Cultural Heritage on the Semantic Web – the Museum24 project

Barnabás Szász[1], Arto Saraniva[2], Katalin Bognár[1], Michal Unzeitig[2], Mikael Karjalainen[2]

[1] University of Debrecen, Department of Computer Science,
H-4000 Debrecen, Hungary
bszasz@gmail.com
bognar@inf.unideb.hu
[2] Artio Ltd.,
FI-42100 Jämsä, Finland
{first_name.last_name}@artio.net
http://www.artio.net/

Abstract. Describing and storing multimedia documents in a way, which would make them easily searchable, retrievable and exchangeable has been a long discussed problem. Most users of the web do not have mature searching strategies. Solutions are needed to improve the performance of full-text retrieval systems to find relevant information for ordinary users. This paper presents a semantic virtual museum publishing diverse cultural heritage material on the web. The system was developed under the project Museum24 – Virtual Museum of Jämsä Region (Museo24- Jämsän seudun virtuaalimuseo, www.museo24.fi). It shows how semantically rich and interrelated content can create a consistent semantic portal. Further, it introduces overall concept on how multimedia documents are being annotated with semi-automated methods and how the metainformation is managed and stored in a CIDOC-MPEG7 based ontology. It also discusses how the public and administrative interfaces are realising better user experience to provide the visitors useful search and navigation services and the maintenance personal quick and efficient workplace.

1 Introduction

The main goal of the Musem24 project is to improve the accessibility of the cultural heritage in the Jämsä region, which includes two small towns Jämsä and Jämsänkoski in Middle Finnland. The virtual museum covers not only the local museums, but mostly a large variety of the heritage beyond the traditional institutions also, it tells the story of Jämsä Region in nutshell. The project started in 2004 and it will be finished in the end of 2006. It was initiated by the Department of Culture of Jämsä and the local heritage societies and funded by local authorities and companies and by ERDF (State Provincial Office of Western Finland). The company Artio Oy was chosen as the main technology partner of this project.

A trend could be observed in the actual scenario of transition of cultural heritage in the cyberspace. It may be worth reminding that museums, libraries and archives of the real world are the result of a process which began many centuries ago. Museums started out in 16th century Europe as promiscuous bodies where art objects, artifacts, natural items, books and documents were integrated and displayed alongside one another. The evolution of museums might be very schematically regarded as a story of ever stronger specialization. There is no reason to store a digital entity according to the same systems used to preserve the object it emulates. In the digital world there are no more museum buildings or rooms, and we are not obliged to reproduce the same structure of materials. [1]
This paper presents the semantic virtual museum – Museum24 – for publishing cultural heritage materials on the web. It is organized as follows: after reviewing the present state of the field, it describes the general architecture of the underlying system, and its components. Next it introduces overall concept on how multimedia documents are being annotated with semi-automated methods and how the metainformation is managed and stored in a CIDOC-MPEG7 based ontology. Further it discusses how the user interfaces are realising better user experience to provide the visitors useful search and navigation services. The final part meditates about further development of the system.

2 State of the Art

One of the fields of applications which will benefit from the recent advances in Semantic Web Technologies is the area of Cultural Heritage Content Management. This field involves the development of applications for the efficient processing, storage, retrieval and exploitation of cultural heritage materials. Museum collections in many cases include a large set of multimedia content with rich metadata.
Several systems which provide access to cultural heritage collections already exist, such as Sculpteur[1] or MuseumFinland[2] which exemplifies how heterogeneous cultural collections from different organisations can be made semantically interoperable by making use of Semantic Web technologies. Sculpteur introduces simplified views of the ontology, appropriate for dealing with specific query types: who, what, when, where and how type queries can be launched in both systems.
These portals are meant for utilizing access and using ontology for describing museum or other collections. To store and manage metadata about ordinary content like stories or other multimedia documents is still an open issue. In such a system the concepts are in the background organized into ontologies and stories about the concepts that are in the focus.
Maintaining such a system needs different tools: content creator tool, annotator tool and ontology editor tool. On the public pages several services could be introduced based on the underlying ontology such as searching and browsing services based on semantic clustering like related content in time (semantic timeline) or place (maps).

[1] Project homepage: http://www.sculpteurweb.org
[2] Accessible via http://www.museosuomi.fi

3 System Architecture

The Museum24-publishing system itself is a Multimedia Content Management System with ontology based metadata store. Next to the presented ontology manipulation and annotation functionality, the Museum24 has all the advantage of the popular CMS's: user role based content accessing, editing and publishing, easy to use WYSIWYG XHTML editor with built in image and link browser, advanced administrator interface, inner bulletin board for editor messaging. All functions can be reached and utilized through a web based user interface, without installing any application on the client side.

