First Steps towards Semantic Web: Experiences from a Commercial Implementation Project

Riitta Alkula

TietoEnator, Government Services Finland, P.O. Box 403, FIN-02101 Espoo, Finland
riitta.alkula@tietoenator.com

In: Proceedings of Älyä verkossa – Web Intelligence Symposium at the Finnish AI Conference (STeP 2004). Heureka - the Finnish Science Centre, Vantaa, September 2-3, 2004.

Abstract. In the Finnish Museums Online project, TietoEnator built a semantic web portal for the National Board of Antiquities. In a commercial implementation project, the needs and aims differ from those of a research project. A production system must be robust, reliable, extendable, easy to maintain and able to handle large amounts of data. It must also have working interfaces to other systems as well as intuitive user interfaces. The experiences from the project are presented, with special emphasis on defining the core concepts used in a semantic web application.

1 Introduction

For memory organizations, i.e. libraries, museums and archives, managing information is the core business. As the amount of information to be handled constantly increases, more efficient information management methods are needed. Therefore, the promise of semantic web to manage information more intelligently and automatically is very attractive.

Still, for a production system, intelligent functionality is only one and often not the most important need. There are also other requirements the system must fulfill: The system must be robust and capable of recover from minor defects (e.g. ill formed input) and easy to maintain in everyday use. When users' needs change, it should be easy to modify, reprogram and add new functionality. The system must also be able to scale up, if the amount of data increases.

In a research project, different methods can be tested and piloted, to find out the optimal solution. In a production system, normally only proven technologies are accepted, because the customer wants full value for his/her money. Therefore, a commercial project usually does not stand in the frontline of technical development.

In the Suomen museot online project (SMOL; the Finnish Museums Online), TietoEnator built a web portal for the Finnish National Board of Antiquities (NBA) combining proven technologies and a semantic web product with ideas and results from a university semantic web research project. The outcome of the project, named as *suomenmuseotonline.fi*, was launched in May 2004. The experiences and lessons learned from the project are described in the following chapters.

2 Needs and Aims of the Finnish Museums Online Project

In addition to a physical visit, it is more and more common to visit a museum virtually via web. Many museums all around the world have digitized their collections and put this digitized information (at least partly) available in their web site. In Finland, too, quite a many museums already have a web site of their own presenting their collections.

Although this truly increases the availability of the information on the museum collections, the user is still forced to search each museum collection (i.e. web site) separately, if he or she wants to have a comprehensive view of the items available on the whole. In this context, searching via Google is not an option, because it reaches only static web pages, not "deep web" resources, such as web pages created from data accessed from museum databases.

To help getting a wider perspective over single collections, there have been built portals bringing collections of many museums together. The portal can also be a starting point helping in locating the objects and after that directing the user to more specific information (or to a specific collection). Examples from such portals are AMOL, Australian Museums and Galleries Online (http://amol.org.au/) and Kunst Indeks Danmark (http://www.kid.dk/).

Another problem is that building up a web site is expensive and therefore possible only to museums having the requisite funds. Therefore, a service-based solution providing a common publishing channel should be a tempting option for some museums, relieving them from the responsibilities and costs of web site building and maintenance.

The idea of developing a common portal for all Finnish museums was presented already in the end of 20th century [1], although the tender process did not start until year 2003. Then, NBA proposed building a common museum portal or search system to solve the accessibility and resource problems of the museums. The portal is free and open to users (citizens and researchers) to browse and search virtual collections, without need to know which (physical) museum actually possesses the object.

After the tender process, NBA selected TietoEnator as the software integrator and Profium as the semantic web software vendor. Also ideas and ontologies from the *Finnish Museums on the Semantic Web* project of the Helsinki Institute for Information Technology (HIIT) were used [2]. Both TietoEnator and NBA were partnering organizations of that project and therefore have access to its results.

The Suomen museot online (SMOL, Finnish Museums Online) project was started in Autumn 2003 and ended in Spring 2004. The web site is available at address: http://www.suomenmuseotonline.fi.

