Museum Collections and the Semantic Web

Maria Nisheva-Pavlova[1,2], Nicolas Spyratos[3], Peter Stanchev[2,4]

[1] Faculty of Mathematics and Informatics, Sofia University, Bulgaria
[2] Institute of Mathematics and Informatics, BAS, Bulgaria
[3] Laboratoire de Recherche en Informatique, UMR 8623 du CNRS, Université Paris-Sud, France
[4] Kettering University, Flint, USA
marian@fmi.uni-sofia.bg, Nicolas.Spyratos@lri.fr, pstanche@kettering.edu

Abstract. The paper discusses some current trends in the area of development and use of semantic portals for accessing heterogeneous museum collections on the Semantic Web. The presentation is focused on some issues concerning metadata standards for museums, museum collections ontologies and semantic search engines. A number of design considerations and recommendations are formulated.

Keywords: Cultural Heritage, Digitization, Semantic Web, Semantic Interoperability, Metadata Standards.

1 Introduction

During the last decades information technologies play a considerable role in lots of successful projects directed to digital preservation of cultural and scientific heritage. The growth of the number of digitized heritage collections increases the necessity of proper software tools assisting the access to these collections and making the best use of them.

An essential characteristic of cultural collection contents is its semantic rich-ness. Collection items have their history and are related to the society, and to other collection items. The collection semantic network is not limited to a single collection but spans over other related collections in other museums. The network of semantic associations can be extended to contents of other types in other organizations, as well. It is advisable to publish digitized cultural heritage collections using semantic portals. Such portals typically provide the end-user with two basic services: (1) a search engine based on the semantics of the content and (2) dynamic linking between pages based on the semantic relations in the underlying knowledge base. Semantic Web technology enables new possibilities when publishing museum collections on the Web [6] such as collection interoperability in content (web languages, standards, and ontologies make it possible to manage heterogeneous collections of different kinds mutually interoperable) and intelligent applications development (more versatile, user-friendly, and useful applications based on the semantics of the collections).

There is still a lack of suitable ontologies for museum collections. Languages such as OWL[1] and ontology editors like Protégé[2] enable the rapid development of ontologies but lots of questions concerning multilingual capabilities and processing of synonyms are still open. We need to consider [4]:

- How to prioritize the ontologies? In particular, which ones should the heritage collection develop and which ones will we be able to borrow from other areas?
- What heritage-based organizations should focus on ontology creation?
- Ontologies often fail to be interoperable. How can we make it to work effectively?
- How do we know what our agent has discovered through its search on the Semantic Web can be trusted? This is especially important when two ontologies may conflict with one another.

2 Semantic Web and Semantic Technologies

The concept of Semantic Web was introduced "not as a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation" [1]. The Semantic Web is considered by some authors as the abstract representation of data on the WWW, based on a set of standards. It is being developed by the W3C[3], in collaboration with a large number of researchers and industrial partners. The main idea was to help the Web to become a truly machine-readable resource and to make the information it contains structured in a logical, comprehensible and transparent fashion [5].

The most significant standards supporting the Semantic Web goals include [2]: Uniform Resource Identifiers (URIs), the Extensible Markup Language (XML), and the Resource Description Framework (RDF) family of standards. The development and maintenance of proper ontologies play a crucial role in the implementation of software for the Semantic Web. According to [3], "ontology is a specification of a conceptualization". Ontologies provide a controlled vocabulary for their domain. This can provide great impact, since users, authors, databases and computer programs can all use terms from the same vocabulary. Ontologies have been used to provide intelligent search support. They can be utilized as a source for semantics-based expansion of the user queries and radically help improving the search results.

"Semantic technologies" is a general term for any software that involves some kind and level of understanding the meaning of the information. For the digitization of cultural heritage collections, most important are the technologies for automatic metadata generation, intelligent search, and multimedia retrieval.

---

1 http://www.w3.org/TR/owl-features/
2 http://protege.stanford.edu/
3 http://www.w3.org/

3 Semantic Web Portals for Museum Collections and Europeana

Several successful projects that utilize Semantic Web technologies to provide intelligent access to cultural heritage collections already exist. Among the most popular are the projects REACH and MuseumFinland as well as the Amsterdam Museum Linked Open Data set.

The objective of the REACH project [11] is to develop ontology-based representation in order to provide enhanced unified access to heterogeneous distributed cultural heritage digital databases. The complete system is composed of the following subsystems: (1) a cultural heritage web portal for unified access to the information and services, (2) a digitization system for the efficient digitization of artwork and collections, (3) a core ontology to describe and organize cultural heritage content, (4) a multimedia content-based as well as ontology-based search engine to offer advanced choices of searching methods, and (5) an e-commerce section for the commercial exploitation of the portal.

