience is required to locate the best service that satisfies the needs of the user. For example, we extended OWL-S with properties such as *Media* that specifies the type of media, such as video vs text, that is used to deliver the service, or the *CostModel* of the service, such as flat rate or a fee-per-use.

2.6 PERCI

The interaction between mobile devices and physical objects is gaining more and more attention since it is an intuitive way to request services from real world objects. We currently see several solutions for the provision of such services, most of these are proprietary, designed for a special application area or interaction technique and provide no generic concept for the description of services requested from real world objects. On the other hand ubiquitous environments with many tagged and networked objects – which are also often referred to as the *Internet of Things* – could provide generic standard methods for physical tagging and interfacing. In our project PERCI (Pervasive Service Interaction) we aim at leveraging Semantic Web technology to enrich and orchestrate such standard tagging and interaction methods. In particular we study how Semantic Web service frameworks such as OWL-S could be extended towards ubiquitous service interaction.

3. Concluding Remarks

In our research – with the vision of contextual intelligence for mobile and ubiquitous service environments in mind – we are exploiting the usage of Semantic Web Technology within several practical projects that we briefly introduced above. These projects are either concerned with fundamental support for OWL-based development and/or with ontology-based services and applications in the mobile computing arena.

In this paper we presented our activities to date, and we highlighted some of the challenges that we are facing right now. We have gained considerable experience on handling contextual information and on the infrastructure that is required. Yet, despite the progress that has been made, we feel that there are still impressive challenges to tackle. Within the context of our projects we have also identified limitations and problematic issues in using Semantic Web Technology. In particular, in exploiting OWL we found limitations in the language specification itself as well as in the tool support that we reported elsewhere [10].

References

1. Mrohs, B., Luther, M., Vaidya, R., Wagner, M., Steglich, S., Kellerer, W., Arbanowski, S.: OWL-SF – a distributed semantic service framework. In: Proc. of the Workshop on Context Awareness for Proactive Systems (CAPS'05), Helsinki (2005) 67–77
2. Luther, M., Mrohs, B., Vaidya, R., Wagner, M.: OWL-SF – distributed owl-based reasoning on objects in the real world. In: Proc. of ISWC'05 (Demo Track), Galway (2005)
3. MobiLife: Project homepage. http://www.ist-mobilife.org (2005)
4. Luther, M., Böhm, S., Wagner, M., Koolwaaij, J.: Enhanced presence tracking for mobile applications. In: Proc. of ISWC'05 (Demo Track), Galway (2005)
5. Böhm, S., Luther, M., Wagner, M.: Smarter groups – reasoning on qualitative information from your desktop. In: Proc. of the 1st Workshop on The Semantic Desktop at ISWC'05, Galway (2005)
6. Wagner, M., Liebig, T., Noppens, O., Balzer, S., Kellerer, W.: Towards Semantic-based Service Discovery on Tiny Mobile Devices. In: Proc. of the Workshop on Semantic Web Technology for Mobile and Ubiquitous Applications at ISWC'04, Hiroshima (2004) 90–101
7. Wagner, M., Noppens, O., Liebig, T., Luther, M., Paolucci, M.: Semantic-based Service Discovery on mobile Devices. In: Proc. of ISWC'05 (Demo Track), Galway (2005)
8. Martin, D., Burstein, M., Hobbs, J., Lassila, O., McDermott, D., McIlraith, S., Narayanan, S., Paolucci, M., Parsia, B., Payne, T., Sirin, E., Srinivasan, N., Sycara, K.: OWL-S: Semantic Markup for Web Services. Member Submission, W3C (2004)
9. Siorpaes, S., Broll, G, Paolucci, M., Rukzio, E., Hamard, J., Wagner, M., Schmidt, A.: Mobile Interaction with the Internet of Things. In: Proc. of Pervasive'06 (Special Track on Late-breaking Results), Dublin, Ireland, May (2006)
10. Liebig, T., Luther, M., Noppens, O., Paolucci, M., Wagner, M., von Henke, F.W.: Building Applications and Tools for OWL – Experiences and Suggestions. In: Proc. of the Workshop on OWL: Experiences and Directions. In conj. with RuleML'05, Galway, Ireland, November (2005)

Semantic Laboratory Notebook: Managing Biomedical Research Notes and Experimental Data

Alexander Polonsky[1], Adrien Six[2,3], Mikhail Kotelnikov[1], Vadim Polonsky[1], Renaud Polly[1], Paul Brey[2]

[1] Cognium Systems SA, 15 rue Commines,
75003 Paris, France
{apolonsk, mikhail.kotelnikov, renaudpolly, vpolonsk}@cogniumsystems.com
[2] Institut Pasteur, 25-28 rue du Dr Roux,
75015 Paris, France
{askc, pbrey}@pasteur.fr
[3] Université Pierre et Marie Curie - Paris 6, 4 place Jussieu,
75005 Paris, France

Abstract. The main raw product of biomedical research is the information contained in laboratory notebooks and the associated computer files of individual researchers. Most of the problems in managing bioresearch information downstream stem from the way this information is initially recorded and stored. Electronic notebooks based on traditional knowledge management approaches have not been widely adopted by bio-researchers – the vast majority still use paper notebooks. We describe deployment of a software system based on the semantic tagging approach that successfully addresses the key adoption problems. This case study also indicates fruitful directions for the future R&D.

Keywords: Semantic Annotation, Semantic Tagging, Knowledge Articulation, Life Sciences.

1 Introduction

A recent article in Financial Times stated that "*R&D productivity* - not R&D investment - is the real challenge for global innovation" [1]. This is especially true for biomedical research, one of the largest global R&D sectors. Biomedical research is highly information intensive, and much of its information management aspects are inefficient due to a low degree of automation.

According to a study by Atrium Research, research chemists spend on average 2/3 of their time on information-intensive tasks such as meetings, literature analysis, writing papers and reports, and less than ¼ on conducting experiments [2]. Biomedical research generates at least as much information as chemical research, and hence we can expect that the work time distribution for biomedical researchers is skewed to at least the same degree. Indeed, getting the right information is critical for every step of biomedical research, from project planning to reporting the results.

Hence, automating its knowledge management aspects can result in a substantial productivity boost.

We have analyzed 40 articles from the relevant academic studies, analyst reports, and professional press to collect and prioritize the various knowledge management needs in biomedical research. We found that the adoption of structured Electronic Laboratory Notebook (ELN) systems is the key bottleneck to adequately addressing these needs. However, the ELNs based on the traditional knowledge management approaches force people to make a choice between 1) flexible but unstructured data entry or 2) rigid but useful organization. This is one of the main reasons why these systems have not been widely adopted by researchers [3]. As a result, a large amount of information discovered during research gets lost over time and is hard to retrieve, understand, analyze, and manipulate [4-7]. We have used a semantic tagging approach to develop a collaborative laboratory notebook software that allows both sufficient flexibility of data entry and thorough organization of the recorded information.

2 Knowledge Management Needs in Biomedical Research

We have randomly selected and analyzed 40 vendor-independent articles (including analyst reports, academic studies, and professional press) on the subject of knowledge management needs in biomedical research. The table below summarizes the results of the analysis. The needs in the right column were grouped into top-level categories located in the left column. The numbers in parentheses represent the number of articles that mentioned the corresponding needs.

Table 1. Literature analysis of KM needs in biomedical research.

Top-level needs with # of citing articles	Component needs with # of citing articles
Collective data management (29)	• search quality (recall and precision, multimedia) (11) • sharing experience, methods, data, analysis, resources (9) • simple and flexible data entry (5) • free access to scientific information (4) • partial (controlled) sharing (3) • information clarity (3) • real-time, persistent availability of information (2) • useful perspectives on information (2) • sharing knowledge organization methods (1)
Data storage (22)	• electronic, as opposed to paper (9) • storing all experimental data (including method details, full experimental history, negative results, "uninteresting" results, replications, unfinished projects) (8) • intellectual property protection (traceability, security) (8) • open and standard formats (5) • support for large data quantities and multimedia (3) • long-term archiving (2)

Data integration (21), across:	• databases and publication archives (8) • disciplines (biomedical research, chemistry, high-throughput screening, drug development, clinical/patient evaluations) (6) • individuals and groups within an organization (different departments, globally distributed sites) (4) • applications and websites (4) • organizations (subcontractors, partners) (3) • subfields (e.g., brain mapping, genomics, transcriptomics, proteomics, metabolomics) (3) • personal information (research notes, data files, emails) (1) • access rights levels (private, group, corporate, public) (1) • domain concepts (pharmaceutical compounds) (1) • business processes (1)
Personal data management (18) (a subset of Collective data management)	• search quality (recall and precision, multimedia) (11) • simple and flexible data entry (5) • information clarity (3) • useful perspectives on information (2)
Project management (18)	• quality of process and innovation (e.g., quality assurance, experimental design) (5) • work evaluation (work/contribution-based as opposed to publication-based, accountability) (4) • task management (efficient coordination, planning, and reliable implementation of a preset sequence of hierarchical tasks, e.g., protocol implementation) (2) • keeping up to date on a project (1) • resource sharing (cost, time, expertise) (1)
Automatic Analysis (12)	• inference rules, validation (compliance, safety checks, and other validity checks) (6) • consistency analysis (results, methods) (3) • decision support (2) • quantitative analysis (2) • interdisciplinary concept mapping (1) • discovery (e.g., new inter-object relationships) (1) • statistical bias analysis (1) • hints, autocompletion (1)
Communication (7)	• clarity (e.g., format consistency) (3) • automatic report and publication-draft generation (2) • multi-channel publishing (2) • open (review-independent) communication channels (1)

The above summary represents the needs as *perceived* by the domain analysts and the bioresearch community. The citation frequency indicates the degree to which a particular need is perceived, thereby giving a rough sense of the need's priority. However, most of the needs in the table are in fact inter-dependent. For example, better integration would lead to improved search which would in turn lead to improved project management and can indirectly improve data integration.

In order to derive the core user requirements in the domain, we propose to classify the expressed needs into 4 requirements categories: 1) constraints: properties that must be present in a software solution; 2) simplicity or ease of use; 3) direct benefits from individual and collective use of the system; and 4) desired side-effect benefits. We can redistribute the needs from Table 1 according to the 4 categories as follows:

1) intellectual property protection;
2) simple and flexible data entry;
3) all the remaining needs from Table 1;
4) free access to scientific information.

The needs in the 3rd category would be best addressed via manipulations of structured data [8]. The classic approach is for the information to be entered into structured forms or templates thereby becoming much clearer to humans and more accessible to computer-aided operations, such as structure-based search, integration, and analysis. Although this approach has worked very well for certain kind of data (structured data), it has not worked well for all data (unstructured data). Indeed, a lot of information entered in a document format is difficult to input into a form. The same is true for information represented as a network, image or sound. Hence, the traditional approach of structured forms creates a conflict between the requirements categories 2 and 3 above.

This is the main reason why the vast majority of information in a bioresearch organization remains unstructured [9]. As we discuss in the next section, the requirements categories 1 and 4 also depend on the degree of information structure, and therefore, are also in a conflict with the 2nd category. Yet, due to the complex and unpredictable nature of research information, the category 2 is key for a successful adoption of an IT solution by researchers [3]. As we show below, a semantic approach can substantially diminish the conflict between these critical requirements.

2.1 The Key Role of Electronic Laboratory Notebooks

The vast majority of biomedical researchers store the raw information obtained in the course of their experimental work in paper laboratory notebooks and private computer files. Only a small fraction of this information remains during the transformation into scientific reports or publications, leading to a large information loss [4,5]. Hence, an ELN plays two key roles: 1) as the first point of information entry, and 2) as a comprehensive repository of research information, where both research notes and associated electronic files can be stored.

The 1st role is important since it is simpler and more efficient to organize information at the time of its entry than afterwards. A semantic structure created at the stage of note-taking can be propagated to the subsequent stages of processing, such as reports or publications.

3 iPad: Semantic Laboratory Notebook for Biomedical Research

The most natural and straightforward way to represent research notes is using the document representation (that is how they are currently recorded). Imbedded in the note documents can be other data formats such images, video, tables, forms, and network data. Hence, what is needed ideally is an integrated environment for structuring and working with structured information that would address the specific needs of all these numerous data types and would allow to view the same information in different representations (e.g., in a tabular format or in a document format). The ELN system we have developed so far, named iPad, uses the semantic tagging approach to allow people to easily structure and work with structured documents.

3.1 Semantic Tagging Approach

Electronic forms have carried over into the electronic environment many constraints associated with their predecessors, the paper-based forms. For example, it is not possible to enter information in between form fields, copy, delete, or move several form fields at a time, add a form field inside another form field, etc. As a result, electronic forms are often not well-suited for structuring complex hierarchical information such as research notes. However, the constraints of the forms are purely historical, they are not required to give information a semantic structure.

The semantic tagging approach is based on the inverse paradigm: as opposed to forcing information into a given form structure, information can be recorded in a free document format and then labeled with the corresponding semantic tags. This approach allows to combine flexibility with structural organization during document authoring.

3.2 iPad Overview

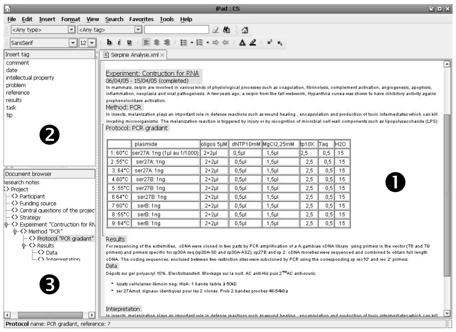
The system is based on a three-tier software architecture comprised of the client applications (standalone iPad Editor and iPad Web Portal), iPad middle-layer Server, and a database. The information is entered via iPad Editor (Fig. 1) and can be viewed either in the Editor or the Portal. It can be stored either in the database or on the users' computers (although, in the latter case, the functionality is quite limited to encourage central storage). The middle layer mediates the transfer of information between the client applications and the database. Different middle-layer adapters allow connecting the system to any database, although currently only the relational database adapter has been developed for connecting to any major relational database (the default is MySQL). The Editor and the Server are implemented in Java (JDK 1.4.2).

Multimedia information (formatted text, tables, images, etc) is entered into the Document Editor (Fig. 1) using traditional document editing functionalities. External files can be attached to the document by drag-and-drop and are displayed as hyperlinks.

The Tag List proposes relevant semantic tags (e.g., project, result, method) depending on the cursor's position within the document. At this time, the proposed tags model the organizational concepts of biomedical research projects and not the discovered biological knowledge (a much more complex problem). For example, the tags are used to clearly mark which method was used to obtain a given result as a part of what experiment or project. iPad also offers a free tag mode in which users can define their own tags. However, these modes are kept separately in order to avoid confusion between the preset tags proposed by the system and those created by users. The free tag mode can be used for ad-hoc organizational needs that were not taken into account by the preset tags.

Only semantically valid tags are proposed and they can be inserted at any place in a document where they are valid. Each tag has a set of attributes that can be filled out in a pop-up form. The possible tags and their attributes are specified in external XML Schema documents and are visualized within the document as specified in external XML-based templates.

The subsequent document structure appears in the Document Browser window. The structure is a hierarchy of inserted tags and can be used to browse the document by clicking on the tag of interest and viewing the corresponding document section in the Document Editor. In addition to the hierarchical relationships, tags (from the same or different documents) can be interlinked with hyperlinks.



Experiment: Contruction for RNA
06/04/05 - 15/04/05 (completed)
In mammals, serpin are involved in various kinds of physiological processes such as coagulation, fibrinolysis, complement activation, angiogenesis, apoptosis, inflammation, neoplasia and viral pathogenesis. A few years ago, a serpin from the fall webworm, Hypanthria cunea was shown to have inhibitory activity agains prophenoloxidase activation.
Method: PCR
In insects, melanization plays an important role in defense reactions such as wound healing , encapsulation and production of toxic intermediates which can kill invading microorganisms. The melanization reaction is triggered by injury or by recognition of microbial cell wall components such as lipopolysaccharide (LPS)
Protocol: PCR gradiant

	plasmide	oligos 5µM	dNTP10mM	MgCl2,25mM	tp10X	Taq	H2O
1: 60°C	ser27A: 1ng (1µl au 1/1000)	2+2µl	0,5µl	1,5µl	2,5	0,5	15
2 :55°C	ser27A: 1ng	2+2µl	0,5µl	1,5µl	2,5	0,5	15
3: 64°C	ser27A: 1ng	2+2µl	0,5µl	1,5µl	2,5	0,5	15
4 60°C	ser27B: 1ng	2+2µl	0,5µl	1,5µl	2,5	0,5	15
5 :55°C	ser27B: 1ng	2+2µl	0,5µl	1,5µl	2,5	0,5	15
6 64°C	ser27B: 1ng	2+2µl	0,5µl	1,5µl	2,5	0,5	15
7 60°C	serB: 1ng	2+2µl	0,5µl	1,5µl	2,5	0,5	15
8 :55°C	serB: 1ng	2+2µl	0,5µl	1,5µl	2,5	0,5	15
9: 64°C	serB: 1ng	2+2µl	0,5µl	1,5µl	2,5	0,5	15

Results
For sequencing of the extremities, cDNA were cloned in two parts by PCR amplification of a A.gambiae cDNA library using primers in the vector (T6 and T9 primers) and primers specific for sp30A seq (sp30A-S0 and sp30A-AS2), sp27B and sp 2. cDNA moieties were sequenced and combined to obtain full length cDNA. The coding sequences, enclosed between two restriction sites were subcloned by PCR using the corresponding sp rec10' and rec 2' primers.
Data
Dépots sur gel polyacryl 10%. Electrotransfert. Blockage sur la nuit. AC anti-His puis 2^{me}AC anti-souris.