3 Phases of the SMOL Project

3.1 Preceding Preparations

The SMOL portal is, actually, just an icing on the cake. It is based on the results from the national digitalization project MYYTTI, where Finnish museums have digitalized their cultural heritage since year 1997 [3]. The vision of Digitisation of cultural heritage committee (KULDI) states that by 2010, the foremost cultural heritage in museums, archives and libraries will have been digitized [4].

Even when only the foremost cultural heritage will be digitized, the task is vast and will continue for years. According to the Ministry of Education, in 1997 there were over 22 million museum objects with their context information, and more than 10 million items of photographic or other picture material [5]. Currently, only a very small fraction of the Finnish museum collections (objects and their metadata) have been digitalized. And from the digitalized material, only a minor part has been set available via Internet, because of copyright restrictions, lack of resources, or other reasons [6].

Another issue in the digitalization process is that it includes more than mere information collection and technical conversion. The descriptions produced earlier for a museum's internal and professional researchers' needs must be updated to meet the needs of a wider user population. This requires a lot of manual and intellectual work. The old data processing wisdom "garbage in, garbage out" applies also here – the better the quality of data and metadata, the easier they are to convert to another digital format or system.

The phases of a digitalization project include, among others [6]:

- Actual digitization: Objects (physical museum items and photos) are digitized by using scanners, digital cameras etc.
- Transforming from manual to electronic classification: metadata is stored in electronic collection management systems (databases) instead of manual catalogs.
- Supplementing the metadata descriptions: for example classification codes and key words must be checked and when required, supplemented and corrected by using commonly used thesauri and classification codes – or ontologies, if available.

After these tasks have been performed, the (meta)data of the objects are stored in a database with links to the corresponding digital image files. In Finland, there are many museum collection management systems on the market, mostly developed by Finnish vendors. The most widely used are:

- Musketti, developed for the National Board of Antiquities by EDS Finland
- Antikvaria, developed for the Finnish Museums Association by TietoEnator
- Siiri, developed for the Museums of Tampere by Profium
- MediaVu by Grafimedia
- Image by WM-data Novo

3.2 The Foundation Stone: Definition of the Core Concepts

As digitized material comes from various museums, common understanding of the basic metadata elements is needed, to enable uniform searching from originally heterogeneous data collections.

In practice, it means identifying, defining and getting agreement on the most important concepts and their meanings between the participating organizations, when each organization uses its own vocabulary and gives different meanings to seemingly similar terms, or names a specific phenomenon or object differently.

To enable this, we have in TietoEnator developed a method for recognizing and defining core concepts/metadata elements. The method was originally created in a pilot project developing web service interfaces for Finnish base registers, where each government authority used a vocabulary of its own. For example, the definitions of concept *address* varied in different registers. In that project, the aim was to enable finding appropriate web services and integrating them intelligently with help of metadata. To create a unified vocabulary between all participants, TietoEnator consultants used the method to find out the most important concepts, to label and specify their meaning, and to define the range of their accepted values [7].

The general idea of the method is to start from a group of data models or schemas and first reach a consensus between them. The most important concepts found are nominated as core concepts (or core elements). Then the next data model is added and compared with the core concept set defined earlier. In principle, the initiative set of data models should be as large as possible – in practice, however, it is advisable to start with a relative small group, in order to have better control over the definition process.

The process is iterative and incremental, i.e. new data models will be added one by one. The aim of this “start small, expand gradually” method is to get as soon as possible a working set of concepts, which then can be piloted in a semantic web application, and which also can be flexibly extended.

Our method recognizes the fact that metadata descriptions are not likely to be developed from scratch, which would be very expensive and time consuming, but by utilizing already existing data models or schemas. The same applies also to ontologies, which are likely to be developed from existing vocabularies like term lists, taxonomies, thesauri etc.

Another issue in using already existing data models or vocabularies is that they generally have been developed for a specific task (e.g. for information retrieval, translation) or for a specific user group (e.g. journalists, researchers), which reflects on the selection of concepts (or terms) and the relations between them. The idea of core concepts is to define a central area that is common for all data models, and preserve all domain-specific, detailed information in the original, more specific data models.

Once the set of core concepts has been defined, each organization maps its own vocabulary (concepts) to and from the core concepts. In this way, the core concepts function as a lingua franca or a bridge between the various data models from participating organizations and information systems.