The purpose of the core ontology is to provide a global model able to integrate metadata originating from different sources. The integration process involves efficient mapping of the available metadata to the concepts and relations of the core ontology. Only one knowledge base has to be used for the development of cross-domain tools and services. While the area of cultural heritage combines very heterogeneous sources of information and material, one of the requirements of the project was that the ontology to be used should be as extensible as possible. In order to meet this requirement, the CIDOC-CRM ontology (see Section 4) was used.

The web portal provides advanced searching capabilities to the users. The users are able to use a variety of searching functionalities such as: ontology-based search, content-based visual search and a novel hybrid ontology-visual search.

MuseumFinland [6] is an ambitious attempt to generate a complete Semantic Web portal bringing more than 15 museum collections together. The corresponding software system transforms collection databases into a virtual Semantic Web space. Its pages are linked with semantic links that are useful for finding information based on its content. It offers to the user a semantic browsing and searching facility in the combined collection knowledge base. This facility is implemented by server-side software, called Ontogator. When the user views the exhibition entry page with a web browser, Ontogator dynamically generates WWW pages with links to other pages of interest. MuseumFinland uses seven domain ontologies: the Artifacts (Object Types) ontology, the Materials ontology, the Actors ontology, the Situations ontology, the Locations ontology, the Times ontology, and the Collections ontology. All taxonomy classes are instances of metaclasses for which properties such as the creator, description, date of creation, etc. can be specified.

Ontogator provides the user with two semantics-based facilities:

- *View-based search engine*. Ontogator shows multiple ontologies used in annotating collection data. By selecting ontological classes from these hierarchies, the user

can define search queries in terms of ontology concepts instead of simple keywords.
- *Semantic recommendation system*. It enables the user to find out explicit and implicit semantic associations within the global collection data and to use these associations for browsing the collection.

The Amsterdam Museum Linked Open Data set [7] is a five-star Linked Data[4] representation of the entire collection of Amsterdam Museum consisting of more than 70 000 object descriptions. The Amsterdam Museum uses a digital data management system to manage their collection metadata and authority files. As part of the museum's policy of sharing knowledge, in 2010, the Amsterdam Museum made their entire collection available online using a creative commons license. The collection can be browsed through a web interface. The metadata of Amsterdam Museum was mapped to the Europeana Data Model (EDM) [13] and is currently hosted on the Europeana Semantic Layer.

The Amsterdam Museum data consists of three parts: (1) an object metadata set consisting of metadata records for the approximately 73 000 objects; (2) a thesaurus consisting of 28 000 concepts used in the metadata records and (3) a person authority file consisting of data about 67 000 persons related to the objects or the metadata. The metadata, thesaurus and vocabulary were all harvested through an OAI-PMH interface[5]. The resulting XML was first converted to crude RDF and subsequently restructured using interactive rewriting rules. Then the Amsterdam Museum specific classes and properties were mapped to those of EDM in order to make the Amsterdam Museum Linked Open Data interoperable with the EDM.

4 Metadata Standards for Museum Collections

The support for different types of interoperability is recognized as one of the main advantages of Sematic Web technologies. Interoperability may be divided into many categories but the following three types are most significant for the cultural heritage collections:

- *organizational interoperability* – refers to cooperation between and within cultural heritage organizations, business goals and process mod-eling;
- *semantic interoperability* – refers to understanding the meaning of in-formation stored in digital repositories;
- *technical interoperability* – refers to interconnection, presentation and exchange of digital objects within a digital repository or between reposi-tories, accessibility and security issues.

Semantic interoperability mostly depends on the development, adoption and use of metadata standards and proper ontologies.

[4] http://linkeddata.org/
[5] http://www.openarchives.org/

Amongst the most popular museum specific standards [8] one should men-tion CDWA, Object ID and MUSEUMDAT.

CDWA (Getty Research Institute, 1990) describes the content of art data-bases providing a conceptual framework for describing and accessing information about objects and images. It consists of 31 categories with 505 metadata types.

Object ID (J. Paul Getty Trust, 1999) is a standard for describing cultural ob-jects. It was developed through the collaboration of the museum community, police and customs agencies, the art trade, insurance industry, and values of art and antiques.

MUSEUMDAT (Zuse-Institut Berlin, 2006 – 2007) is a harvesting format optimized for retrieval and publication, meant to deliver automatically core data to museum portals.