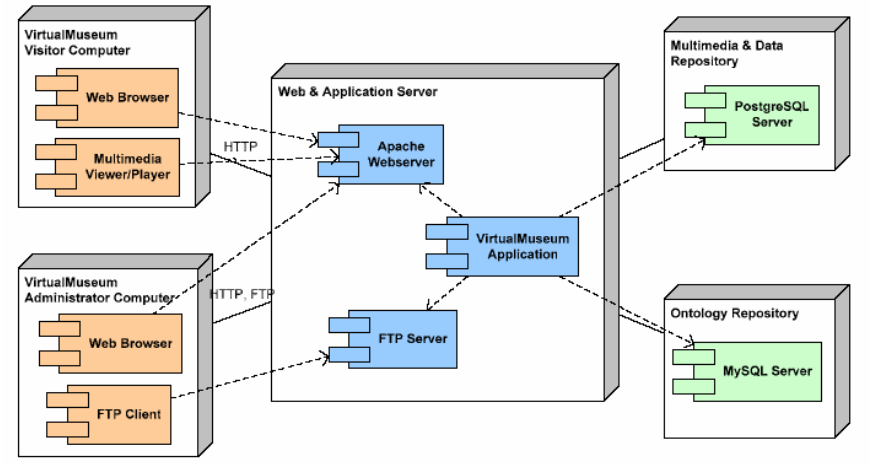

Fig. 1. The system architecture

The concept of the system is typical tree-tier architecture. The diagram in Figure 1 shows the deployment of the first version of the Museum24 system. The server part of the system consists of three nodes (servers) – Web & Application Server, Multimedia & Data Repository and Ontology Repository. The client part is formed by nodes representing museum administrators' workstations and personal computers of the virtual museum visitors. The system was made independent on the underlying operating system. Every node may even be run on a different operating system. For the Application server, the only limitation is availability of the third-party tools necessary for the application. However, all the most popular systems such as Linux clones or other UNIX-based systems (Solaris, MacOS, HP-UX, IBM AIX or FreeBSD) as well as Windows are supported.

4 Ontology Administration Issues

Content annotators are widely used technologies, but most of them use only textual labels, simple words or phrases, or so called tags to describe the meaning. While tagging is quite popular nowadays, the services like Flickr[3] or del.icio.us[4] are mushrooms starting up, still the ontology editors have much too complicated UI's and methods for a broader usage. However, folksonomies which are core elements in these services have weaknesses aroused from uncontrolled vocabularies. [2]
In Museum24 we combined the simplicity of these tagging services and the power of underlying ontology. The annotation is done by referring to ontology individuals which are created on demand. The annotators are domain professionals which make the result even better then the quality of average folksonomies. Since the annotators' task is not to classify the content but gather Named Entities from the document which is a more obvious task, the result is also clear and independent from personal term-usage. We have developed tools to support finding the correct term for defining new individuals.

4.1 Named Entities

In Museum24 the content are the multimedia documents which include: articles, images and other media files. These materials are related to Named Entities which are nodes in the ontology. We call named entities or concepts the representative values of real world actors, objects, events and places. Those concepts are interconnected in the ontology as well as linked to the multimedia materials also. The Named Entities annotate the multimedia documents, which also have representative entities in the ontology, allowing the annotation to be stored in the ontology as well.

4.2 Collaborative Ontology Building

The CIDOC object-oriented Conceptual Reference Model (CRM) [5] – which was chosen as metadata structure in Museum24 – represents ontology for cultural heritage information. It was developed by the ICOM/CIDOC Documentation Standards Group. To allow storing information about the content and format of multimedia documents, the base structure of the ontology was extended with the MPEG-7[6] class hierarchy suggested by Jane Hunter. [3]
Since the applied ontology of the system consists of 95 classes but no predefined instances, populating the ontology is the task of the content creators. Because none of the authors are information architects, a multi level collaborative ontology building process was introduced. [5] In the next stage period of the Museo24 project different

[3] Flickr homepage: http://www.flickr.com
[4] Del.icio.us homepage: http://del.icio.us
[5] CIDOC homepage: http://cidoc.ics.forth.gr
[6] MPEG-7 overview: http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

user roles will also be introduced because some tasks of the process need ontology expert.