- lysats cellulaires:-témoin neg: HisA: 1 bande faible à 50kD
- ser 27Amut: signaux identiques pour les 2 clones. Prob 2 bandes proches 48-54kDa

Interpretation
In insects, melanization plays an important role in defense reactions such as wound healing , encapsulation and production of toxic intermediates which can kill

Protocol name: PCR gradiant, reference: ?

Fig. 1. iPad Editor: (1) Document Editor, (2) Tag List, (3) Document Browser.

Once the information is entered in a structured way, iPad gives the user a large set of functionalities to benefit from the resultant structure, including structure-based browsing and search (with user-friendly interface), information perspectives (also called "semantic lenses" [10]), automatic report and publication draft generation (using mapping between XML Schemas), and ability to visualize the information in a variety of ways on iPad Web Portal (using XSLT).

It turns out that the ability to structure documents also addresses the issue of intellectual property protection (requirements category 1). Parts of documents that contain sensitive information can be specially tagged and processed appropriately (e.g., printed and signed). This substantially decreases the amount of work since only a small part of the electronic information has to be processed in this way.

In addition to the valuable functionality, formal and guided document structure improves information clarity. This has a side-effect benefit: research notes could eventually be shared freely on the Web since they would be sufficiently structured to be understood by other scientists and to retrieved using structure-based search [8] (requirements category 4).

4 Case Study at Institut Pasteur

iPad has been developed in collaboration with Institut Pasteur (Paris, France), a world-renowned biomedical research Institute. It has been used for over a year by a research group of 4 people as well as 3 individual researchers at the Institute. The number of users is gradually expanding.

Although we have not yet conducted a formal evaluation, user feedback has been positive and confirms our assumptions. Users have noted the improvements in 1) information clarity (both within their own and others' notes), 2) research quality due to useful perspectives on their work, 3) report and publication writing, 4) information retrieval, 5) information sharing, and 6) integration.

We have also confirmed our view that the semantic tagging paradigm is not straightforward for users from the beginning and requires a tutorial. The User Interface is the key factor determining the learning curve. Despite being new, the tagging paradigm as implemented in iPad becomes sufficiently intuitive after a couple hours of training and practice.

The system has been used by several individual researchers independently from a research group, showing that it provides benefits that are independent from collective utilization. This is important for its adoption since it avoids the prisoner's dilemma issues commonly present in collaborative systems.

5 Future Work

Directions for future development are numerous:

- Migration from the read-only Web Portal to a Web-based editing environment (a Semantic Wiki) to improve information availability
- Integration of an environment for structuring and working with network information (i.e., biological processes)
- Migration from XML to RDF in order to better support semantic relationships
- Adoption of the peer-to-peer paradigm to increase collaborative flexibility
- Integration with linguistic algorithms for semi-automatic tagging
- Adoption of falksonomy techniques for constructing dynamic collaborative ontologies of biomedical concepts and using this ontology to achieve a greater degree of information structure.

In addition to the technical development, we plan to complete a formal evaluation of iPad in both academic and industrial research settings.

Although iPad's functionality has been focused on the needs of biomedical scientists, due to the generality of iPad's approach and architecture, it can add value in domains other than biomedical research. For example, iPad has already been used to structure and manage generic (non-research) project notes. More work needs be done to evaluate the scope of iPad's applicability.

Acknowledgements. This work was in part conducted within NEPOMUK project supported by IST FP6 grant from the European Community.

References

1. Schrage, M.: For innovation success, do not follow where the money goes. Financial Times (2005)
2. Michael, E.H.: It's Not About the Paper. Scientific Computing & Instrumentation (2004)
3. Michael, E.H.: Electronic Study Management. Scientific Computing (2006)
4. Butler, D.: A new leaf. Nature Vol 436 (2005)
5. Knight, J.: Null and void. Nature Vol 422 (2003)
6. Phillips, M.L.: Do you need an electronic lab notebook. The Scientist (2006)
7. Sarini, M., Blanzieri, E., Giorgini, P., Moser, C.: From actions to suggestions: supporting the work of biologists through laboratory notebooks. Proceedings of the 6th International Conference on the Design of Cooperative Systems (2004)
8. Berners-Lee, T., Hendler, J.: Scientific publishing on the 'semantic web', Nature Web Debates (2001)
9. Building Blocks of an Enterprise Content Management Business Case for Life Sciences. First Consulting Group (2004)
10. Neumann, E.: A Life Science Semantic Web: Are We There Yet? Science Vol 2005, Issue 283 (2005)

Semantic Business Automation[⋆]

Jens Lemcke and Christian Drumm

SAP Research Center CEC Karlsruhe
SAP AG
firstname.lastname@sap.com

Abstract. In this paper, we aim to investigate how semantic Web services can improve standard business process management tools. Based on a standard SAP process in the area of logistics, we show how current approaches support business flexibility via manual modeling tools. Our application of semantic Web service technologies on top of today's business process management tools enables the automation of major tasks of business process management.

1 Introduction

The main purpose and central challenge of *business process management* (BPM) in today's companies is to help keeping track with increasingly dynamic markets. This changing business environment demands for more and more flexibility of the companies to adapt their own business to changing market requirements and to improve interoperability with potential business partners.

In this paper, we show how *semantic Web service* (SWS) technology can improve standard business process management software. We do this on the basis of the real-life order-to-cash business process in the logistics domain between the two business partners shipper and carrier. Upon a purchase request of a customer, the shipper procures the requested good from its depot, packs and labels it, hands it over to the carrier and possibly provides after-sales services such as package tracking to its customer.

Most of these process steps require heavy interactions between the shipper and the carrier. In order to compete on a quickly changing logistics market, it is essential to a shipper that it can easily switch between different carriers which steadily adopt their conditions and service offers over time. In Sect. 2, we sketch how current business process management software on top of a *service-oriented architecture* (SOA) facilitates this business flexibility. This is mainly achieved by providing manual modeling tools. Section 3 then details how modern semantic Web service technologies can be applied on top of existing business process management software. We show how major parts of business process management tasks can be automated using this technology.

[⋆] This material is based upon work partially supported by the EU funding under the projects DIP (FP6 - 507483) and ATHENA (FP6 - 507312). This paper reflects the author's views and the Communities are not liable for any use that may be made of the information contained therein.

2 BPM-Based Implementation

In this section, we will provide a high-level description of how current BPM tools are used to implement the order-to-cash scenario involving a shipper and a carrier as lined out in Sect. 1. In general, a BPM-based implementation of a business process involves two main components: i) a process modeling tool, and ii) a process execution engine capable of executing the modeled processes. In our case, the process modeling tool is called *Maestro* [1], and the process execution engine is *Nehemiah* (see Fig. 1).

Using the Maestro tool, a domain expert would create a graphical representation of the process executed at the shipper side. Note that this graphical representations is not yet linked to any of the shipper's business systems or any services of a carrier. Therefore in a second step, the domain expert manually connects the single process steps of the business process to services offered by either the internal or the partner's business systems. Connecting different services and systems usually requires a mapping between different message formats. Consequently, the domain expert also needs to create the necessary mappings converting between the input and output messages of the different business systems involved.

After the business process has been modeled manually and the involved business systems have been connected to the different process steps, the business process is stored into the process repository. During run-time, the process execution engine retrieves a process from this repository and executes an instance of it upon an incoming request.

The main advantage of BPM-based implementations is the design-time flexibility. After the business process has been modeled using the graphical editor, services implementing different process steps can easily be exchanged during design-time. For example, integrating a new business partner into the collaboration only requires to adopt the connections of the partner services to the appropriate process steps as well as the development of the necessary message transformations. However, BPM-based implementations do not provide any additional flexibility during run-time, as the process execution engine simply executes predefined processes. Therefore, dynamic exchange of carriers during run-time based on the availability of their services is not possible with current BPM-based solutions.

3 Added Value through Semantic Web Services

The previous high-level description of a BPM-based implementation of a business process shows several limitations of current solutions. The most prominent ones are: i) necessity for manual development of message mappings, ii) manual creation of the *collaborative business process* (CBP, Sect. 3.3), and iii) flexibility limited to design-time.

Using technologies developed in the semantic Web services area, these limitations of current BPM-based implementations can be overcome. In the subsequent sections, we will first describe our overall architecture for integration of semantic Web service technologies into current BPM tools (Sect. 3.1). Following this, we will describe in detail how this architecture enables i) the automatic generation of necessary message mappings (Sect. 3.2), ii) the automatic integration of the public processes of different partners into one CBP, and iii) the flexible service selection during run-time (Sect. 3.4).

3.1 Solution Overview

Our overall architecture consists of two parts: A design-time, and a run-time component. During design-time, we want to simplify the creation of the CBP as much as possible. After loading two public processes, the Maestro tool therefore should generate the CBP automatically (if possible) and present this as a proposal to the user. Furthermore, the tool should generate the message mappings necessary for invoking the involved Web services. Figure 2 shows the design-time architecture of our enhanced Maestro tool. We assume that the representations of the two business processes not only contain the process flow but also the *XML schemas* (XSD) of the input and output messages associated with each process step. After loading the two public processes specifications into our tool, the *lifting engine* generates two things: i) an alignment between the message elements of the XSDs and the domain ontology, and ii) a semantic description of the public processes. In the next step, the *mapping engine* uses these alignments between the ontology and the XSDs to generate a list of possible mappings. Now, the *composition engine* takes this list of possible mappings and the semantic process descriptions to generate the CPB which is finally presented to the user via the Maestro tool. After the user reviewed and applyed possible modifications on the CBP, the result is stored into the central process repository.

During run-time we want to enable the dynamic selection of services based on different criteria. In our scenario, we would, e. g., like to be able to select the carrier offering the cheapest price for a given shipment request. Therefore we introduce a component called *semantic service selection.* Based on the concrete request, contractual information modeled in the domain ontology, and a selection goal, the best process is being selected from the process repository, instantiated and executed.

3.2 Automatic Generation of Message Mappings

In order to create an alignment between the domain ontology and the input and output messages as depicted in Fig. 2, the lifting component executes a set of elementary matching algorithms. These matching algorithms exploit the information available in one XML schema and the ontology (like, e. g., element and concept names) to create a similarity matrix. This similarity matrix associates each pair of the XML schema and ontology entities (e_S, e_O) with a similarity value. Based on this similarity matrix, an alignment including a mapping expression between the XML schema and the domain ontology can be calculated $(A_{S \to O})$.

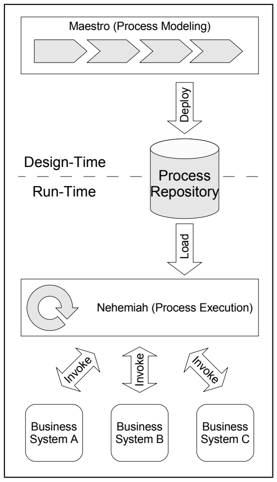


Fig. 1. Current Solution Using BPM Tools.

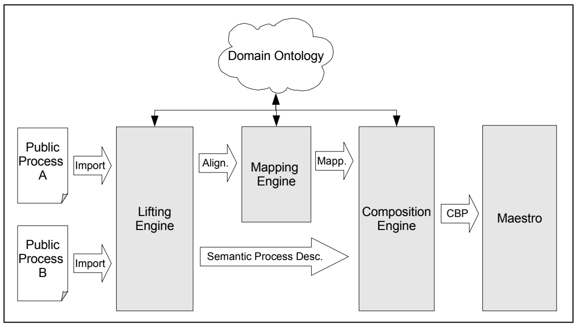


Fig. 2. Design-Time Architecture of the Enhanced Maestro Tool.

The automatic generation of message mappings is performed by the mapping engine. This component takes the alignments created by the lifting engine as input and generates executable mappings between XML schemas. In order to create a mapping between S_1 and S_2, the mapping engine takes the alignments $A_{S_1 \to O}$ and $A_{S_2 \to O}$ as an input. For each mapping element in $A_{S_1 \to O}$, the mapping engine searches for a mapping element in $A_{S_2 \to O}$ that relates a schema entity of S_2 to the equivalent ontology entity. If such an entity is found, the mapping expression is used to determine how the schema entities of S_1 and S_2 are related. This in turn creates a new mapping expression that is added to the mapping $map_{S_1 \to S_2}$. Mappings are not generated between each pair of schemas but only between input schemas of one process and output schemas of the other, and vice versa.

3.3 Automatic Integration of Partner Process Steps

For the automatic process composition by the composition engine connotated in Fig. 2, the public process description of the shipper as well as the available WSDL descriptions of the carrier services need to be transformed to a format that the composer can work with.

The composer technology we are going to use bases on the semantic Web services composition approach described in [2, 3]. For each partner which is to be integrated in the composed process, we therefore need a semantic Web service interface description consisting of the following parts: i) the messages communicated by the semantic Web service given as ontology concepts, and ii) behavioral constraints between the single message exchanges of the semantic Web service. In other words, the behavioral constraints can be understood as a workflow diagram, like an UML 2.0 activity diagram, containing control nodes, like decision, merge, fork and join. The activities in this diagram would be connected to input and output nodes representing the messages communicated. Here, each message is not understood as a technical XML schema description, but an ontology concept for which the corresponding XML schema can be nominated later on.

The main task of a composer is to combine the sets of behavioral constraints to a CBP of all parties involved. The composition engine thus basically compares the inputs and outputs that are defined as ontology concepts in the two behavior descriptions and connects them where possible. In each such connection, a transformation activity node is incorporated that defines a conversion which possibly needs to be performed in the real-time execution of the combined process. This conversion is given by the mapping function $map_{S_1 \rightarrow S_2}$.

The result of the composition is a business process—the CBP—that contains the process steps of both parties, their interconnections via mapping activities, and those inputs and outputs that could not be interconnected. The composition therefore is successful, when there are no inputs and outputs left that could not be connected to corresponding communications of the other party.

3.4 Semantic Service Selection

After discussing how semantic Web service technologies can be used to improve the design-time of current business process management solutions, we will now investigate how they can be used to improve their run-time.

As stated in the solution overview, the semantic service selection is responsible for selecting the best fitting carrier for the current shipping request during runtime. The realization of this component is described in detail in [4]. It is based on an approach for semantic Web service discovery introduced in [5, 6]. For applying this approach, an abstract service capability is described based on the domain ontology. The abstract service capability is carrier-independent and covers all possible Web service capabilities within the domain.

Additionally, a successful offline negotiation between a shipper and a carrier is required. The result of this negotiation phase is a contract between that carrier and shipper describing the provided service capabilities by that carrier. Each contract is modeled as a sub-concept of a service capability concept of the domain ontology. These semantically described contracts are stored as OWL documents in a separate repository. The concrete shipping request created at run-time is then described either as an instance or as a most specific sub-concept of the abstract Web service capability according to the domain ontology. A shipping request can be fulfilled by a carrier Web service if the the concrete request subsumes a Web service capability.

If more than one contract matches the concrete request, an additional selection step is required in order to choose between the available carriers. This step usually requires run-time invocation of the carrier Web services in order to get information necessary for the selection according to the goals of the requester. A *selection goal* specifies the criterion to nominate the best suiting carrier, e. g., best price or shortest delivery. Since these two parameters are subject to frequent change due to the competition on the carrier market, we decided not to design these parameters in the semantically described contracts.

After the selection is done, the process associated with the selected carrier is loaded from the process repository, instantiated and executed in the Nehemiah component (see Fig. 2).

4 Summary and Outlook

The proposed carrier/shipper-scenario trys to keep the described system simple in order to be able to concentrate on the important steps first. The important aspect is mainly to examine how semantic technologies can beneficially be applied to real-world business scenarios. We identified the automation of the so far manual business process integration as the main area of contribution for semantic Web technology. The solution extends and therefore bases on standard BPM modeling tools.

Furthermore, we abstain from the requirement of business partners to adhere to exactly the same software component interfaces. Thus, mediation comes into play and its interacting with the semantic Web service composition is a second aspect to focus on using this scenario.