In the SMOL project, we adapted this vocabulary unifying method to metadata concepts (or fields, or elements) from three Finnish museum collection management systems, namely Musketti, Siiri, and Antikvaria. Dublin Core was used only selectively, because it did not have all the elements regarded necessary in the SMOL portal. When defining the searchable elements or access points, a schema based on Dublin Core Culture & Simple was used as a reference. The schema has been developed for Minerva network coordinating national digitization in European Union [8].

The project group found and defined altogether twenty (20) core concepts. This set of concepts was then described as an RDF schema. Also an input interface for the museums sending their data was defined. This data interface was an XML schema, accompanied with documentation and an XML example file [9].

As a result, three types of concepts (or elements) were specified:

1. *Search fields*. These were access points or field labels used in the search window. Four of them corresponded to DC Culture High Level Elements [8]: *what*, *who*, *where*, and *when*. Two of the access points were specified in the SMOL project: *material*, and *keyword.*
2. *Core concepts*. The basic concepts, with which other concept types are mapped. For example, a search term written in search field *who* will be mapped with the contents of metadata fields labeled as *subject*, *user*, or *creator*. Another example: in Musketti, *Esineen/hankintaerän kontekstitiedot: pääasiallinen käyttöpaikka* ('description of an additional place of usage') is mapped to the core concept *Muu käyttöpaikka* ('other places of usage'). Some of the core concepts are also presented in the search result window (list view), to present an overall view of the retrieved items to the user.
3. *Additional concepts*. These concepts provide more detailed or museum-specific information. In the SMOL portal, they are used, for example, to present more detailed information in object description window (catalog card view).

Each participating museum (in practice, collection management system) will map its element set to the set of SMOL core concepts. (Museum data output modules are described in chapter 3.4.)

We believe this approach to be both appropriate and practical. Now we are neither restricted to the smallest common factor between the different museums nor confused with the variety of their vocabularies. By using core concepts we enable coherent searches and uniform presentation of the basic features; with additional concepts, each museum is able to show its objects in a way that does justice to the richness and variety of the material.

3.3 The System Functionality in General

The core of the SMOL system is based on the following products:

- Semantic content (RDF) management platform – Profium Semantic Information Router (SIR)
- Java Engine – Apache Tomcat
- Database – Microsoft SQL Server

Also some other software products were used (e.g. Log4J logger component from project Jakarta), but they are not listed here.

Profium SIR is a Java-based application. Its services can be accessed through servlets or Java Server Pages (JSP) on a web application server. As we in system development prefer Model-View-Controller (MVC) paradigm, we did not use any JSP coding. For example, in generating the user interface, we used a Java servlet that gets an XML file and the XSL stylesheet as input, and invokes an XSLT transformation that generates the web pages.

As we know from visiting traditional physical museum collections, visual appearance and presentation of museum objects is very important. Consistently, also a virtual museum should have an illustrative and attractive look-and-feel. Therefore, much effort was made to graphic design, which was done by a senior art director. The web page functionality was designed by a technical site builder specialized in XSL transformations.

In storage phase, content objects (i.e. XML files and image files) are fed to the SIR input handler. SIR has many general input handlers, including one for XML files. Because processing of SMOL image files required some extra functionality not included in the out-of-the-box handler module, we programmed a custom museum picture adapter to process SMOL data. In our application, SIR first sends the XML files to the Lucene fulltext system to be indexed. After that, SIR generates metadata (i.e. directed RDF graphs, or RDF triples) from the input, and stores both the content objects and metadata.

In the search phase, a user feeds search terms in the appropriate fields in the search window. From them, an RDF query is generated. The query is basically a dynamic XML file, which is based on query language (RDFQL) developed by Profium. The search terms in the query are matched with the values presented in the RDF triples.

From matching records, a list view is generated. When a user selects an item by clicking its link in the list, a new window opens. This detailed information window presents the core elements as well as museum-specific additional metadata elements with all images linked to that metadata record.