After realizing that the common steps of ontology maintaining in other systems like Protégé are too complicated for our content creators, a simple ontology based annotation tool was developed, leaving the annotation and ontology maintaining process separate. Most of the difficulties came from the complexity of the CIDOC CRM class hierarchy. The class structure was examined and four properties were kept in the process to be able to describe the content in different aspects as shown in Table 1. All four properties are mapped to the same CIDOC property "*is about*".

Table 1. The four type of annotation properties

Property	Aspect
About whom?	describing persons, groups, legal bodies
About what?	describing different kinds of physical and conceptual stuffs
About when?	describing a time period
About where?	describing a place

The number of the allowed operations was reduced, leaving the annotation process consisting of the following three steps:

Named Entity recognition

In the annotation process the most important step is collecting all the concepts which are related, which are describing the content of the media document. It means that this step must include a strong searching and text parsing method.

Named Entity classification

The second step needs a domain expert because sometimes it is not obvious a concept which class of the ontology belongs to. New NE's created during the annotation process have no ontology class, which decides their properties through the selected class. The expert during this task sees a list with non classified items and is able to attach to a class from tree-like list.

Named Entity description

Finally, the concepts need to be interlinked in the ontology in order to be able to use reasoning on them; this is done by setting up properties. This task needs domain experts and is separated from the annotation step. Creating links is a two step method: First, the expert has a list of non described concepts and uses either existing individuals to define value for properties, or creates a new one by defining label and choosing class. To decide if such named individual already exists in the ontology, an AJAX[7] based LiveSearch tool was developed, which immediately checks the typed name against the ontology and suggests existing concepts. The result can be seen within a second, and the expert can select one existing individual from the list. This

[7] More about AJAX: http://www.developer.com/design/article.php/3526681

method can help to avoid the problem of individuals with multiple different names. The second step in the link creation process is setting up the relation type. The class of subject and object limits the class of properties. At times it is obvious and sometimes just a few options exist.

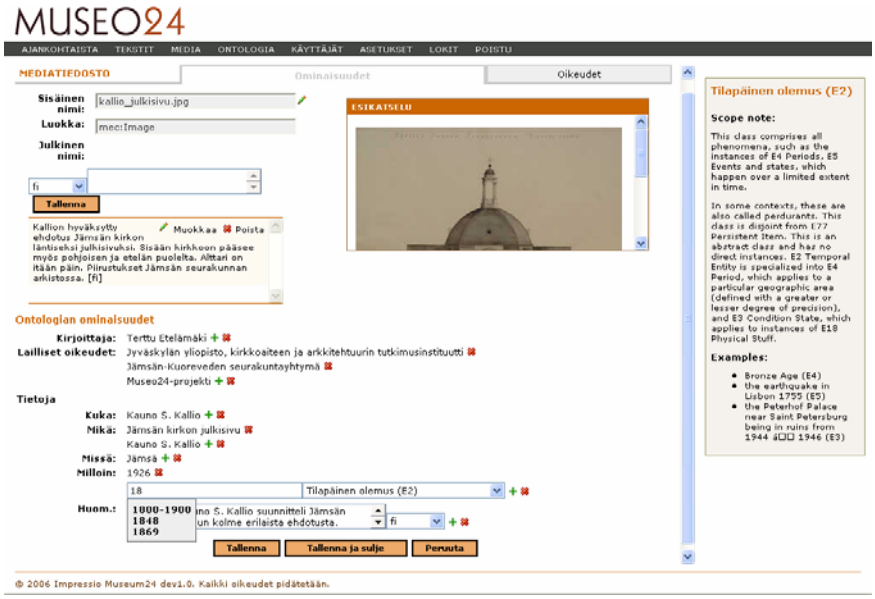

Fig. 2. The media annotator tool with activated LiveSearch result box

4.3 Semi-Automatic Named Entity Parsing

Annotation in Museum24 means creating relations between named entities and multimedia documents. The first step of the annotation process is collecting related NE's from the document; the second is to set the relation type. The CIDOC-MPEG7 ontology limits the relation types; the most time consuming part is the NE recognition.

To overcome the limitations of manual annotation, a semi-automatic system was developed. Semi-automatic systems, as opposed to completely automatic systems, are used because it is not yet possible to automatically identify and classify all named entities within textual documents with complete accuracy. Building a completely automatic annotation system is an open research problem. Another open question is the automatic annotation of non-textual documents.

Even though the Named Entity Recognition systems are getting matured, our approach for a semi-automatic annotation engine is a rather simple solution. After

submitting a textual document, a parser tries to find NE's in the document and offers the result as annotation. We except that all NE occurs at least once with full name and in that way can be recognised. The document editor has the opportunity to accept, remove, and extend the result. Since in the Finnish language the word suffixes could change the stems, this method is not completely accurate and needs manual work.