After the successful implementation of this first scenario, the described setting can be extended to a more comprehensive application in a later version. In the current proposal, the composed business process is being created during design-time. When a carrier changes its conditions, the process of composition needs to be executed again in order to compute the potential adoptions to the process instance. In a more dynamic implementation, this step could be executed each time a customer requests a shipment. This way, the system would immediately and automatically incorporate changes to the carrier capabilities.

References

1. Greiner, U., Lippe, S., Kahl, T., Ziemann, J., Jkel, F.W.: Designing and implementing crossorganizational business processes - description and application of a modeling framework. In: Interoperability for Enterprise Software and Applications Conference I-ESA. (2006)
2. Albert, P., Henocque, L., Kleiner, M.: Configuration-based workflow composition. In: ICWS, IEEE Computer Society (2005) 285–292
3. Albert, P., Henocque, L., Kleiner, M.: A constrained object model for configuration based workflow composition. In Bussler, C., Haller, A., eds.: Business Process Management Workshops. Volume 3812. (2005) 102–115
4. Friesen, A., Namiri, K.: Towards semantic service selection for B2B integration. In: submitted to Methods, Architectures & Technologies for e-Service Engineering (MATeS) at ICWE. (2006)
5. Li, L., Horrocks, I.: A software framework for matchmaking based on semantic web technology. In: Proc. of the Twelfth World Wide Web Conference. (2003)
6. Preist, C., Cuadrado, J.E., Battle, S., Grimm, S., Williams, S.K.: Automated business-to-business integration of a logistics supply chain using semantic Web services technology. In: International Semantic Web Conference. (2005) 987–1001

USE of SEMANTIC TECHNOLOGIES
AT Agence France-Presse (AFP)

Stéphane GUERILLOT

13 place de la Bourse
75002 PARIS, France
Stephane.guerillot@afp.com

Abstract. Adding value to content of various media types and selling and delivering customised content to various types of customers can be the definition of any news agencies' business today. But the insertion of taxonomies in the information workflow is a challenge that needs prototyping and experimental phases because of the impact on the production cycle itself.
With the increasing pressure of real-time information, this is a practical example of semantic web applications in an "adapt or die" context.

1) Introduction

A Long Tradition of Newsgathering

AFP is the world's oldest established news agency, founded in 1835 by Charles-Louis Havas, the father of global journalism.
Today, the agency continues to expand its operations worldwide, reaching millions of individuals via thousands of subscribers such as radios, televisions, newspapers, administrations… from its main headquarters in Paris and regional centers in Washington, Hong Kong, Nicosia and Montevideo. All share the same goal: to guarantee top quality international service tailored to the specific needs of clients in each region.

AFP provides a 24/7 worldwide multilingual news coverage and produces daily an average of 2 millions words, 1000 photos, 50 news graphics, as well as video reports and multimedia products for the Web. French, English, German, Spanish, Portuguese and Arabic are the languages used.

A news agency is to report about the news, the facts with speed and reliability. *Accuracy* is the guiding principle. It has to get to the heart of the issues and to provide coverage and analysis of world events. As well as general news from around the world, AFP is offering economic and financial news, sports coverage, human interest stories, celebrity news, science, culture, new technology, lifestyle and offbeat items.

To do so, AFP is relying on 1.200 reporters, 200 photo-reporters and 2.000 stringers based in 165 countries.

2) The Challenge

With the Web development, news agencies are no longer THE player for news and information management. As a direct impact on our day to day production cycle, our customers as well as our own journalists are expecting more and more functionalities and help to work with the huge pressure of the information, its volume and its continuous flow.

It is always on and those "users" can easily be submerged by the amount of information to process and manage. As an example, our French clients are receiving more than 1200 news items a day and between 500 to 800 digital pictures. Our Image Forum online database is offering more than 7 million pictures plus an even wider selection of the daily production.

The characteristics of the search engines on the Internet can lead us to believe that information search and accurate match are now possible in an efficient way and that full text search tools are no longer the ultimate solution in our business environment.

Our Clients (encompasses the internal clients – our journalists – and our subscribers) have several access tools to the information including Web platforms and News Editorial systems and their requirement is going *beyond the language barrier* and the information *nature*.

They want to:

- Subscribe to selections by themes,
- Browse quickly and efficiently through a large corpus of multimedia documents, including archives.

Our main goal is to provide, within AFP, the solutions within its production cycle and help for the implementation of the changes required to offer those potentials to our Clients via our next generation of Multimedia Editorial System.

3) Search and Filters

In the news business, the search and implementation of real-time filters are based upon concepts and topics.

The selection is made on information of a known nature (Text, Picture…) and precise categories which can be taken from a list of Subject Matters (such as *alpine skiing* or *cinema*) or Named Entities of different classes (*City*, *Person's Name*). Sometimes it could also be related to a specific Genre of information such as *magazine* or *obituaries*.

So Nature, Subject and Genre are the main criteria for selection of a news item.

We should then offer:

- On our real-time and archive platforms the possibility to quickly find information regarding the Vatican, Benoit XVI, Ségolène Royal, Tom Cruise, the Oscar 2006 awards as easily as the information related to football or bio technologies;
- To easily browse between concepts as per example between Antoine Deneriaz, the Subject detail of alpine skiing and an Event such as Turino 2006;
- To allow our Clients to simply create either on our platforms or even in their own Information System, the search interface and alert filters based on the news items metadata as made available by AFP.

4) What is available today?

A news agency is a Factory for News and the Journalists are, as in many cases under pressure to release their production as quickly as possible and accurately. They are willing to enrich their reports if the tools made available for them are efficient, simple to use and do not interfere too much with the constant request made to them to beat the clock.

Our content specificities are essential when it relates to mark-up and enrichment.

1. AFP is working from production to distribution in 5 different languages. All the production, in every language, is made available to each and every desk. They constantly are exchanging information, asking for details, explanations or follow-up.
   It is then essential to provide them with search and selection tools operating through multiple languages at a same time.
2. Most of the news items are related to subject matters (such as *Politics*, *Art*, specific *Sports*) or persons (G.W. Bush…) or events (European summit) or organisations (Political party, Microsoft…) or "products" such as *Da Vinci Code*.
   The news items are often short and concise. The main issue is mentioned in the first two paragraphs which means the first 100 words. Then the rest of the news item is more about the context or background information.
3. AFP is producing between 5 and 10.000 documents per day. During the peak hours we could have one document validated every second.

Production:

Up to now, the "production" is organised by Nature. We have multiple production lines running in parallel. For example, Text and Photo are managed and processed with their own systems from the journalist (photographer), the keyboard and the camera, to the desk.

The main driving line is then concentrated on what is called the *slug* or *slug line*. It is limited to 24 or 64 characters (depending on the distribution data format) and contains a set of keywords including named entities and controlled vocabulary taken from standard lists. This set is adapted permanently to follow the news focus.

For the Text production line, the slug line can be specific to one desk and or a language and or the desk specialisation (such as sport, business…). Because of this required flexibility it is impossible to use them as a universal and reliable reference in the general news domain but only in the sport and the economy and finance production lines.

On the Economic & Finance desk, we have also included an automatic parser to search for company or organisations names appearing in the news item to generate the ISIN code and by that make the Client search or indexing more reliable and accurate.

<Slug> Japan-IT-camera-company-Canon <Slug>

<Title> Canon latest to pull out of film cameras </Title>

TOKYO, May 25, 2006 (AFP) - Canon Inc. said Thursday it would stop developing film cameras, joining a growing number of high-tech firms pulling out of the sector as digital cameras take over.

"The situation is very difficult for new (film-based) cameras," Canon president Tsuneji Uchida told Jiji Press in an interview.

The announcement by Japan's largest digital camera maker followed similar exits from film cameras by rivals Nikon and ailing Konica Minolta.

…

hih/sct/mtp

Our Text Editorial systems then provide with semi-automatic categorisation based on the words entered in the *slug line*.

Our Photo Editorial system is also providing for additional manual categorisation to bring consistency with the slug used in the Text services plus the inclusion of person’s names, locations, precise subject matters or details and keywords taken from our AFP taxonomy for Images.

In both cases, we are using the IPTC News Code taxonomy (hierarchical with 3 levels).

Distribution:

AFP news services are either distributed as complete services in a define language (managed by one editorial desk) or as a selection of news item sorted by one of our filtering tool.

This last tool is working from concepts (topics) that are defined with the words used in our unstructured text and stored in a reference database. They are then used against each news item validated in our production line and are nearly applicable across any defined language (but Arabic).

To show the limits of such a solution, the continuous evolution of our *slug* words can lead to complex situations: If “Katarina” is used in a slug line during the hurricane reports in 2005, how should it be considered in the future? Is it then becoming a generic term or a synonym?

When made available to our Clients, they want, from their interface, to apply automatic filtering for alerts and to ease their work. Imagine how you will deal with thousands of news items arriving in your mail box on a daily basis. With a standard full-text searching tool you might find too much hits and even worse miss some important ones.

When applied on Photo services and because of the caption writing style (often reduced to one paragraph and less than 150 words) this is critical.

If you add that the end user’s interface is most of the time offering the multi-criteria or advanced search as an option and that the Boolean logic or operators are not well managed by the human being behind the screen, you would understand the value of enriched content applied to News.

5) Our project

The classification and enriching services enable the association of Metadata to the documents.

Some of the information is filled through assisted input (scroll, combo-box) by the user, other information can be assumed from the content.

These services depend exclusively on the document’s content and are independent of the editorial platform; they should be accessible through the network and can be called by the users whether they are producers, editors or archivist.

To allow the users to work offline, the necessary tables/data to run these services locally are downloaded on the editor’s workstation.

The classification and enriching services are the same for the production and editorial functions.

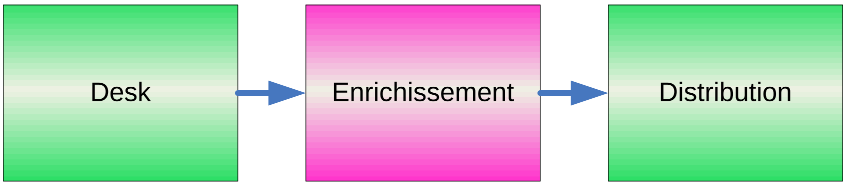
They must comply with the timely constraints that are part of the user's tasks/job.

Automatic indexation

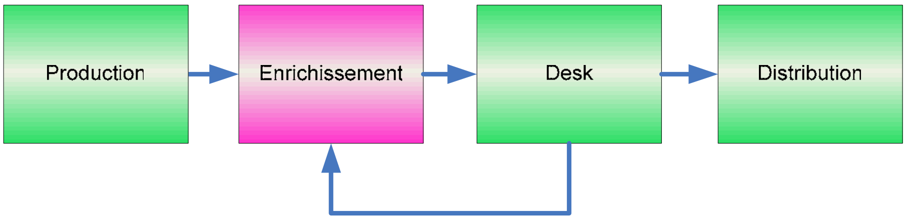
To answer the need of search be concepts in various technical environment, within AFP platforms as well as at our Client's, we have concluded that we had to enrich the information at the desk level.

It is this information which is later on in the distribution cycle made available to our Clients, on line, on the Web or within their professional applications.

Several projects were studied and one prototype is based on a purely automatic enrichment. Those enriched documents can then be sent to our selection engine with an acceptable percentage of errors.



If we later on wish to refine this process we shall have to include in the loop a control by the journalists and, ideally this would be done without delay during the validation process.



Learning process

The initial project is also based on the use of a selection of less than 100 elements from the 1300 terms already defined in the IPTC taxonomy – News Codes.

For each category a set of carefully selected news items in a semi-automatic mode. They will be used as references or learning corpus by the system to assign one or multiple categories by using similarities.

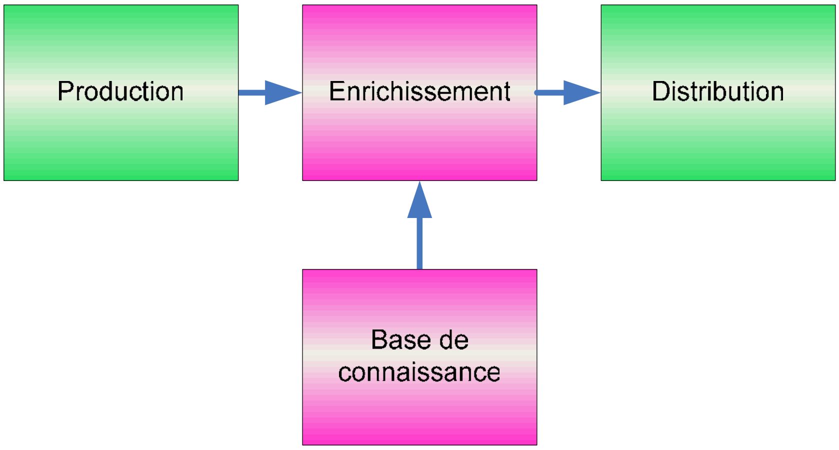
A post processing refinement using additional rules could be added for some categories when the first process is not conclusive.

The benefits of a Knowledge base

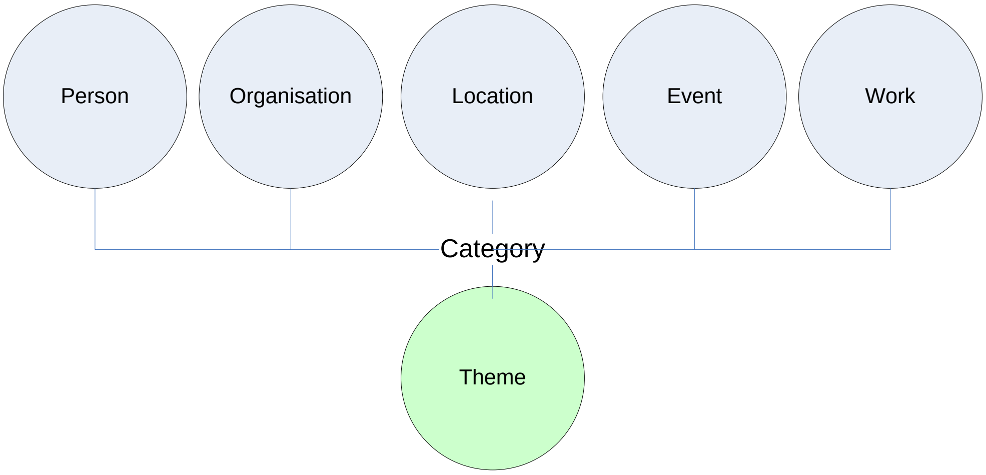
A knowledge base is included with a set of selected Named Entities. It is similar to a dictionary with locations, organisations, companies, persons, events… as needed when appearing within one of the editorial system at AFP.

Each entity can have properties such as alias or descriptions in different languages or specificities such as the year of birth for a person.

This knowledge is bringing the necessary consistency and reliability needed for all the content processed by our various systems.



Those named entities can also be connected through a semantic network with complex topics. This is essential if we want to connect multiple elements such as a company name to a location, a person, his role within a company…



The knowledge database is then a reference for the generation of interrelated content with internal and external links.

The system is also able to infer Named Entities not pertaining to the Knowledge database or relations between Named Entities from linguistics rules. Those potential new entries could then be proposed as new candidate to an administrator which is essential in a constantly moving environment as the one found in the News industry.

6) Conclusion

Adding value to content of various media types and selling and delivering customised content to various types of customers can be the definition of any news agencies' business today.

We know how to gather and produce the basic information but our main goal is now to increase the value of our products and services to our Clients - being "mono" or "multi" media - by a wide and efficient use of metadata within our news items and links (internal and external).

As we have at the same time to safeguard the specificities of our business and remain highly competitive with speed and accuracy in our reports, this project is in fact forcing a redesign of our Editorial systems in a broad sense.

We have anticipated the current trend and for the past 6 years have been heavily involved in the development of new standards at the IPTC level as well as in enhancing cooperation with manufacturers on pilot projects.

Two different – but also complementary strategies – have been tested to reach this target.

- Rely on the journalists' willingness to enrich the production
- Implement new tools to apply the content enrichment in the background with standard taxonomies and ontology.

At the moment most of the people involved in the projects have the feeling that the standardisation will bring a lot of benefits to both the internal and the external Clients and that the cost for it will be acceptable (especially the cost seen by the producer at his end of the information chain).

We still expect that those new developments will also help increase the service quality as perceived by the Subscribers as it should ease browsing through the massive amount of information brought to them and allow for an efficient access to the pertinent documents.

One of the key for the success of these new features is also the possibility to include links between news items of the same nature as well as of different nature leading to real Multimedia content management as early as possible in the production chain.

AFP is committed to support standards, particularly those from the IPTC and the W3C.

After the implementation of NewsML™ as our standard format for distribution of XML and multi-media content, our strong implication in the development of the NewsML 2G new format should be seen as a clear sign of the news agencies to jump on the semantic Web bandwagon.