We have aimed to a robust system where new material can be easily added with minimal modifications, when new museums and their collections are added in the system. Once the museum takes care that its identification as well as the identification of the content objects are unique, all the data from various museums can be identified and handled appropriately. This is practical to the NBA, because it does not need to do any additions or modifications to the SMOL portal, if the new input files are formed according to the XML schema.

3.4 Data Interfaces to Museum Collection Management Systems

For museums, it is important that the amount of manual editing and conversion is minimized. Once the material is available in computer-readable form, the data should, in principle, be collected and converted into any format with a single keystroke. The same applies to maintenance: one should maintain the data only in one system, from which updated data is sent to other systems, instead of maintaining the same data in many different systems.

NBA decided that actual data handling and corrections should be done only in the museum collection management systems for the following reasons:

- The museums do not have to learn new systems, but can use collection management software they are already familiar with.
- The museums have to maintain only one system: if the data in SMOL is not valid, the corrections are done only in the museum collection management system and corrected data is then re-sent to SMOL.
- SMOL is only a publishing or presentation system, not a mission-critical system of any single museum; if SMOL for some reason would break down, the original data is in safe in the collection management system and can be reproduced from there.

Technically, there is not anything advanced in our interface modules; it is just a matter of traditional programming. For the acceptability of the SMOL concept, however, these modules are of major importance: the more smoothly the data collection and transformation process goes, the easier it is to a single museum to join to the SMOL portal.

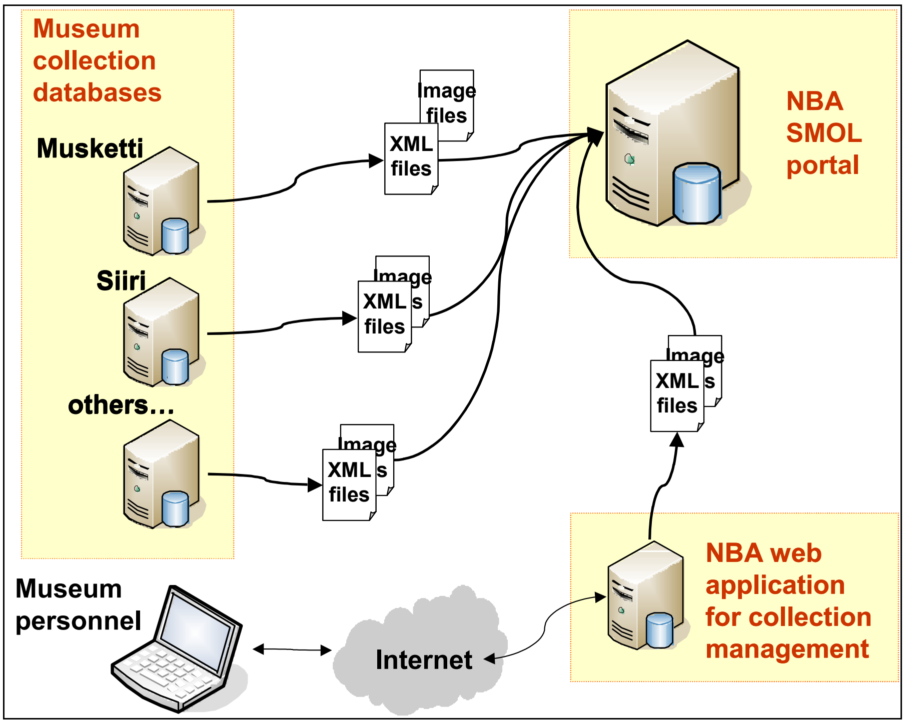

Fig. 1. The process of producing XML files (metadata) and associated image files for SMOL portal from various collection management systems.

Two output modules have already been developed: one for Musketti by TietoEnator and another for Siiri by Profium. Other modules for major Finnish collection management systems are under consideration.

In both Musketti and Siiri, the concepts – in practice fields - were analyzed and mapped with the SMOL core concepts and their accepted range of values. The mapping and transformation rules are programmed into the output modules, which search and collect the metadata, make the transformations between the value domains and attach the specified XML tags around the data (Figure 1).

For example, in Musketti, *valmistusaika* ('time of production/preparation') of a museum object may be expressed as exact dates, or as years (even centuries). In SMOL, we present time information only in years. When time data in Musketti is presented as exact dates, transformation rules guide the output module to select only the value of the year from the date field.