4.4 Embedded Metadata Extraction

Current efforts in multimedia format standardizations have recently provided functionality to embed image metadata into actual image files. For example, the JPEG file format provides support for embedding a standard set of descriptors in the file header, defining metadata elements including file size, width/height, pixel density, etc. Additionally, there are extensions to this element set, such as the Exchangeable Image File Format (EXIF[8]), which includes camera specific information (camera make, model, orientation, etc.). MP3 audio files have also meta-information – so called ID3 tags – about author, title, play length, bitrate, etc. Our approach takes advantage of such existing metadata by extracting and importing this information into the ontology via MPEG-7 descriptors.

5 Ontology Based Information Retrieval

Since the four categories of content descriptors were introduced, the next step is to use them as facets on the public pages as described in [4]. A facet is a kind of view of the materials, which allows grouping them in a multidimensional way. The idea is to organize the concepts and individuals of the underlying ontology into orthogonal category hierarchies called facets. For the time being we are using one of them for building the Semantic Timeline.

5.1 Public Search

The act of searching can be distinguished between two forms: an item known to exist may be searched with the intent to locate it, and an item whose existence is uncertain may be searched, in order to ascertain whether it exists or not. This second form of information consuming is often called "browsing".
In the first case, actors are able to describe the object they are looking for; the only issue is a common language with the search system itself. Using different descriptors might foil to find relevant objects. Introducing taxonomies or synonym dictionaries helps find the terms of the common language.
If the actor has no clear idea about which object(s) to locate, but has some thoughts about the order to select categories the only issue is to select the filter options. Too many options or too deep option-hierarchy might confuse the user, whereas too less

[8] Unofficial EXIF Homepage: http://www.exif.org/

might foil to select one. By browsing, the task is to collect all the concepts which have some interest, which results in a set of related stories, images, videos, etc.

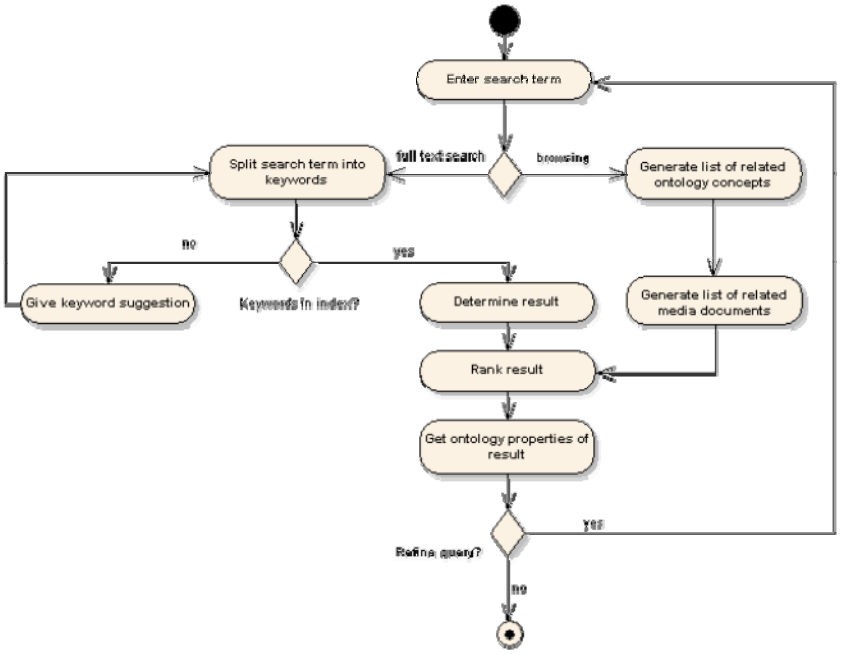

Fig. 3. The search process

Between the two forms of searching no clear border exists, hence the reason we have introduced the hybrid model in the Museum24 as depicted on Figure 3. Normally the search is done by matching keywords. By operators keywords could be included, included optionally, excluded, quoted and stemmed (manually, later with stemming algorithms). An algorithm checks the words to determine whether they were mistyped and offers correct ones if needed. All textual materials (articles, notes, annotations, ontology individuals) are indexed and included in the search process. The result is ranked by word importance which is calculated by position, font, and role of the word. Because the words in ontology concept labels have higher score, the annotated items are on the top of the ranking. The most common ontology properties are collected in a hierarchical tree form in order to offer browsing and filtering possibilities. Selecting one of them extends the search query with the ontological category. This on-the-fly-built category tree contains fewer elements then the whole ontology hierarchy.