Glossary

Desk

A desk is an editorial entity in charge of receiving the production (news items) from the field or the specialised production department, to sort them, to edit them and to validate them:

- The sorting action consists in removing the information that will not be used in an editorial product managed by the desk
- The editing action consists in checking that the copy conforms to the Agency's editorial rules and is adapted to the customer's requirements. The editor can enrich a news item by adding some background information. Corrections, translation and truncation are common actions carried on the desks. Any change that could possibly change the angle or meaning of the text is made in accordance with the producer.

Document

A document is an editorial object from one nature (text, photo, graphic, video and multimedia) which follows a manufacturing process carried out by journalists and/or photographers.
It has a particular Type (NewsItem IPTC).
It goes through various stages that correspond to its editorial lifecycle: *in production, produced, in edition, published, archived, deleted.*

Event

An event is defined as « what happens and what can be potentially covered » (as news).

The event can either be planned or not. The unpredictable can take priority over the entirety of planned events.

Explicit or dynamic collection

A collection is either explicit or dynamic. An explicit collection is a set of structured elements; each of these elements can be a document or a collection.

A dynamic collection contains in addition and executable query on a corpus. Once the query has been executed, the collection that contains the result of the query becomes explicit.

Explicit Editorial Link

An explicit link is an editorial object linking two editorial objects together. It is created by actors in order to enrich the editorial products and to bring them added-value. It has a source, a destination, and a lifetime (optional).

Editorial Product

An editorial product is composed by a document stream or by collections prepared by the journalists and photographers and validated for their delivery to clients (stream service, Internet Journal…).

IPTC

The International Press Telecommunications Council (IPTC) was founded in 1965 to safeguard the telecommunication interests of the world's press. IPTC develops and maintains technical standards to improve the free exchange of news which are adopted by virtually every major news provider world wide. www.iptc.org

Metadata

Metadata is data associated to the document (but is not a part of its strict content) and enables to classify the document with various criteria.

News Codes

Is an IPTC standard to assign metadata values from predefined common sets (Subject Codes & Qualifiers, Genre, news status, news types …)

NewsML™

NewsML is a media independent IPTC standard for describing news in an electronic service environment. NewsML defines an XML based language for expressing the structure of news, associated metadata, and relationships between news, throughout their lifecycle.

The current NewsML version is v1.0, ratified in October 2000 by IPTC members.
Its new generation NewsML 2G to be released in 2007 will allow to package content across media and content types. www.newsml.org

Process

A process is the entire chain of tasks performed by a group of actors.

Production Services

The coverage of various news events is made by the journalists and photographers from production services and gives birth to the creation of news items. The same process applies through all media.

The collected information is sorted, re-read, corrected, adapted, sometimes rewritten in the production department prior to being sent to the desks.

The basic principle at production level is to create and complete the documents, as fast as possible, according to AFP’s rules.

The guiding and assistance functions of the editorial system help the journalists and the photographers to respect these rules.

Acknowledgments. This document was made available thanks to the contribution of Laurent Le Meur, head of the MULTIMEDIA LAB of AFP, chairman of the IPTC News Architecture Working Party.

Training Management System for Aircraft Engineering: indexing and retrieval of Corporate Learning Object

Anne Monceaux[1], Joanna Guss[1]

[1] EADS-CCR, Centreda 1, 4 Avenue Didier Daurat – 31700 Blagnac France
{Anne.Monceaux, Joanna.Guss}@eads.net

Abstract. Training management in a company may benefit of a better integration with competence management outcomes. This paper is about an initial exploration of this proposal. It proposes a specific approach to support the indexing and retrieval of training courses with regard to the professions' target competences. This approach is grounded on Learning Object metadata, and semantic web (SW) technologies enabling advanced search and reasoning on Learning Object description. We intend to implement it using the KINOA prototype platform that contains an annotation editor and a semantic search server. The approach requires that a semantic Learning Object repository is built on several existing data sources. Standards from IEEE LOM and AICC are used as a starting point for the building of the semantic learning object repository and extended to fit with our needs and context.

Keywords: Training management, Learning Object, Semantic search technology.

1 Introduction

Training significantly contributes to the companies' ability to react on requirements of fast changes markets, customer needs and successful business process. Nowadays, industries have a high demand for well-trained teams and in the same time face continuous changes in their work processes and tools. Not only is continuous education an important process but it is managed on a contractual basis. Therefore, training management activity is a usual responsibility of Human Resources departments (HR). Actions and decisions about training are hold by HR according to the company objectives. The important requirement for training management is that it supports developing and maintaining the right range of skills and competences needed for the employees' jobs.

In order to support continuous education of engineers participating in an aircraft program, a training management process has been implemented within an Aeronautic company. This process is supported by a training management tool. It supports training courses management, each training course description is captured, referenced and maintained, as well as employees' training history management; each training request is captured and traced, if validated, until the corresponding training session has been hold. These memorized data are intended to be reused when dealing with a new training request.

It is identify as a need that training retrieval and selection through this data be linked to the skill and competence development target. Yet, training offer/selection processes are not integrated with other existing company systems such as competence management, HRMS (Human Resource Management Systems), or CMS (Content Management Systems) systems.

This paper relates to a study that currently flows into an EU project called LUISA[1] aiming at search, interchange and delivery of Learning Objects (LO) in a service-oriented context. We restrict it to one subject raised by the definition of training services: the proposal of using competence gap analysis as a driver in the training selection process. In this framework, we first assimilate information about trainings to Learning Objects. Therefore we start by defining Learning Object and the main types of systems that make use of them in section 2. A prior problem we face is how to retrieve trainings in relation with skills and competences as needed to fit our needs and context. We propose an approach that relies on semantic modeling of training and competence management concepts, and indexing technology by means of metadata. The third section presents training selection use cases and necessary underlying conceptual model for search context expression. Then, we present our approach in the fourth section. It was built to illustrate possible indexing and retrieval of Learning Objects created on the basis of existing HR databases and materials. The approach includes the following major steps that are explained in more details in this paper:

- specify the set of metadata from available resources (that represent mostly unstructured knowledge) according to AICC definition
- model the learner's context along several dimensions (personal, organizational, topical, …) and their knowledge requirements
- based on this model, contextualize the training elements (viewed as Learning Object according to IEEE definition)

2 State of the art

Learning objects (LO) can be defined as instructional components or "objects" that can be reused in various contexts for technology supported instruction. They were first introduced in the object-oriented paradigm of computer science. These objects are thus intended to be retrieved and reassembled by instructional designers. They include in first place instructional materials and contents: a LO is "an independent and self-standing unit of learning content that is predisposed to reuse in multiple instructional contexts" (Polsani [1]). Learning resource, on line material or instructional component are all terms that are used to mean much the same as "learning object" in this acceptation. But needs related to computer-based instruction

[1] LUISA is an acronym for Learning Content Management System Using Innovative Semantic Web Service Architecture.

demanded enlarging LO acceptation to cover many kinds of other knowledge elements related to computer-based instruction. For Barritt and Alderman [2] not only is a LO "an independent collection of content and media elements" but also "a learning approach (interactivity, learning architecture, context) …"

In addition to reusability, second fundamental idea is that LO digital entities deliverable over the Internet. This leads to ground LO indexing and retrieval implementation in semantic web technologies, defining metadata schema for their description. The term metadata refers to a collection of keywords, attributes and descriptive information. The search, retrieve and reuse of LO thus rely on their previous description by use of metadata.

To facilitate the adoption of this approach the Learning Technology Standards Committee (LTSC) of the Institute of Electrical and Electronics Engineers (IEEE) promotes a popular metadata schema in the domain: "Learning Objects Metadata Standards" (LOM) [4]. A LO is defined as "any entity, digital, or non digital, which can be used, reused, or referenced during technology supported learning" [3]. Each one can be described using a set of more than 70 attributes divided into 9 categories responsible for general, technical, or educational aspects of the resource.

The Aviation Industry CBT (Computer Based Training) Committee (AICC) is currently working on an IEEE LOM compatible metadata collection adapted to aviation industry training [5]. The AICC schema specifies a distribution of attributes into 11 categories and includes additional vocabulary compared to LOM.

Different kinds of systems can make use of LO in company context. Learning Management Systems (LMS) are software to support the management and monitoring of company training management activities. Current LMS enable organizations to manage learners (students, employees) keeping track of their progress and performance across various types of training activities. They usually include a catalogue of available and/or relevant courses, materials, and training events, all entities that can be represented in the form of LO. Learning Content Management System (LCMS) are software to support the management of instructional materials and contents. LCMS main target users are authors of training materials and instructional designers. The business problem they aim at solving is the storage and sharing of reusable contents to support the creation and delivery of new learning materials throughout the company. Training contents are managed in the form of LO.

Therefore learning object metadata become the fundamental element for both LCMS and LMS complementary technologies. The grounding of Learning Object indexing and retrieval implementation in semantic web technologies fits well with our objective (using competence gap analysis as a driver for training's selection). The main interest relies in the innovative features of SW architecture allowing linking of metadata elements with the ontological representation required for search context consideration.

3 Use case description

As underlined above, our goal is that trainings' selection can be based on the opportunity they offer to bridge the gap with the profession target competences. In this section, we present the existing information sources, use cases to illustrate their use, and requirement to provide contextual description for training search.

3.1 Training related information sources

In our current context, training related information and materials (final learners, training modules, sessions, etc.) are stored and managed in several data sources, mainly:

- SAP database allow the management of the training course descriptions and of the employees' training history. In this database Human Resources capitalize the information about the available training courses: each training course is referenced and described by means of a label, a summary, source organism, etc. and keep track of requested, planned, rejected, accepted or completed training sessions for every employee.
- An intranet Catalogue is published based on the training database. It corresponds to a selection of the core offer, build on the more usual and recurrent trainings.
- Independent repositories contain some training programs and training materials. These materials are edited and managed independently in form of MS PowerPoint or Word charts.

3.2 Competence related information source

As regards the description of competences (abilities to perform some activities), skills and knowledge (knowledge and know-how that must be demonstrated for a given competence), a profession's competence and skill index has been defined. It is structured by main fields (families of activities) in the company, such as Engineering, Information System or Architecture. A competence and skill combination build a profession profile. To illustrate this description, we take the example of the 'Application Architect' profession in Figure 1.

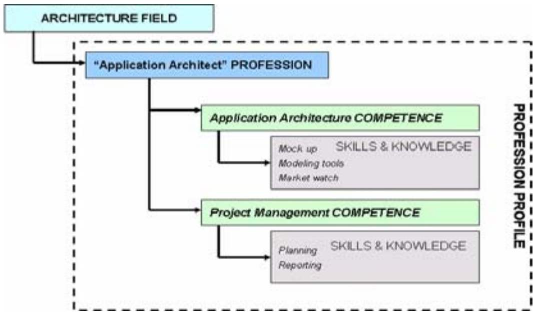


Figure 1- Profession competence index (excerpt)

The reference competence index supports the deploying of competence management throughout the company. Any job or position in the company (meaning a position hold by an employee) can in its turn being defined in terms of target "proficiency levels" defined on the basis of the competences associated to each profession: we call it profession profile. Requirements for the target profession profiles are provided by operational managers. They are used afterwards as input for comparison with actual employees' profiles, leading to identifying possible competence gap. Human Resources capture and maintain these competence grids in a dedicated database, for subsequent competence management actions among which training activities.

3.3 Training selection use cases

Two use cases have been identified where training selection involves competence gap analysis.

- Engineers express individual training needs and requests. Then, training managers study each request and accept it or not according criteria such as budget compatibility and availability of a training course relevant to the expressed needs, but also an assessment of the well-found nature of the request.
- Employees have annual interviews with their direct manager; these interviews aim at allowing comparing their skills and competences with their profession profile. At this occasion, competence gaps are stated which lead to expression for training needs and eventually to selection of trainings.

We comment the first case 'As Is' procedure illustrated in Figure 2. The employee doesn't access the entire training database but may browse the intranet core catalogue. He addresses his request to Training Manager (by phone call, e-mail, etc.). The Training Manager analyses the expressed needs using available resources (the training history and database, profession's competences and skills index) and finally proposes (or not) a training course referenced in the database. He has no access to the actual employee profile. Finally, the selection relies on the Training manager's experience, his knowledge of training prerequisites and goals, and his ability to recall fitness to a particular Profession.

To go on with the second case: interviews are clearly situations where training needs are defined. Yet, available resources to support possible gaps and trainings needs identifying with regard to expected position profile being not linked:

- Position Competence Assessment Grid driving comparison between a position profile and an employee's one
- Web Training Catalogue (an online view over the training database that comprehends core training offer, not competence indexed).

Although this use case typically involves the knowledge of the competences related to a given profession and proficiency level required, this knowledge can not be exploited to query the training catalogue. This second case finally results in addressing a request to training manager as described in first case.

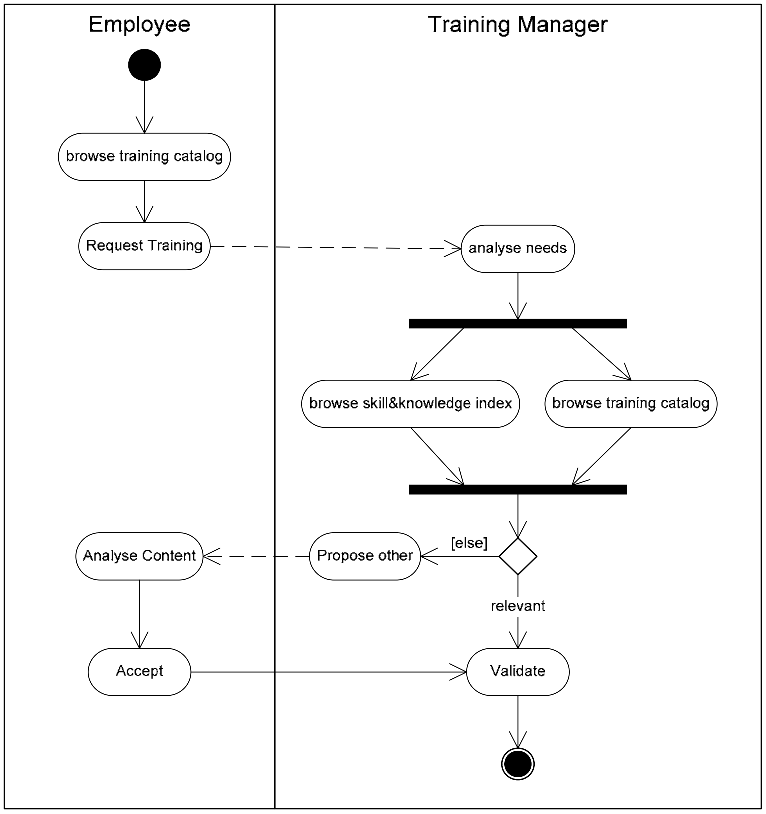


Figure 2 - As Is training request procedure

As a conclusion, we shall underline that actual solution cannot address the need for linking the competences related to a given profession and training courses or programs: this would mean very costly evolutions or modifications of static models coded in databases structure. But semantic web technologies tackle with this type of problems: advanced search functionalities using a global ontology allow combining and even deducing new knowledge from existing resources.

3.4 Requirement for a context aware search

Requirements are cascaded as follows:

1. Training selection service should contribute reducing competence gap,
2. Training search function to support selection should take competence gap description as a criteria,
3. Training description (in form of annotated LO) should include the competence term.

No metadata are currently intended to support description of Learning Object associated competences. Consequently we need to refine LO description to enable retrieval of trainings that fit with profession profiles. To do so, training goals are assimilated to the Learning Object target competences; prerequisites are the required competences that condition request validation and registering to a session.

The key point towards context-aware learning object delivery in our context is that both trainings goals and prerequisites must be described in terms of competences. This is where we face a different kind of problem related to the cost of manual annotation in time and resources, especially when training database is continuously evolving to mirror update offer.

This raises a secondary requirement: the possibility of supporting LO annotation.

4 Towards semantic search and annotation support

With regard to the requirements, we propose implementing a semantic search function over a repository built on the several data sources available. The primary advantage we see in this approach is the possibility of crossing information currently independent. In a second stage, we intend to make use of the allowed advance search to support new annotations.

The actual implementation of the model is not reached yet. Thus this section provides an initial exploration of the approach, which in our view includes:

- define the needed set of metadata to annotate LO,
- create an ontology as a unifying model for existing information,
- export and transform data from existing sources,
- define search involving mapping over the annotation files and the ontology that provide cross views over the resource, and inference support for annotating the LO.

4.1 Semantic search platform architecture

The implementation relies on the KINOA platform [6] first developed to support shared ontology-based annotations on documents. It allowed similar implementations in other application domains. The main reason for this choice is that KINOA integrates Corese Semantic Search Engine [7], which we would experiment to retrieve corporate learning objects. The architecture is shown in Figure 3. It contains:

- Digital repository with resources expressed in RDF language: training courses, sessions, but also, employees' profiles and professions. It is built with information exported from Training, Training History and Competence (Interview) databases and transformed according to training ontology,
- The ontology expressed in RDFS language
- Semantic Search Server (Corese) that will index these RDF resources files and uses the ontology and inference rules to support search functionalities.
- Search and visualization interfaces (based on the ontology).