The selection and transformation process is automatic, so the museum personnel do not need to know anything about internal technical procedures. They just need to decide, if digitized information on a specific object should be sent to SMOL or not.

For museums not using major Finnish collection management systems, NBA provides a web-based collection management application, which is offered to the museums as a service. The museums can feed their metadata and respective image files via web pages to a database, from which the SMOL XML output files are generated.

3.6 The Ontology

In the SMOL project, we did not create an ontology of our own, but used one produced in the HIIT project *Finnish Museums on the Semantic Web* [2], namely MAO ontology [10]. It has been developed on the basis of MASA, the Finnish thesaurus for museums [11].

To enable the users to browse the ontology, we programmed a simple ontology viewer. It uses the functionality of a SIR module called SchemaViewer, via the API of that module. The user interface of the ontology viewer is generated via XSLT transformation - in the similar manner as other SMOL web pages. With this functionality, the user is able to browse the ontology hierarchy up and down, seeing one level of the hierarchy at a time.

Currently, the user invokes the ontology browser by pressing a link in the search window. A new window opens, in which the user can browse the ontology. When the user finds an appropriate keyword (term), he or she marks the keyword and clicks *"Add"* button. The ontology browser window closes and the selected term is automatically inserted in the keyword search field in the search window.

4 Experiences and Problems

In general, we feel that the common 20/80 per cent rule applied also in this project. Our experienced developer programmed the basic functionality (80 per cent) very fast, but fine-tuning of the last features (20 per cent) took more time than anticipated.

As in all new technology projects, immaturity of software and scarce documentation causes unexpected problems. There are not many similar applications and, therefore, there is not available such help from programmer community and newsgroups as there are with more mature products. When we encountered problems, finding out the reasons for them took time. For example, testing with extensive systematic test material revealed some problems with Lucene free text indexing, and both TietoEnator and Profium spent a lot of time in locating the reason for that.

Testing with original material from NBA also revealed that different persons feed data in Musketti collection management system differently, thus causing some inconsistency in the material. Names of the image files could contain spaces or other special characters, which caused problems when forming URL addresses based on image file names.

Another problem was that some objects did not have a name, therefore missing also the link to the detailed information window of that object. We also decided to ignore references to missing image files in the input phase, if an XML file in other respects was correct. Otherwise we would have had to reject quite a many input files, making the browsing of error and exception logs tedious. This kind of problems will vanish, as the quality of input material stabilizes after updating the data in museum collection management systems.

When developing XML output modules, the problems were not so much involved with the RDF schema, but with many-sided and complicated Musketti database schema. It had a lot of normalizations, making it difficult to find out, from which database field each piece of information should be selected.

A seemingly minor, but irritating problem was the character encoding of the XML files. Some twenty years ago, we tackled with ASCII encoding, when transferring text from one computer system to another, because different operating systems and programs coded Scandinavian alphabet (*å*, *ä*, or *ö*) differently. Nowadays, ISO Latin 1 (ISO 8859-1) is universally accepted and Scandinavian characters are, normally, handled properly. But NBA has museum objects also from other countries, for example from Morocco and Siberia. The names of these objects cannot be properly expressed in plain Latin character set. In principle, Unicode (UTF-8) encoding should solve this problem. In practice, we noticed that changing some parameters in web application servers and other programs from default ISO Latin 1 to UTF-8 caused unexpected problems with system response times and system stability. Therefore, we decided to abandon Unicode for the time being and keep to ISO Latin 1 encoding.

The same problem applied also to the Protégé ontology editor [12], although conversely. The MAO ontology we received in December 2003 was encoded in ISO Latin 1, whereas the default setting of the Protégé editor was UTF-8. Because ISO Latin 1 and Unicode code Skandinavian alphabet differently, characters *å*, *ä*, and *ö* were displayed incorrectly in Protégé window. Another problem with Protégé is that there seems to be some inconstancy between the versions. For example, an ontology developed by using Protégé version 1.8 did not open in Protégé version 1.9, but opened well in Protégé version 2.1.