5.2 Related Content – Semantic Folders

The content of the Museo24 system is organized into a hierarchical folder structure. A folder and each element in a folder could be in two different states: published (when it is visible for the visitors) and unpublished.
The content of the folders could be defined in an explicit way: copying it there, creating it there, or implicitly setting a filter rule. A filter rule is a combination of properties. Properties could be combined by conjunction and disjunction. To define a property the user should first choose the type from the hierarchical type list. The user should than choose the value from the possible values list which depends on the previously selected property type. All the elements fulfilling the filter will be displayed in the folder implicitly. A semantic folder can be seen as a stored semantic search query.

5.3 Semantic Timeline

In the Semantic Timeline we are using the 'About when?' properties to sort and group the materials in a time order. The timeline is divided in ages. Each age is described by a professional and the visitor has a basic idea about the happenings internationally and locally the same time. Opening the ontology viewer she/he will find the list of articles and other multimedia documents which are related to this time period.
Some manual work is needed in the ontology editor to define which time period entity covers the other. Using predefined time periods by experts could reduce the amount of work needed to keep the ontology up to date.

6 Conclusion and Further Work

In this paper we have shown our comprehensive approach for building semantic portals focusing on three issues. First, we examined ontology population with collaborative methods, the reason we introduced it, and the problems we have faced during the introduction period. Next, a semi-automatic media annotator tool was introduced. Last, we discussed the benefits of background ontology for portal visitors.
All the presented functionality is big improvement compared to ordinary Content Management Systems. Here, we have made the argument that there are many big open issues that have hardly been dealt with so far. Automatic and collaborate ontology building processes are inaccurate and allow anomalies in the ontology, such as duplicated concepts, mistyped names, wrong property values, etc. Since the ontology population is still in early phase, we could discuss the accuracy and correctness of the ontology at a later date. Future questions include how to handle these issues and what tools we need in order to maintain efficiently the underlying ontology and how to make the "feeding" process as simple as possible, "keep it simple…" as ice-hockey coaches keep on saying.
In further work, we plan to add rule based reasoning in the ontology based search process to enhance it's functionality by utilizing additional relationships to the "*is*

about" property between concepts and media documents. Introducing new tools based on the ontology facets like Semantic Map is also part of our aim.
At present, the Museum24 portal is accessible via: http://www.museo24.fi, only the titles in English so far.

References

1. Galluzzi, P., The hybrid digital Library, Berlino, 2003. http://www.zim.mpg.de/openaccess-berlin/Paolo_Galluzzi_211003.pdf
2. Mathes, A.: Folksonomies - Cooperative Classification and Communication Through Shared Metadata, Computer Mediated Communication - LIS590CMC, Graduate School of Library and Information Science University of Illinois Urbana-Champaign, December 2004 http://blog.namics.com/2005/Folksonomies_Cooperative _Classification.pdf
3. Hunter J., Combining the CIDOC CRM and MPEG-7 to Describe Multimedia in Museums, DSTC Pty Ltd., University of Queensland, Australia, 2002. http://metadata.net/harmony/MW2002_paper.pdf
4. Hyvönen, E., Junnila, M., Kettula, S., Mäkelä, E., Saarela, S., Salminen, M., Syreeni, A., Valo, A., Viljanen, K.: Publishing Museum Collections on the Semantic Web - the MuseumFinland Portal. Proceedings of WWW2004, New York, Alternate Track Papers and Posters, USA, 2004. http://www.cs.helsinki.fi/u/eahyvone/publications/ www2004/museumFinland_poster.pdf
5. Korpilahti, T., Hyvönen, E.,: An Architecture for Collaborative Ontology Library Development. Proceedings of 16th European Conference on Artificial Intelligence (ECAI2004), Workshop on Application of Semantic Web Technologies to Web Communities, 2004. http://www.seco.tkk.fi/publications/2004/korpilahti-hyvonen-an-architecture-for-collaborative-2004.pdf
6. Addis et al., New Ways to Search, Navigate and Use Multimedia Museum Collections over the Web, in J. Trant and D. Bearman (eds.). Museums and the Web 2005: Proceedings, Toronto: Archives & Museum Informatics, published March 31, 2005 http://www.archimuse.com/mw2005/papers/addis/addis.html
7. Uschold, M. and Jasper, R. A Framework for Understanding and Classifying Ontology Applications. In Proceedings of the IJCAI-99 Workshop on Ontologies and Problem Solving Methods (KRR5). Stockholm, Sweden, August 1999.