Training resources can be enriched and modified by an annotation editor (not represented on the figure).

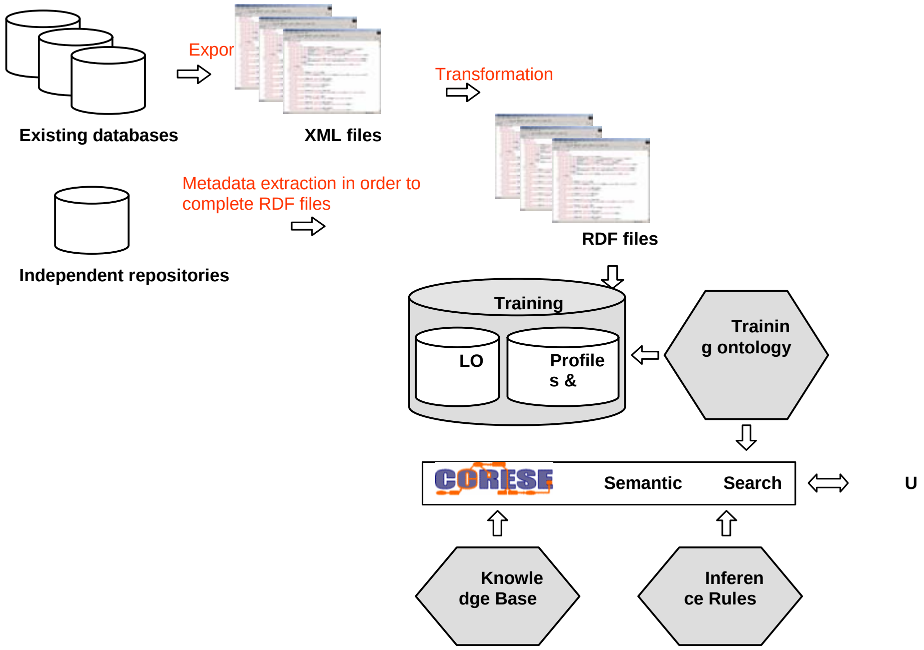


Figure 3 - KINOA search platform architecture

4.2 LO metadata definition

The existing training resources will have to be described before their integration into LO digital repository. We refer to the AICC schema to define three types of corporate learning objects in our context.

- Assignable Unit: Training materials used in training sessions (mainly documents)
- Structured training Package: Training courses descriptions (Training database)
- Training Program: Description of some logical linking between the training courses, often in form of graphical representations in MS PowerPoint format

This structuring conforms to the typology of LO granularity levels defined in AICC standard (Figure 4). Each component type comes with a specific set of (standardized) metadata, not represented in the figure.

Level	Term	Description
1	Asset	pieces of content or assessments that usually can’t be used by themselves, such as images, animations, text, video, questions.
2	Launchable resource	a grouping of one or more assets bundled together for a single launchable resource, such as a web page.
3	Assignable unit	a self-contained “chunk” of data consisting of one or more assets or launchable resources. An assignable unit is the first level of aggregated objects where assets are combined for a particular stand-alone purpose. An assignable unit is the lowest level that

		can communicate with an LMS.
4	Structured training package	a digital description of Assignable Units, Launchable Resources, and Assets, including off-line activities (simulator sessions, classroom sessions, etc.). Sequencing information and the structure may be hierarchical with many levels, or flat.
5	Training program	a collection of structured training packages related to a specific syllabus, or curriculum. It includes a digital description of the structured training packages, as well as sequencing information for the structured training packages.

Figure 4 - AICC Learning Objects granularity levels

The definition of the metadata depends of the foreseen application (and ontology) and new metadata may be proposed in the process. For example, Training course database contains attributes that either match with the LOM or AICC standardized schemata (label, identifier, summary, source organism, etc.) or require the developing of a local metadata (target profession, concerned subsidiary company).

Related to the session object stored in Training history come other existing LO metadata such as cost or duration. But more interesting in our context, it provides the link between Training and Competence management database. This will be further explained.

4.3 The ontology

The domain application ontology is needed to semantically describe the metadata and describe relations between concepts related to the several data sources. It aims at provide the appropriate model of manipulated concepts and help establishing cross data sources relationships.

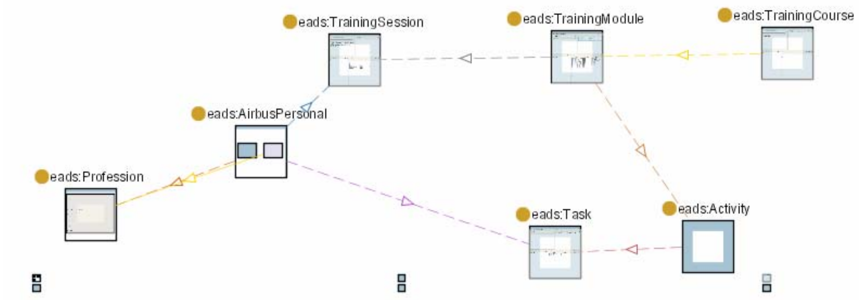


Figure 5 - Training selection ontology (extract)

We are currently developing this model. It will evolve during validation and experimentations with HR training managers.Main concepts are Field, Function, Profession, Competence, Skill, Knowledge, Proficiency Level, Employee, Training Course, Training Session, Training program... An analysis of information sources allows identifying naming conventions to express the modelling concepts, especially to extract some semantics (relation between data). Schema in Figure 6**Erreur !**

Source du renvoi introuvable. shows an abstract of some data related entities and established relations. The modelling notation is UML entity-relationship-diagram.

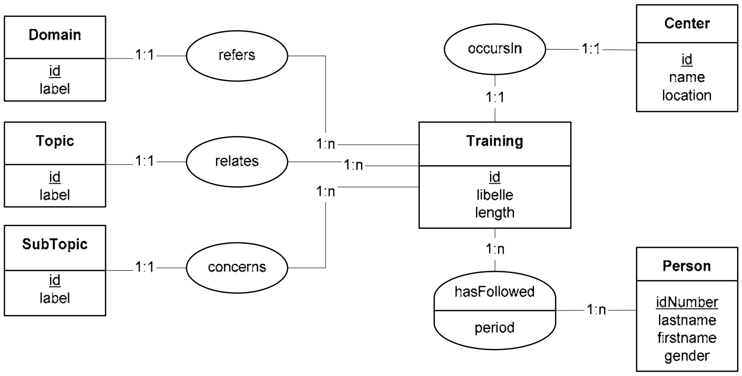


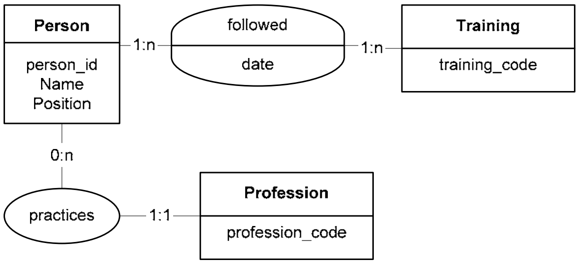
Figure 6 - Training History Database Schema

4.4 Data transformation

Based on the export from the databases, annotations files are built in accordance to the metadata schema and the ontology. A transformation of these initial data results in RDF files as input to the KINOA platform presented (see Figure 3).
The transformation step specifies for every metadata describing the LO what is to be created from available databases.

4.5 Search description

The core feature of Corese Search Server [7] is the ontology based search and inference rules processing. Following our idea a search for Training must involve a description of profession profile's competences. The model of the domain shoes linking concepts, Session and Person, between respectively Profession profiles and Training Course (a session being a particular implementation of a training course)..



It can be used to search and filter training courses adaptively to the employee. For instance, based on the employee's competences (already acquired) and his profession profile (competences required), we can deduct the training modules of interest.

As said above, another objective is to complete missing information by using inferences rules. Existing Training course description (metadata) do not usually contain information about target competence or profession. Reasoning on the employee's profession and its linked competences enables to propose some missing metadata (list of possible 'target' competences to be related to a given training module). The definition of rules will be done in collaboration with HR actors.

5 Conclusion and Perspectives

The use of competence related information is a way to improve the efficiency of training management. We propose an approach to support the indexing and retrieval of training courses with regard to the professions' target competences. This approach is grounded on Learning Object metadata standards and semantic web technologies. We intend to implement it using the KINOA prototype platform that contains an annotation editor and a semantic search server. Ontology base search enables search by type of concept and by relations between concepts. Moreover, specific knowledge of a domain can be added to the data of the repository by using inference rules.

So far, conceptual models and implementation steps have been defined. The next steps, besides actual implementation, will be to work with Training Managers to assess the relevance of searches and to define inference rules that will allow complementing the annotations.

References

[1] Polsani, P. "Use and Abuse of Reusable Learning Objects", Journal of Digital Information, Volume 3, Issue 4, 2003

[2] B. C. Barritt, B.C. and Alderman, Jr. F. L. "Creating a Reusable Learning Object Strategy". Pfeiffer, New York. 2004

[3] IEEE-LTSC Learning Object Metadata Working Group. Scope and purpose. Retrieved Mai 2006. Available at http://ltsc.ieee.org/wg12/

[4] Final Draft Standard for Learning Object Metadata (LOM). IEEE, Retrieved Mai 2006. Available at http://ltsc.ieee.org/doc/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf

[5] "Aviation Industry Metadata Description DRAFT", Draft version 1.6, January 2006

[6] Longueville B., 2005, "KINOA: a collaborative annotation tool for engineering teams", International Workshop on Annotation for Collaboration, Paris, France.

[7] Corby O., Dieng-Kuntz R., Faron-Zucker C., 2004, "Querying the Semantic Web with Corese Search Engine", Proceedings of 16th European Conference on Artificial Intelligence (ECAI'2004), subconference PAIS'2004, Valencia, pp. 705-709.

Use of Ontology for production of access systems on Legislation, Jurisprudence and Comments

Authors:

Jean Delahousse (Mondeca - jean.delahousse@mondeca.com)
Johan De Smedt (Wolters Kluwer Belgium – johan.desmedt@wkb.be)
Luc Six (Wolters Kluwer Belgium – luc.six@wkb.be)
Bernard Vatant (Mondeca - bernard.vatant@mondeca.com)

Abstract

Wolters Kluwer Belgium publishes about specialized areas related to legislation, jurisprudence and doctrine. The paper reports on an effort to transfer knowledge, scattered over a divers set of classification, coding and index generation systems, into a central thesaurus system, modeled and controlled by an ontology.

Copyright

1 Wolters Kluwer Objectives

Wolters Kluwer Belgium publications focus on specialized areas related to legislation, jurisprudence and doctrine. Specialized areas include tax, human resources, environment, safety and security and non profit organizations.

1.1 The Challenge

At WKB, we started with a diverse set of specialized paper and CD publications. For each of these specialized products, end-of-book indexes are maintained and produced by dedicated tools. Typically these tools manage part of the editorial and/or production flow only for a small set of publications in the same specialization area.

WKB has decided to bring this content on-line and keep it up to date instantly. The on-line search quality must provide accurate access to the specialized content. Moreover, the cost for editorial index maintenance should not increase. The effort invested in classifying content must be reused in all production flows.

The access mechanisms for on-line products have some new challenges. Whereas an end-of-book index exclusively handles a specialized area, on-line content can be packaged and may be subscribed to as requested by the market (end-user needs). The new access mechanisms need to go across specialization areas.

A complicating factor in this process was the diverse systems used to mange the content itself. Content was kept and maintained in different system ranging from file systems, over typesetting system to specialized CMS.

1.2 Strategy

The overall strategy had several axes. Those most relevant for this paper are:

- Reducing the number of CMS and introducing the Content Unit as the abstraction of a reusable document;
- Introducing the concept thesaurus model. The system had to be able to support the integration of the existing indexing systems;

- Reorganizing the editorial and production workflow of the publishing process so that content enrichment is done upstream and not bound to production deadlines.

1.3 The Action List

Consequences of the above strategy for the access structures were ambitious: Migrate from a indexing process embedded in the production phase to:

a content driven classification embedded in content acquisition and
an automated, thesaurus based index generation at production time

Production process changes:

- Automating index generation as much as possible. Corrective procedures have to become rule based and executed by the index generation process.
- With this automated index generation tool, change the editorial workflow procedures. Set up a content driven workflow for classification effort.

Provide classification services on content in multiple Content Management Systems

Objectives: classify content wherever it is managed

Service on-line production as well as CD, DVD and paper based productions, all from the same thesauri and classification.

Objective: reuse the classification information on all channels where the content is being published. The index generation process must be controlled by the delivered content. Classification information of the published CU must feed into the on-line indexing and search engines.

Reduce the number of indexing systems and merge different thesauri when possible

Objectives:

Replace the different classification systems by one new system
Facilitate the merger of thesauri
Reduce classification effort by using multilingual thesauri

2 Implemented solutions

Before detailing the thesaurus management system (TMS) we need to clarify some concepts and their naming as used in the project implementation. Some of these concepts were already touched upon in the previous chapter.

2.1 Introduction: Basic Concepts

The **Content Unit**: A unit of text identified and managed by a CMS. The CU corresponds to an abstraction of a document. The abstraction means a document can be identified independent of its version history. Each document has at least some general meta data like an *identifier* and a *title*. Example: `a law`.

The **Content Unit part**: An identifiable part of the CU. The identifier of the CU part is managed in the scope of the CU. Typically, CU-part identifiers are not changed when its text changes or when it's encapsulating CU is versioned (exception: when the CU-part is removed, the identifier is removed and will not be reused). Each CU has some form of sub-title or label. Example: an article.

The *Content Unit part* **Range**: The range specifies a contiguous text within a CU. Comment may be added to the range to describe it in a human readable form. Example: `art. 3 – art. 5`

Classification: The link between a text in a CU (the CU, a CU-part or a Range) and a set of thesaurus terms, used to characterize the text. Example: in art 3, `points 3, 5-7`

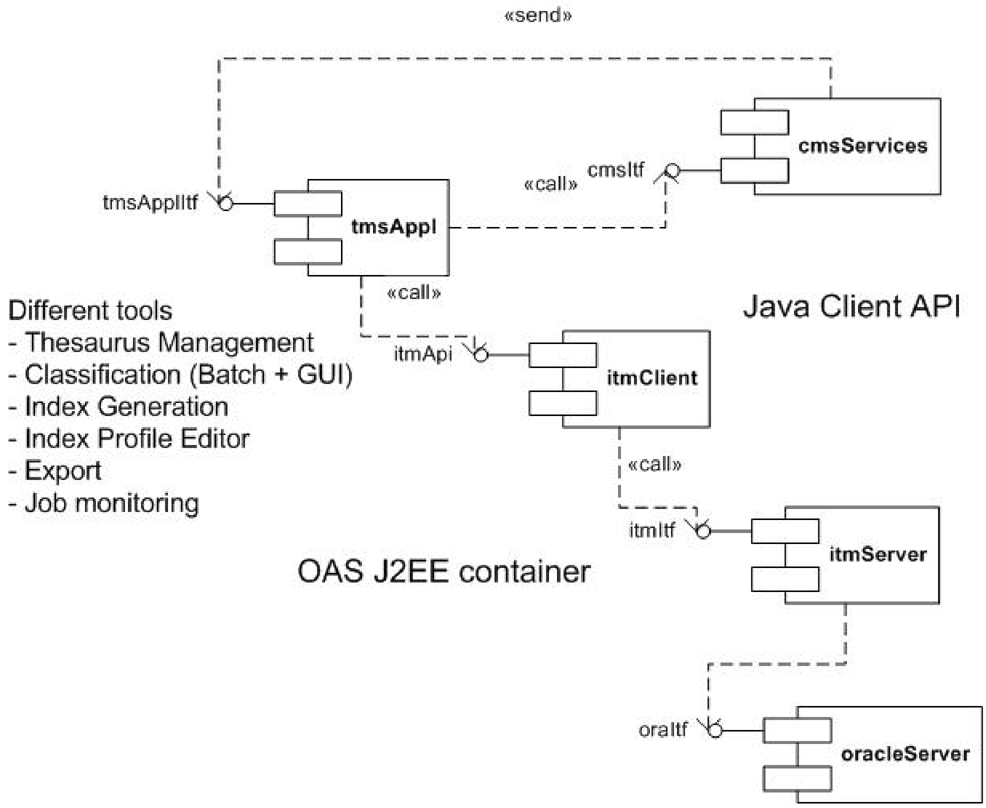
Folder: publication specific set of **content references**. Each content reference identifies the CU (or CU-part) and the location where the CU-part is included in the publication.