5 Futher Development

In the near future, an obvious aim is to increase the amount of data in the SMOL system. As the system has been designed as flexible, adding new museums and collections should not be any problem. What we need is output modules to other collection management systems, in addition to current Musketti and Siiri output modules. It is also possible that the data elements (namely museum–specific additional concepts) collected from Musketti and Siiri will increase, making the content even richer it is now. That means modifications only to the output modules, not to the SMOL portal itself.

Visual layout and functionality of the ontology browser should be improved. As already discussed in the chapter 3.6, we developed a simple ontology browser just to view the MAO ontology. The user is able to pick up only one search term at a time. It would be interesting to have a tool, which would automatically enhance the search to all subcategories and/or related terms of the selected term. For example, when a user picks the term *vehicle*, subcategories like *car*, *boat*, *airplane*, etc. would be added automatically to the query. This feature is currently investigated in the subproject *Ontology-based query interface* of the HIIT research project *National ontology project in Finland* [13]. If the results are promising, we will consider, how this feature could be implemented in SMOL, too.

References

1. Oppimisen, luovuuden ja osaamisen Suomi II. Opetusministeriö, Helsinki (2000). Opetuksen, tutkimuksen ja kulttuurin tietoyhteiskuntaneuvottelukunnan loppuraportti. ISBN 952-442-200-X (The Final Report of the Council for Promotion of Information Society through Education, Research and Culture 1997-2000.) URL: http://www.minedu.fi/julkaisut/pdf/Oppimisen%20Suomi.pdf
2. Finnish Museums on the Semantic Web. (Home page of the project.) URL: http://www.cs.helsinki.fi/group/seco/museums/.
3. Myytti, museot yhdistävät ja yhtenäistävät tietojaan. Suomen Museoliitto (Finnish Museums Association). URL: http://www.museoliitto.fi/projektit/myytti.htm
4. Kulttuuriperintö tietoyhteiskunnassa: Strategiset tavoitteet ja toimenpide-ehdotukset. Opetusministeriön julkaisuja 2003:24. Opetusministeriö, Helsinki (2003). ISBN 952-442-514-9. (Cultural Heritage in Knowledge Society. Final report of the Digitisation of cultural heritage committee, KULDI). URL: http://www.minedu.fi/julkaisut/kulttuuri/2003/opm24/opm24.pdf
5. Opetusministeriön tietostrategioiden tilanne. Opetusministeriön tietostrategioiden työryhmä. Opetusministeriön työryhmämuistiot 1997:26. Opetusministeriö, Helsinki (1997). URL: http://www.minedu.fi/julkaisut/julkaisusarjat/tietostr.html#3.
6. Hongisto, Vesa. Suomen museot online. Helsinki (2004). A presentation held in publishing the MuseoSuomi and Suomen museot online web sites on March 8, 2004 at the National Museum of Finland.
7. Rekisteripoolin sanastotyö: Sanastojen yhtenäistämisprosessi. Rekisteripooli, Helsinki (2004). URL: http://www.rekisteripooli.fi
8. 24 Hour Museum Metasearch project: schemas. System Simulation (2003). URL: http://www.minervaeurope.org/DC.Culture/XMLSchema/1.0/MetasearchSchema.pdf

9. Suomen museot online: Sanastokuvaus. URL: http://www.nba.fi/tiedostot/e8c3398d.pdf
10. Hyvönen, E. et al. A Content Creation Process for the Semantic Web. Proceedings of OntoLex 2004, Ontologies and Lexical Resources in Distributed Environments. A workshop of the 4th International Conference on Language Resources and Evaluation, LREC 2004. Lisbon, Portugal (2004). URL: http://www.cs.helsinki.fi/u/eahyvone/publications/contentCreation.pdf
11. Leskinen R. (ed.) MASA, Museoalan asiasanasto. National Board of Antiquities, Helsinki (1997). URL: http://www.nba.fi/fi/masaetusivu
12. The Protégé Ontology Editor and Knowledge Acquisition System. URL: http://protege.stanford.edu
13. National Ontology Project in Finland. (Home page of the project.) URL: http://www.cs.helsinki.fi/group/seco/ontologies/