Index: In the presentation, refers in particular to back of book indexes.

An **index Profile** is a set of rules applicable for a publication, It is invoked on a folder referencing CUs that are classified by a thesaurus. The result of applying these rules is a back-of-book index.

2.2 System Architecture

The thesaurus system is a set of applications, built around a J2EE service (ITM from Mondeca). The general architecture is depicted in the picture below. Some TMS applications may access other services as well. The CMS is a typical example of such a server. The TMS application defines an interface for its CMS access. Next to the depicted API interface, ITM also offers web interfaces.



2.3 Manage multiple Thesauri

The business provides us with a synergy problem. Occasionally, multiple end-of-book indexes are the basis for thesauri. Some were conceived in well separated and specialized domains. Others however have a great overlap (e.g. French and Dutch indexes on the same content).

To handle these two synergy problems, the thesaurus management module has two strategies: **Merging thesaurus terms** to re-organize a thesaurus and **Mapping thesauri terms** to enable integration among thesauri. Both strategies are implemented as API and in the UI.

2.3.1 Merging thesaurus terms

Merging thesaurus terms allows cleaning-up poorly formed thesauri. Typically the poor thesaurus structures result from the migration of index based classifications. Merging terms, deprecates one term for its preferred term. The following sub-actions may be parameterized:

Classifications of the deprecated term are taken over by the preferred term.

Associations (RT-RT and BT-NT) may be taken over by the preferred term.

Names and synonyms of the deprecated term may be taken over by the preferred term.

2.3.2 Mapping thesauri terms

Mapping thesauri terms allows transfer of classification information between terms of different thesauri. The mapping is registered. The classification information is copied (not moved). Names may be copied (becoming synonyms). BT-NT associations may be traversed and copied (deep mapping) or not.

2.4 CMS interfaces

In general, the CMS interface provides functions to:

- signal new or changed CU or CU meta data (CMS to TMS)
- report classification change (TMS to CMS)
- get a read-only copy CU (standard CMS service)
- get extra metadata of a CU (standard CMS service)

The interfaced CMS systems are:

- BRONS: A proprietary WKB CMS handling legislation and jurisprudence
- SigmaLink: A commercial CMS handling specialized documentation and comment.

Content in other legacy system are migrated to one of these CMS in order to upgrade their production flow to the on-line product strategy.

2.5 Classification export

Classification export is typical for on-line productions.

Thesauri are regularly updated completely. A thesaurus delivery includes synonyms and all associations.

The CU classification is delivered per CU. Each CU uploaded to the web portal has its own classification information.

2.6 Index productions

This is typical for end-of-book publications used in books, loose-leaf and on CD or DVD.

Index production is initiated by the publication build system (typically, a sub-system of the CMS, occasionally an independent system). The publishing system establishes a folder with CU-references (see above). Index generation starts from this folder. It generates the default thesaurus view for the CU in the Folder, excluding all terms that are not leading to a classified CU. Subsequently rules are applied to transform this default view. Consequences of the applied rules are:

. Name selection of the thesaurus term (commonly used names, domain specific names);
. Synonym selection for the thesaurus term;
. Term exclusions from the index;
. Index depth selection (indexes typically have a maximum depth);
. Copying and moving of thesaurus branches in the index;
. Finally, when all terms have been decided upon, only RT-RT associations connecting terms in the remaining thesaurus are retained.

2.7 Productivity Tools

Thesaurus editor: Manages thesauri (merge and mapping functions), descriptors (names, synonyms) and associations

Classification editor: Interactive tool for research and corrective action on classified content.

Index profile editor: a tool to edit a generated index. Corrections are saved as rules, applicable in an automated way on subsequent index productions.

Batch job import and export: Interactive and command-line interface to an application to launch batch jobs, parameterized by publication constraints.

Batch job monitoring: the operator tool to follow-up and manage batch job productions. The framework also monitors interactive users on the TMS.

3 Ontology based information modeling

An Ontology Model to meet WKB needs

We are now presenting a quick non-formal review of the ontology used to meet the above presented business needs. A formal presentation is available in separate OWL/RDF documents, developed using the Protégé-OWL ontology editor.

3.1 Thesaurus Ontology model

3.1.1 Context and approach used

The thesaurus ontology model is a formalization of the meta-model implicitly defined by commonly used thesaurus standards, such as ISO 2788:1986[1] and ISO 5963:1985[2]. Although developed in the framework of WKB project, this ontology is not domain specific, and can be applied to any monolingual or multilingual thesaurus built upon the above standards. It has been used since in a variety of other Mondeca projects, involving multilingual resources published by international organisms, such as the Thesaurus on Tourism and Leisure activities (World Tourism Organization), the UNESCO Thesaurus, or the European Environment Thesaurus (GEMET), as well as several domain-specific thesauri for various industrial customers.

A preliminary important remark is to stress that the approach used is **not** to convert the thesaurus structure of concepts into an ontology, where descriptors would be re-factored as classes, and generic-specific relationships as sub classing relationships. Such an approach has been sometimes used in the past, and brought about confusion and misunderstanding between the ontology community and thesaurus community. Fortunately, those communities are now on a more constructive track, as proven by the current W3C SKOS[3] project. It brings about the ontology viewpoint as a meta-level, describing the generic structure of a thesaurus, and independent of the domain and content nature (descriptors). Actually the thesaurus ontology model was built in 2004, when SKOS was still only a by-product of the SWAD-Europe[4] project and not yet on the mainstream W3C Semantic Web Activity track where it is now. Anyway, at that time, SKOS was

[1] http://www.collectionscanada.ca/iso/tc46sc9/standard/2788e.htm

[2] http://www.collectionscanada.ca/iso/tc46sc9/standard/5963e.htm

[3] http://www.w3.org/2004/02/skos/

[4] http://www.w3.org/2001/sw/Europe/reports/thes/

already closely monitored by Mondeca and its approach qualified as relevant. Hence, the core structure of the thesaurus ontology model is very similar, and can easily be mapped into SKOS core vocabulary.

3.1.2 Core structure

The core structure elements map quite exactly the standard thesaurus constructions and their SKOS representation. They allow a formal representation of any Thesaurus in a semantic format, independently of any specific application. They form the *declarative* part of this module.

The core class is “Descriptor”, which maps to the SKOS class “Concept”.

The various terms, or lexical forms, used to denote a descriptor, are attached as attributes, either preferred terms or synonyms, with a specification of the language. Abbreviations, definitions, and scope notes are defined the same way.

Hierarchy of descriptors is described using a generic “broader-narrower” relationship. This hierarchy has no added semantics regarding the original thesaurus, given its main if not only main purpose to help navigation and search of resources indexed by descriptors, and to be reproduced in index hierarchies. Technical attributes, such as “Top Term”, “Leaf Term” and “Level” can be computed from this structure.

For the same purpose, a loose associative relationship type is defined (related terms).

A “Thesaurus” class is defined, allowing integration of several thesauri, each declaring its own language, and linked to its various application profiles (see below). The “Thesaurus” class maps quite neatly to the SKOS class “Concept Scheme”.

3.1.3 Functional elements

The *functional* elements are defined as extensions of the declarative core, to meet specific process requirements. They are not exhaustively described here, some of them are presented to make clear the distinction between declarative and functional ontology.

Functional elements linked to the “Descriptor”.

“Classifying Term”. This boolean attribute indicates whether or not indexing of resources is allowed directly on the current descriptor. Typically it will be set to “false” for higher levels of the hierarchy. The value of this boolean is used to control the indexing interface.

Deprecation mechanism. A “Descriptor” can be changed through the Thesaurus Management interface into a “Deprecated Term”, with attributes “Replaced By” (pointing to a valid descriptor) and “Valid Until”. The resources which used to be indexed on the deprecated term are redirected to the replacing one, along with its lexical forms, which are kept as synonyms.

More functional elements are linked to the “Thesaurus” class itself. They are mainly linked to the publication process, and singularly the index generation, which is likely to use the same thesaurus structure or elements in different ways. Those elements are defined in the “Index Generation” module, they can be customized at will, and extended to meet more specific business requirements. They leverage the declarative part of the thesaurus ontology model, which is standard and stable, but are independent of the content of the thesaurus itself.

For example, the “Index Profile” class is bearing attributes asserting rules under which thesaurus elements are used and published in a given application context.

Thesaurus (one or more) used, and language used (for multilingual thesaurus)

Specific subclass of descriptors used (for example if the thesaurus is organized along semantic fields)

Levels of the thesaurus which will be used, if one wants to exclude too generic, or too specific descriptors.

Indexing attributes used (different types of attributes can be used to index content units against the same thesaurus)

Exclude specific descriptors, or at the opposite promote them as top entries.

Include or not synonym entries, and/or use specific synonym types (such as abbreviations), or customized synonyms defined in the profile.

Include or not associative relations

Sort and display options (use the default alphabetical order, or specific sort keys)

3.1.4 Remark on OWL species used

The index profile elements are somehow providing an extra layer of description of the thesaurus structure. Note that this description will make assertions about thesaurus ontology elements, such as "use this or that attribute", or "use this or that class". Such declarations have an impact of the characteristics on the ontology, and the species of OWL used: using properties or classes as values of other properties make them to be considered as individuals. For example an "index profile" for the "authors' table" will declare "dc:creator" as value of the property "indexed attribute". In the OWL/RDF description, in this situation, "dc:creator" is defined both as property and individual, which makes the ontology OWL-Full. On the other hand, if the ontology is defined by modules, "dc:creator" is an individual only in the index generation module, whereas it is a property in the content unit module. The two descriptions of "dc:creator" correspond to different aspects of the same RDF resource in different application perspectives.

3.2 Reusable CU ontology model

For a publisher a Reusable Content Unit is:

a valuable asset, stored in a CMS and qualified with administrative metadata (source, editor, date..)

the basic component to build a product which implies to have all the useful attributes describing the Content Unit to build semi automatic process able to select the relevant Content Units for a product publication (validity of the Content Unit, subject, Content Unit language…)

the basic content component an end user will be able to access using knowledge based information tools such as book index, taxonomies or faceted search. This implies to manage trough metadata classification links to the thesaurus level.

The Content Units management ontology is focus on delivering the proper services to the publisher when building "intelligent" end users products with the best productivity and reactivity. The ontology model for Content Unit reflects those needs:

Manage relationships between the Content Unit ontology and the Content Management system: Content Unit Identifier enables to manage the relation between the ontology repository and the CMS where the files are stored. A specific connector is created for each Content Management System to solve the Content Unit Identifier into a file address in the CMS.

Enable Content Unit classification on thesaurus using metadata. In our case, depending on the Content Units origin a single French-Dutch legal thesaurus or two different thesaurus, one in French the other one in Dutch are used.

Assure Content Unit Ontology independence with Publication ontology. Content Units are used in several publications at the same time. It was decided to keep a strict independence with Publication ontology. Publication ontology is able to relate to the Content Units trough Content Reference, when Content Units are not aware of their use and position in a publication.

3.2.1 Return of experience:

The Content Units Ontology is a technical tool for the publisher to manage content units and content unit classification, it should be independent of the product they will be used for and of end user needs for content navigation into the final product.

Content Unit Ontology model is the most business dependant part of the ontology model as the attributes (metadata) of Content Units depend on the publisher needs but also on the existing metadata schema. This is not too much of a constraint as the other part of the ontology model: Thesaurus ontology model and Publication ontology model don’t depend on Content Units ontology model.

3.3 From Machine centered ontology (Thesaurus and Content Units ontologies) to human centered ontology for book publication

Ontology model for publication needs to describe all the components of a product issue by the publisher: table of content, selected Content Units to include in the product, product indexes enabling reader to access content from an organized list of subjects.

The requirement for the “Publication ontology” was to be generic for any type of product, to be able to produce products in continuity with actual processes, to make the process change invisible for clients and last to support both French and Dutch publications.

“Product” class enables to manage descriptive attributes on a product, independently of the periodical publications. Attributes describe the publication media, the product language and links to the set of rules used for index automatic generation

“Publication” class describes the container that will link all the publication components for a specific publication. “Publication” has attributes such as publication date, publication manager… and is the binding point for the three components of the publication: Table of Content, selected Content Units and Publication index.

Reusable Content Units can be selected for numerous publications, for each publication there is a need to manage specific attributes for the Content Unit, such as a specific name depending on the context in which the CU is included. The answer was to create a class of object “Content Reference” that will be used in the “Publication” and have a link to the Content Unit. The Content Reference will support all specific attributes related to the Content Unit in the publication.

Table of Content is the backbone of the publication; it enables to organize the selected Content References into a meaningful product. Ontology model reflects the hierarchical organization of Content References. The need to respect an order between Content references explain the attribute “order” which is recalculated each time a new Content Reference is inserted in a table of content.

Index of a book can be considered as a human centered view of a subset of the thesaurus. It is a subset of the thesaurus as it must only list the terms of the thesaurus relevant to the selected Content Units. It inherits from the thesaurus the hierarchical organization of terms (Broader / Narrower terms) but also the non hierarchical relationships (Related terms) and the synonyms.

Index is customized for each product. Humans use index searching with their own term in first level index entries. Humans appreciate to have first level index entries organized in alphabetical order and to find the word they are looking for even if the index entries redirect them to the preferred term.

This mean relevant terms placed in branches of the thesaurus should be put at top level in the index, synonyms of first level index entries should be listed in the index first level also, all first level index entries should be listed alphabetically. Also the highest hierarchical levels of the thesaurus may not be relevant to the end user, this means a need to suppress highest level terms or lower level terms, or even to crunch some intermediate level of the thesaurus to make a sound full index for the publication.

Index can than be seen as a reengineered view of the thesaurus for a selected set of Content units. On the ontology model level this implies to build specific classes of objects to model the index: “index entries” instances related both Terms in the thesaurus and the classified Content Units, index entries hierarchical relationships to describe index entries hierarchy.

Another requirement for index generation is to build a reusable application able to adapt to Content Unit metadata schemas, to adapt to the publication language and to adapt to the publication media (paper, CD Rom).

Index generation is an ongoing process, done for each publication of the same product: the productivity requirement is to capitalize on a set of general rules for the index generation and on a set of local rules that can be enriched at each index generation process. The ontology models the general rules as attributes of the class “Index” (language choice, number of level of the index, create index entries for synonyms). Local rules are described in the class “Profiles” enabling to memorize specific rules to apply to a term of the thesaurus if it is used in the index of a specific product.

The automatic index generation process is a two steps process, first run enable to build a simple index based on the thesaurus structure. This simple index structure is used to visualize and edit local rules to apply during index generation such as: exclude terms, moved the term to the top level of the index, use an other word for the index entry than the one used in the thesaurus…), the second step will generate the final index applying general and local rules.

The resulting index is stored in the ontology before being exported using API or XML serialization.

There are several lessons to retain from the Index generation process:

First is this need to build separate ontology models

for a machine centered application which role is to manage the complete collection of Content Units and their classification terminology independently of any usage

for a human centered usage into a publication made of a selected set of the Content Units, a Table of Content and an Index as knowledge based navigation tool.

Second lesson is the possibility to automatically generate the Human centered ontology from a machine centered ontology using transposition rules and get a result as good as if it was edited manually.

Third lesson is that Human centered knowledge navigation on content depends on the final media: Index is very adapted to books, but navigation based on multiple taxonomies, faceted search and full text searched are very rich tools for Web/Intranet access to contents. As book index, taxonomy navigation, faceted search and rich index can all be issued from the Machine centered ontology: Navigation Taxonomy can be issued from the thesaurus structure, faceted search is based on content unit metadata, and full text search needs semantic extension using terms relationships, terms synonyms, and terms translation from the thesaurus.

3.4 A reusable Ontology Architecture

Aspects of the solution stressed in previous section show the general interest of the approach used, and can be summed up by the following points. These aspects can qualify the system presented as a ontology-driven architecture[5], meaning that architecture of the system and architecture of the ontology are fully integrated.

The ontology model is modular. Each module answers a set of business requirements, generally meeting a consistent set of tasks and possibly users (either humans or systems). Examples of such modules are Thesaurus Management (building, edition, and update), Indexing (using the thesaurus to index resources), and Index Generation (extracting relevant descriptors to build an index).

Each module is implemented in specific parts of the information system architecture, and for that purpose every module has *declarative* elements, and *functional* elements. The declarative elements describe the business objects independently of their use, in other words what is generally called the *domain ontology*, whereas the functional elements describe parameters and rules on how the information system handles those objects.

Each module is independently re-usable. For example the Thesaurus Management module can be used independently of the Index Generation module.

The same ontology element (RDF resource) can be used in different modules with different semantics, either declarative or functional, which can appear at first sight as inconsistent, but are in reality orthogonal or complementary, as the above quoted “dc:creator” example has shown. Only when merging all the modules does the resulting “OWL-Full-ness” of the ontology appear.

4 Project Status / Results

4.1 Status

The implementation and technical deployment is finished.

8 indexing system have been migrated, including the 5 mayor systems

- 2 mono lingual indexing systems have been migrated into 1 bilingual thesaurus
- 2 bilingual indexing systems have been migrated into 1 bilingual thesaurus
- 4 monolingual indexing systems (two nl, two fr) have been migrated. Cleaning and merging are ongoing.

The migrated indexing systems have resulted in thesauri used to support:

- 2 operational bilingual production sites
- 80 productions on CD end end-of-book

Remaining indexing systems are expected to be fully integrated in one of the above thesauri, either as a specialized sub-thesaurus, or by mapping them into one of the above thesauri.

[5] http://www.w3.org/2001/sw/BestPractices/SE/ODA/060103/

4.2 Costs

	ICT		Editorial		
Item	Days	Cost (€)	Days	Cost (€)	
From book to online	525	348,750	210	92,400	
Software package, Analysis, Implementation, Testing	450	306,000	60	26,400	
Documentation and user guidance	75	42,750	150	66,000	
From indexing systems to one thesaurus	270	183,600	1,000	440,000	
Data conversion	110	74,800			
Initial data-cleaning of thesaurus	60	40,800	500	220,000	
Migration to one thesaurus	100	68,000	500	220,000	
Optimization of the index process performance issues	120	81,600	-200	-88,000	per/year for all indexes
Total	915	613,950	1,010	444,400	

4.2.1 Remarks:

The analysis and testing is done by the ICT department of Wolters Kluwer Belgium. The same department has implemented the ITM software package, fully integrated with two content management systems (one legacy system and Sigmalink of Empolis). The software package can be re-used within Wolters Kluwer world wide with minor changes if they use Sigmalink.

The cost for interfacing with the European internet Platform of Wolters-Kluwer has been included (from book to online). This platform was not yet reliable at the beginning of our TMS project. Setting up a production ready system for fully automated flow took us several man months for two main sites (Human resources and Safety & Environment).

Data conversion from the old indexing systems and data-cleaning to one multilingual thesaurus (per country) is always needed. We expect the same cost for every Wolters Kluwer sister.

Companies which have already investigated in maintaining one thesaurus will have no editorial cost.

Maintenance cost of the old indexing software systems is not included in the above mentioned table.

During the production of huge indexes a performance issue appeared. It took us (Wolters Kluwer and Mondeca) several months to redesign the architecture. Query optimization and multi threading solved the problem.

(end of report)

Data Integration using Semantic Technology: A use case

Jürgen Angele, ontoprise GmbH, Germany
Michael Gesmann, Software AG, Germany

Abstract

For the integration of data that resides in autonomous data sources Software AG uses ontologies. Data source ontologies describe the data sources themselves. Business ontologies provide an integrated view of the data. F-Logic rules are used to describe mappings between data objects in data source or business ontologies. Furthermore, F-Logic is used as the query language. F-Logic rules are perfectly suited to describe the mappings between objects and their properties.

In a first project we integrated data that on one side resides in a support and on the other side in a customer information system.

Introduction

Data that is essential for a company's successful businesses often resides in a variety of data sources. The reasons for this are manifold, e.g. load distribution or independent development of business processes. But data distribution can lead to inconsistent data which is a problem in the development of new businesses. Thus the consolidation of the spread data as well as giving applications a shared picture of all existing data is an important challenge. The integration of such distributed data is the task of Software AG's "crossvision Information Integrator" one of the components in the crossvision SOA suite [crossvision].

Information Integrator is based on ontologies. Using ontologies Information Integrator solves three major problems. First of all it provides all means to integrate different information systems. This means that comfortable tools are available to bring data from different systems together. This is partially already solved by systems like virtual or federated databases [Batini et al. 1986]. Information Integrator is more powerful compared to most of these systems as it not only supports databases but additional sources like web services, applications etc. The second problem which is solved is that Information Integrator allows reinterpretation of the contents of the information sources in business terms and thus makes these contents understandable by ordinary end users and not only by database administrators. Finally this semantic description of the business domain and the powerful mapping means from the data sources to the business ontology solves the semantic integration problem which is seen as the major problem in information integration. It maps the different semantics within the information sources to the shared conceptualization in the business ontology.

Within Software AG Information Integrator was used for a first project Customer Information Gateway (CIG) whose mission was to integrate data that on one side resides in a support information system and on the other side is stored in a customer information system.

Conceptual Layering

Conceptually Information Integrator arranges information and the access to information on four different layers (cf. fig 1):

- The bottom layer represents different data sources which contain or deliver the raw information which is semantically reinterpreted on an upper layer viz. ontologies. Currently relational databases, Adabas databases and web services are supported.
- The second layer assigns a so called “data-source ontology” to each of the data sources. These “data-source ontologies” reflect only database or WSDL schemas of the data sources in terms of ontologies and can be created automatically. Thus they are not real ontologies as they do not represent a shared conceptualization of a domain.
- The third layer represents the business ontology using terminology relevant to business users. This ontology is a real ontology, i.e. it describes the shared conceptualization of the domain at hand. It is a reinterpretation of the data described in the data-source ontologies and thus gives these data a shared semantics. As a consequence a mental effort is necessary for this reengineering of the data source contents which cannot be done automatically.
- On a fourth layer views to the business ontologies are defined. Basically these views query the integration ontology for the needed information. Exposed as Web services they can be consumed by portals, composite applications, business processes or other SOA components.

The mappings between the data-sources and the source ontologies are created automatically, the mappings between the ontologies are manually engineered and the views are manually defined queries. Mappings provide ways to restructure information, to rename information or to transform values. Up to now, we do not consider and do not plan to consider approaches which try to automatically derive such mappings [Rahm and Bernstein 2001].

This arrangement of information on different layers and the conceptual representation in ontologies and the mediation between the different models by mappings provide various advantages:

- The reengineered information in the business ontology is a value on its own. The representation as an ontology is a medium to be discussed easily by non-IT experts. Thus aggregating data from multiple systems this business ontology provides a single view on relevant information in the user’s terminology.

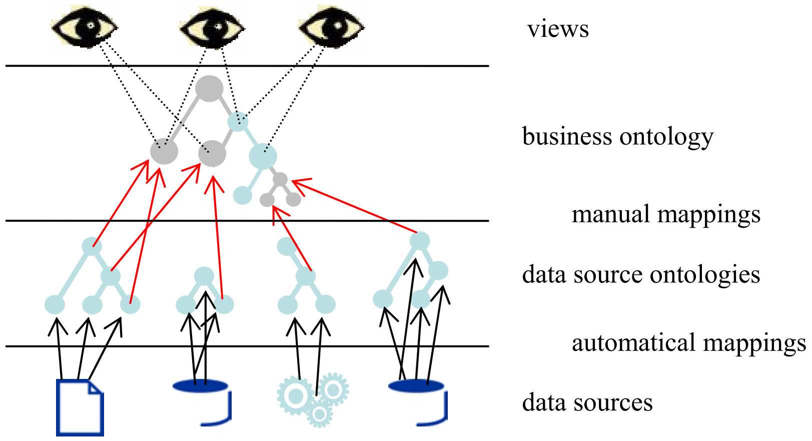


Fig 1. Conceptual Layering of Ontologies

- It is easy to integrate a new data source with a new data schema into the system. It is sufficient to create a mapping between the corresponding source ontology and the integration ontology and thus does not require any programming know-how; pure modelling is sufficient.
- The mediation of information between data sources and applications via ontologies clearly separate both. Thus changes in the data source schemas do not affect changes in the applications, but only affect changes in the mediation layer, i.e. in the mappings.
- This conceptual structure strongly increases business agility. It makes it very easy to restructure information and thus to react on changing requirements. Only the business ontology and the mappings have to be modified. Thus it minimizes the impact of change, eases maintenance and allows for rapid implementation of new strategies
- Ontologies have powerful means to represent additional knowledge on an abstract level. So for instance by rules the business ontology may be extended by additional knowledge about the domain. Thus the business ontology is a reinterpretation of the data as well as a way to represent complex knowledge interrelating these data. So business rules are directly captured in the information model.

Tool Support / Architecture

The crossvision Information Integrator provides a full fledged tool environment for defining models, for mappings between these models and for running queries (cf. fig 2). IntegratorStudio is an ontology engineering environment based on OntoStudio^{TM}.

It allows for defining classes with properties, instances of these classes and rules. Import capabilities generate "source ontologies" from underlying data sources. A powerful mapping tool allows users to interactively define mappings between ontologies by graphical and form based means (cf. fig. 3). Rules may be defined with graphical diagrams. IntegratorStudio supports F-Logic [Kifer, Lausen, Wu 1995], RDF(S), OWL for import and export. Queries which define the mentioned views can be generated and may be exported as web services.

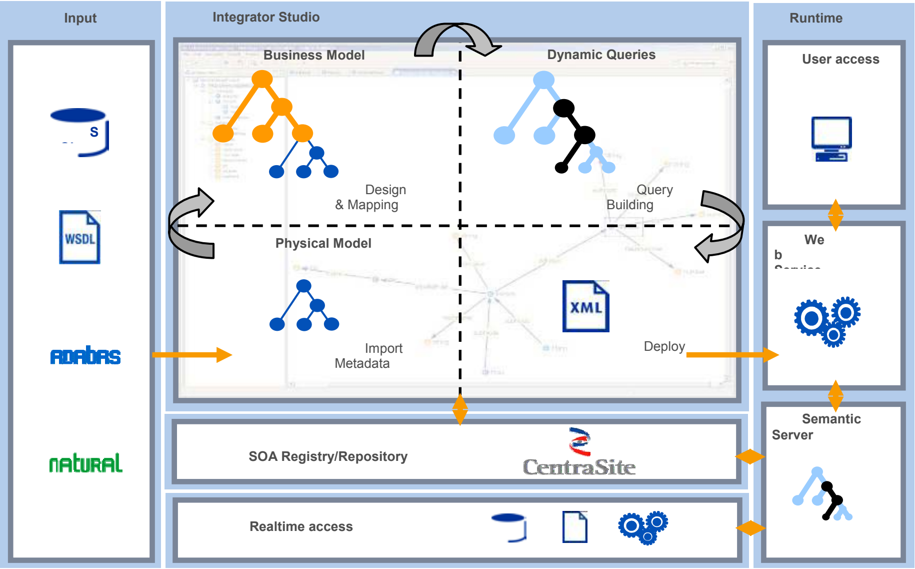


Fig. 2 Architecture of the crossvision Information Integrator

SemanticServer, the reasoning system, provides means for efficient reasoning in F-Logic. SemanticServer performs a mixture of forward and backward chaining based on the dynamic filtering algorithm [Kifer, Lozinskii 1986] to compute (the smallest possible) subset of the model for answering the query. The semantics for a set of F-Logic statements is the well-founded semantics [Van Gelder, Ross, Schlipf 1991].

Meta data like ontologies, their mappings, web service descriptions and meta information about data sources are stored in the CentraSite repository. Also, IntegratorStudio stores information about exported web services in CentraSite. During startup the inference engine SemanticServer which is based on OntoBroker^{TM} loads the ontologies from the repository and then waits for queries from the exported web services. These queries are evaluated by SemanticServer and are online translated into calls to access connected data sources.

Thus SemanticServer represents the run-time engine, IntegratorStudio the modelling environment and CentraSite the meta data repository. SemanticServer is also integrated into IntegratorStudio thus enabling immediate execution of queries to the ontologies.

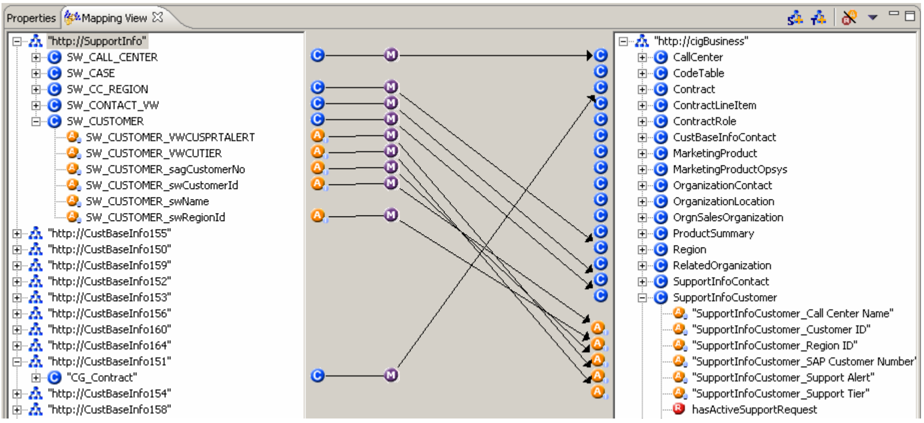


Fig. 3 Mapping Tool in crossvision Information Integrator

Use Case: Customer Information Gateway

Within Software AG the Information Integrator was used for a first project whose mission was to integrate data that on one side resides in a support and on the other side in a customer information system. The support system stores customers, their contact information and active or closed support requests in an SQL server. The customer system provides information about clients, contracts etc. in an Adabas database. The integrated data view is exposed in a browser based application to various parties inside the company, for instance to support engineers.

For illustration purposes we first sketch a very simplified excerpt of imported data and the business ontology. Throughout the following examples we use F-Logic syntax.

First of all there are two classes which have been generated by the mentioned automatic mapping from Adabas files:

F151CONTRACT [F151AA=>string; F151AE=>date].
F87CLIENT [F87AA=>number; F87AB=>string; F87AC=>string].

The cryptic names reflect the internal structure of Adabas files. The names “CONTRACT” and “CLIENT” have been specified by the user during the mapping process. Currently, the semantics of properties is only application knowledge.

Furthermore, we consider two tables from the SQL database. The generated classes are:

CUSTOMER [id=>number; name=>string; addr=>string].
CASE [caseId=>number; customerId=>string; forCustomer=>CUSTOMER].

The business ontology shall contain three classes:

Customer [name=>string; address=>string].
SupportRequest [id=>number; status=>string; issuedBy=>Customer].
Contract[contractId=>string;contractEnd=>date;contracEndFormatted=>string].

In the sequel we present some examples on how we used rules within our ontologies and derive some requirements and use cases for rule languages to be used in such a project.

Data source import

In Information Integrator user-defined built-in predicates implement access to external data sources. In the sequel we abstract from a concrete syntax of these built-in predicates. Instead we illustrate this by a generic predicate "dataAccess":

dataAccess(ci, "tablename", "rowid1", X, "rowid2", Y, ...)

where *ci* describes all parameters that are needed to call the data source, *tablename*, *rowid1, rowid2* are names of some database tables or table columns. *X, Y* are the names of variables which are to be bound by the built-in predicate.

In our example there are rules for ever class in the source ontologies which import data from external data sources. Two of these rules are:

FORALL X, Y c("F151",X) : F151CONTRACT [F151AA→X; F151AE→Y]
← dataAccess(ci, "F151", "AA", X, "AE", Y).
FORALL X, Y c("CASE", X) : CASE[caseId→X; customerId → Y]
← dataAccess(ci, "CASE", "caseId", X, "customerId", Y)].

Every functional model needs to describe relations between objects. Object properties are used to express these relationships. Object identifiers serve as object property values which are similar to foreign keys in relational databases. The foreign key definitions in a schema descriptions are used to generate object properties in source ontologies:

FORALL X, Y X[forCustomer→c("CUSTOMER", Y)] ← X:CASE[customerId→Y].

Source to Business model mappings

It is very easy to define that an object in the data source model is also an object in the business model. Similarly mappings between properties in both models can be expressed. The following example combines both mappings for contract objects:

FORALL X, Y, Z X : Contract [contractId → Y; contractEnd → Z]
← X : F151CONTRACT [F151AA → Y; F151AE → Z].

If the underlying data from the external sources contains such information, it is also easily possible to describe that two objects are the same. For example a client in the customer information system and a customer in the support information system represent the same object, if these have the same name and address. Please note, surrogate values as unique keys are typically not viable object identifiers across independent data sources. Therefore, we need to identify new identifiers:

FORALL X, Y, Z c("Customer", Y, Z) : Customer [name → Y ; address → Z]
← X : CUSTOMER [name → Y; addr → Z].
FORALL X, Y, Z c("Customer", Y, Z):Customer[name → Y; address → Z]
← X:F87CLIENT[F87AB→Y; F87AC→Z].

Often in independent data sources similar data can be encoded in a different ways, e.g. different data types or type systems. Then functions are needed which implement transformations:

FORALL X, Y Y[contractEndFormatted → X]
← EXISTS Z (Y : Contract [contractEnd → Z] and date2string(Z, X)).

where date2string() transforms a date from one format into another one.

Also, object properties need to be mapped to the business ontology:

```
FORALL X, Y, Z1, Z2 X : SupportRequest [ issuedBy → c("Customer",Z1,Z2) ]
  ← X : CASE [ forCustomer → Y ]
    and c("CUSTOMER",Y) [ name → Z1; addr → Z2 ].
```

The inverse reference is also often needed. But because the foreign key constraint in SQL systems does not provide a name for the inverse relation this is currently postponed to application development. N:M relationships, implemented by two 1:N foreign key relations in SQL systems, could also be expressed directly.

All these simple types of mappings are essential for specification of business ontologies on top of data source or other business ontologies. Most of them can be described in the Information Integrator with graphical means, i.e. developers do not need to see the F-Logic syntax.

Queries

To lower investments for learning new languages and to avoid impedance mismatches rule- and query-language should be the same. Information Integrator uses F-Logic for ontology definitions and as the query language. But, queries in the data integration scenario are much like database queries. Primarily we want to retrieve data. We are not so much interested in explanations or in information about which variable bindings lead to a result. This focus on data access requirements sometimes leads to quite complex query formulations. One example is different handling of not existing values (null values) in SQL and F-Logic. Another example are user defined projections. In order to minimize the number of expensive interactions between client and server we database folks tend to create queries which return complex structured results. Object relations should be contained in the result. E.g. for one customer having multiple contracts each having contract items, then the query result should contain the information which contract item belongs to which contract within a single result per customer.

Performance

Because the integrated view is used in an application where e.g. support engineers expect fast answers for even complex queries while talking to a customer, the performance of the rule and query processing is extremely important. In some cases response times in the range of a few seconds are not accepted. In our first project a lot of effort was spent to improve the responsiveness of the system. Problems that showed up here are very similar to query optimization problems in database systems.

Just for illustration we give two examples. First, the data source mappings as shown above always addressed only a single database table or file. However, a system that implements access to external data sources only via such single-table access rules will not achieve sufficient performance. Instead access operations should use the data source's query capabilities like join-operations. As a second example, the rule engine sometimes first retrieved all data from a table and then continued with the evaluation of filters. Instead, filters need to be identified first and given to the query which reads data from the database.

Summary and Outlook

A data model in Information Integrator consists of ontologies. Data source models describe the structure of data that resides in external data sources. Business ontologies provide a conceptualization of business entities. F-Logic rules are used to define mappings between ontologies. Furthermore, rules are the first choice to express semantics that is not immediately available within the data and otherwise had to be implemented in queries or applications. F-Logic is also used as the query language.

With the exception of mapping rules the business ontology of our first project does not contain many other rules. Access to information in these models is more data retrieval and not so much knowledge inference. Much effort during this project was spent on performance improvements.

With an increasing number of web services where some simply expose data, we also need to support data integration for such web services in our crossvision SOA suite. We are currently working on the mapping of web services and their structured XML data to source ontologies.

The crossvision Information Integrator based on ontoprise OntoStudio^{TM} and Ontobroker^{TM} is the first step for Software AG in the field of semantic technologies. Recently we joined various EU research projects like NeOn (Lifecycle Support for Networked Ontologies) [NEON], "Business Register Interoperability Throughout Europe" and "SemanticGov: Services for Public Administration" [SemanticGov]. All these projects address concrete business cases. With our participation in these projects we intend to achieve deeper understanding of needs for adequate tooling and runtime systems when using semantics technologies for data integration. On the other hand we will contribute our knowledge about data-intensive processing.

References

[Batini et al. 1986] Batini C., Lenzerini M., Navathe S.B. *A Comparative Analysis of Methodologies for Database Schema Integration.* ACM Computing Surveys Vol. 18(4):323-364, 1986

[Belkin 1980] N.J. Belkin. *Anomalous states of knowledge as a basis for information retrieval*. The Canadian Journal of Information Science, 5:133--143, 1980.

[crossvision] http://www.softwareag.com/crossvision

[Kifer, Lausen, Wu 1995]. Logical foundations of object-oriented and framebased languages. Journal of the ACM, 42; (1995) 741–843

[Kifer, Lozinskii 1986]. A framework for an efficient implementation of deductive databases. In Proceedings of the 6th Advanced Database Symposium, Tokyo, August (1986) 109–116

[Jaro 1989] M. A. Jaro. *Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida*. Journal of the American Statistical Association 84:414–420, 1989.

[Jaro 1995] M.A. Jaro. *Probabilistic linkage of large public health data files (disc: P687-689)*. Statistics in Medicine 14:491–498, 1995.

[NEON] http://www.neon-project.org

[Rahm and Bernstein 2001] E. Rahm, P. Bernstein. *A survey of approaches to automatic schema matching*, VLDB Journal 10(4):334-350, 2001

[SemanticGov] http://www.semantic-gov.org

[Van Gelder, Ross, Schlipf 1991]. The well-founded semantics for general logic programs. Journal of the ACM, 38(3); July (1991) 620–650

Integrated Access to Biological Data.
A use case

Marta González

Fundación ROBOTIKER,
Parque Tecnológico Edif 202
48970 Zamudio, Vizcaya – Spain
marta@robotiker.es

Abstract. This use case reflects the research on different and innovative ways to handle biological data repositories by means of semantic and artificial intelligence technologies such as ontologies, intelligent agents, semantic grid, etc. The human genome sequencing has given rise to a great number of biological data repositories that once analysed will be very essential for the study of diseases, pharmaceutical research, new treatments and for the development of new bio products. The problem faced is the huge quantity and heterogeneity of this kind of data and the also huge number and diversity of ontologies defined to model biological data.

1 Introduction

The aim of this use case is to provide an unified access point to diverse biological data repositories: accessible through internet (Nucleotide Sequences, amino acid sequences,…), corporate databases, results of experiments (DNA-chips), health cards, medical literature sites…This unified access has to be provide with the purpose of generation and extraction of knowledge from biological data by means of ontologies, combining them (ontology merging) and/or associating them (ontology mapping) to be exploited by means of annotations, intelligent agents, semantic web services and/or semantic grid.

Currently, a great diversity of biological data repositories exists: databases accessible through Internet, corporate databases and microarrays experiments results among others. Equally exists a great diversity of ontologies to model this data. Therefore the situation the researchers has to face with is a lot of disperse data and different disconnected and poor friendly tools to access such data, therefore the researches have to confront great difficulties to aggregate all the data to carry out the research tasks in an integrated way.

Up to now ontologies in biology were considered as mere guides for data structure, with the only purpose to access to the more adequate documents and articles to the researcher interests. This new vision will allow, combining and associating existing ontologies in the biological field, an integrated modelling of the biological data sources (genomics, proteomics, metabolomics and systems biology).

Once modelled, the annotations, intelligent agents, semantic web agents and the semantic grid will offer a centralised access point to extract and generate knowledge from the biological data repositories.

2 Biological Data Inherent Features

The inherent features of the biology field are: huge quantity of disperse-distributed-and autonomous data with great difficulties to be integrated due to their differences in: terminology (synonyms, aliases, …), syntax (spelling, file structure, …) and semantics (intra-/interdisciplinary homonyms)[1]. As instance it is complicated to know if a database table called “Species” is the same table called “Organisms” in other different database.
The following figure (Fig 1) illustrates the semi-exponential growth of DNA databases along the years:

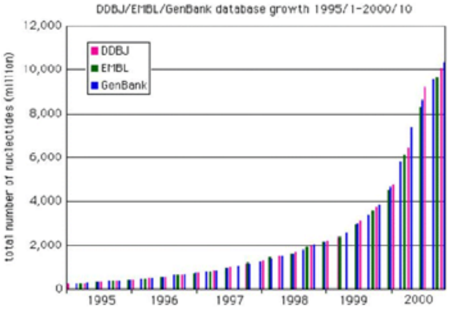


Fig. 1 Semi-exponential growth of DNA databases along the years (Source [2])

In order to highlight the inherent biology features it will be cited the most important biological data repositories and a short state of the art related to semantic technologies that could be applied to the biological domain.

2.1 Biological Data Repositories

Currently, more than 500 biological data repositories exist (584 according to "The Molecular Biology Database Collection: 2004 update”), the following repositories are considered the main or reference ones:

2.1.1 Nucleotides Sequences

These three databases are synchronized and daily updated.

- EMBL (United Kingdom)Nucleotide Sequence Database: it is the main European nucleotide sequences repository (http://www.ebi.ac.uk/embl/).
- GenBank (USA): GenBank is the National Center for Biotechnology Information (NCBI) genetics sequences(http://www.ncbi.nlm.nih.gov/Genbank/index.html).

- DDBJ (Japan): it is the unique DNA database in Japan, officially certified to gather DNA sequences for researches. (http://www.ddbj.nig.ac.jp).

2.1.2 Amino acid sequences

- SwissProt: protein sequences database (http://us.expasy.org/sprot/).
- PIR (Protein Information Resource): protein sequences database (http://pir.georgetown.edu/).
- PDB (Protein Data Bank): repository for the processing and distribution of 3-D biological structure.(http://www.rcsb.org/pdb/).

2.1.3 Gene Expression

- GDX , once of the first databases that integrates diverse gene expression data types and that was developed before biochips irruption.
- ExpressDB relational database containing yeast and E. coli RNA expression data. (http://arep.med.harvard.edu/ExpressDB).

2.1.4 Scientific literature

Many organisms offer scientific literature freely or by subscription. **MEDLINE** is health information from the world's largest medical library, the National Library of Medicine.

PubMed[8], a service of the National Library of Medicine, includes over 15 million citations for biomedical articles back to the 1950's. These citations are from MEDLINE and additional life science journals. PubMed includes links to many sites providing full text articles and other related resources.

UpToDate[9] Specifically designed to answer the clinical questions that arise in daily practice and to do so quickly and easily so that it can be used right at the point of care. The Topic Reviews are written exclusively for UpToDate by physicians for physicians. The content is comprehensive, yet concise, and is fully referenced. It goes through an extensive peer review process to insure that the information and recommendations are accurate and reliable.

2.1.5 Corporate Databases

Companies owns corporate databases with their research labours results stored, as instance, biochips experiments results. At the current situation the results of biochips experiments are stored at private researchers' databases. The boost that these techniques are winning in the biomedical research domain, along with their use extension, is helping as an important engine for the development of public databases with data from experiments; where these data can be stored for later analysis and comparison.

2.1.6 Health Cards

The Health Cards appear as a future possibility to store data that can be computer read and that it is issued to patients or sanitary professionals to facilitate the medical care attention. To store data in a card that can be computer accessible various technologies exist: magnetic strip, integrated circuit memory cards and optical memory cards. The utility of these cards could be administrative tasks, emergency health cards, specific care records and patients general medical records.

The information to be stored is under discussion, in Europe some efforts are focused on patients mobility comfort. This comfort it is desired to be reached by forgetting paper forms and adopting health cards. This means an unification of data to be stored, the media to be used and medical nomenclature. Efforts as SNOMED or GALEN are faced to this last objective.

Another use of the health cards could be the possibility to match genetic patient data with biological databases to know as instance the probability of a patient to suffer a certain disease. Also this could be applied to patients communities for statistical research in order to define priorities and strategies for health and social security and policy, planning and evaluation of preventive measures and care services and cost-benefit calculations between preventive measures and therapeutic actions.

3 Use of ontologies

Currently a huge quantity of ontologies, in the biological domain, have been defined along with specific purpose data mining tools. The data mining tools, that elaborate analysis models, as instance gather genes or experiments in function of different patterns, solve the immediate problems that the researchers have to face with, while ontologies model knowledge for the future research assistance. The problem lie in ontologies definition phase, in definition or not defined at all. In some cases a browser allows to look for information at the repositories by using an ontology[1]. Some initiatives are focused on prospective analysis of diverse biological repositories interaction, as instance: BIOINFOMED, BIRN network, e-BioSci, HKIS, INFOGENE, PRIDEH-GEN[3].

Text mining and natural language understanding in biology can also profit from ontologies. Where currently mostly statistical and proximity approaches are applied to text analysis ontologies can support parsing and disambiguating sentences by constraining grammatically compatible concepts.

To eliminate semantic confusion in molecular biology, it will be therefore necessary to have a list of the most important and frequently used concepts coherently defined so that e.g. database managers, curators and annotators could use such set of definitions either to create new software and database schemas, to provide an exact, semantic specification of the concepts used in an existing schema and to curate and annotate existing database entries consistently.

Efforts such as SNOMED provide us with vocabularies of 36,000 medical definitions and concepts, and ICD10 is the International Classification of Diseases. This kind of efforts allows to own a common language for interoperability, facilitating scientific research labours.

The following ontologies can be considered as a subset of the most widely known ontologies in the biological field:

Gene Ontology (GO) [4]Gene ontology is a controlled vocabulary that has been developed into the project OBO. GO describes how the gene products behave in a cellular context. Currently three ontologies are published at Internet: Biological

Process, Molecular Function and Cellular Component representing biological targets that a gene product contributes, biochemical activities of gene products and places where the gene products can be active. Currently some databases are annotated with GO terms.

The Microarray Gene Expression Data (**MGED**) [5] Society is an international organisation of biologists, computer scientists, and data analysts that aims to facilitate the sharing of microarray data generated by functional genomics and proteomics experiments. The current focus is on establishing standards for microarray data annotation and exchange, facilitating the creation of microarray databases and related software implementing these standards, and promoting the sharing of high quality, well annotated data within the life sciences community. The available ontology is in OWL format, it is composed by standard terms for the annotation of microarray experiments. These terms will enable structured queries of elements of the experiments. Furthermore, the terms will also enable unambiguous descriptors of how the experiments were performed.

The purpose of NLM's Unified Medical Language System® (**UMLS®**) [6] is to facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health. To that end, NLM produces and distributes the UMLS Knowledge Sources (databases) and associated software tools (programs) for use by system developers in building or enhancing electronic information systems that create, process, retrieve, integrate, and/or aggregate biomedical and health data and information, as well as in informatics research.

3.1 Ontology merging and mapping

Once the main ontologies are identified we need some mechanisms to join them in order to access to diverse data repositories by means of a unique ontology.

Ontology **Merging** allows to know which concepts in an ontology A are the same than in other ontology B. Detecting common concepts and allowing the “jump” between ontologies. The ontology merging could be used by a company that wants to use de facto standard ontologies -GO, MGED,...- associating them to specific company’s ontologies. This way proprietary repositories, as instance repositories with experiments results, can be “linked” with public ones. Also ontology merging allows to merge the repositories described by such ontologies (once repositories are annotated).

Ontology **Mapping** allows to sum up one ontology C with other ontology D to obtain a more complete ontology. Due to the fact that a protein could be implied in cell signalling as in a biological process, summing up two ontologies, one describing cell signalling and other describing biological processes can give us a general overview of a protein function.

3.2 Database annotation

Once we have been able to merge or map the ontologies, as described above, we will have to be capable of linking these resulting ontologies with public or proprietary data repositories through semantic annotation. Deep annotation[7] is a framework to be taken into account at this stage. Deep annotation is a framework to provide semantic annotation of large sets of data. It is used to describe the process that allows to derive mappings between information structures using information proper, information structures and information context.

Annotation is relevant for scientific databases, because scientific databases have been developed with the researchers community in mind, trying to stimulate cooperation. Some of the most known databases annotated are: GenBank, H-Invitational, Invitrogen, InterDom, KEGG and USCSC Genome Bioinformatics. GeneOntology has been used to annotate a great number of different databases, as instance: SGD, FlyBase, GO Annotations@EBI Arabidopsis, GO Annotations@EBI Human, GO Annotations@EBI Mouse, GO Annotations@EBI PDB, GO Annotations@EBI Rat, GO Annotations@EBI UniProt, GO Annotations@EBI Zibrafish,...

4 Conclusions

As semantic web technologies are still developing, also the tools needed to implement such technologies are also in a developmental stage, limiting the application of such technologies to the biological domain. Future advances in semantic web technologies could be applied to solve the immediate Scientifics needs: data aggregation and interoperability, unique entry point for data and processes, agreement in terminology, syntax and semantics related to biological data, semantic data annotation to turn human-understable data into machine-understable data, inference languages to extract and generate knowledge from aggregated data,…

References

1. RZPD Deutsches Ressourcenzentrum für Genomforschung GmbH http://www.bioinfo.de/isb/2002/02/0017/main.html#img-1
2. School of Mathematical Sciences (Israel) http://www.math.tau.ac.il/~rshamir/algmb/00/scribe00/html/lec05/node3.html
3. Infogenmed Project Web Site. http://infogenmed.ieeta.pt/webdata/related-projs.html
4. Gene Ontology http://obo.sourceforge.net/main.html
5. MGED Ontologu http://mged.sourceforge.net/Ontologies.shtml
6. Unified Medical Language System – UMLS (http://www.nlm.nih.gov/research/umls/)
7. On Deep Annotation. S.Handschuh, S.Staab, R. Volz, WWW2002, University of Karlsruhe,2003.
8. PubMed Website, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
9. UpToDate WebSite http://www.uptodate.com/service/index.asp