Ontologies play a significant role in providing semantic interoperability of museum collections. They may be used for [9, 10]:

- *Information integration*. A core ontology, which incorporates basic enti-ties and relationships common across the diverse metadata vocabularies, might be useful for integrating information from heterogeneous vocabularies and uniform processing across heterogeneous information sources.
- *Deriving knowledge*. Ontologies organize the terms in heterogeneous domain vocabularies in a form that has a clear and explicit semantics and can be reasoned over. This process is fundamental in deriving new knowledge.

CIDOC CRM (Conceptual Reference Model) [12] is the most widely used ontology for the cultural heritage sector. It has been under development since 1996 and is currently being agreed as an ISO standard. The CIDOC CRM can be used as the basis for data exchange between systems, as a reference guide for the design of new cultural heritage information systems, and as the basis for integrated query tools and mediation systems' data schemas.

The CIDOC CRM is specifically intended to cover contextual information about the historical, geographical and theoretical background in which individual items are placed and which gives them much of their significance and value. As a formal ontology, it can be used to perform some types of reasoning.

5 Recommendations for the Future of Museum Collections on the Web

One of the directions of research and development activities in the area of building online museum collections should be their personalization. Personalization refers to providing differentiated access to information and services according to the user's profile and thus it helps to realize museums' educational, cultural and marketing functions.

Irrespective of the serious achievements in areas like semantic search, information extraction, etc., it is still difficult for people to find within the WWW the right information at the right time and at the right level of detail. In order to find a solution to this problem, researchers from different communities have developed systems that are

able to adapt their behavior to the goals, tasks, interests, and other features of their individual users and groups of users. The result is what we normally call adaptive or personalized systems. Another possible approach is to provide museum managers and users with some flexible tools for opinion mining and sentiment analysis that are applicable to the variety of blogs, social networks, forums, and other online systems, giving opportunities to share comments or evaluation of specific products, in particular museum collections available on the Web. Systems like SENTISITE [14] might play this role to a satisfactory degree.

SENTISITE is an experimental online system for opinion mining and sentiment analysis of short texts retrieved as results of keywords-based search in a given web page or in an online search platform. It is aimed to enrich the possibilities of using the internet by providing a software tool that retrieves and analyzes the sentiment of online texts which interest the users. Besides the factual information it provides an insight at the emotional aspect of themes, events personalities, and objects that are searched online. Two types of classification problems are being solved by SENTISITE – determining the neutrality of a text and determining the polarity of the sentiment in a text. Finally, the text is classified in one of the following three classes: positive, neutral or negative.

A number of Machine Learning algorithms have been experimented for the purpose of sentiment analysis of text documents. The researched and tested algorithms (Naïve Bayes Classifier, Support Vector Machines and K-Nearest Neighbors) provide proven good results in the field of text categorization. Techniques to improve the classification behavior of the algorithms have also been studied. They basically aim to compensate the disadvantages of the chosen formal presentation of a text as a vector of its words. Experiments were made to remove the non-informative features by evaluating the attributes–words with mutual information measure and χ^2 measure. The latter approach showed significant improvement in the accuracy of each of the tested algorithms. Other techniques have also been studied: filtering of stop words, adding the most informative bigrams to the classification features, analyzing negative semantic structures in a text, and adding features based on them. The chosen algorithm is Naïve Bayes Classifier with elimination of the non-informative features, based on the estimation made by the χ^2 measure.

The user interface of SENTISITE is implemented as a website. It allows the user to search for keywords in an online search engine, in a web page or a website. A report is generated in which the discovered results are grouped according to their sentiment estimates. It is possible to save both the search results and a history of the searches initiated by a particular user. A user-friendly interface is developed to monitor the sentiment changes, to display their graphical representation and to choose the time interval to be considered. The user is enabled to gain a clear idea of the sentiment and opinions on a specific topic within the WWW and to track out the trends of alteration of these emotional estimates.

6 Conclusion and Future Work

The existing results in application of Semantic Web technologies to digital preservation of and access to museum collections may be evaluated as promising. They demonstrate good exploitation of the underlying knowledge and satisfactory retrieval results when searching through the collections. But most successful teams currently deal at the level of the individual institution. We believe that in the near future the Semantic Web community could cooperate in handling heritage in ways that accurately reflect the society needs as well as the needs of each particular user. This will make cultural and scientific heritage far more accessible to people and will lend a helping hand to the adequate implementation of the multitude of its roles.

References

1. T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web. Scientific American, May 2001, pp. 34–43.
2. N. Shadbolt, W. Hall, T. Berners-Lee, The Semantic Web Revisited. IEEE Intelligent Systems, Vol. 21, No. 3 (May/June 2006), pp. 96-101.
3. T. Gruber, Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human-Computer Studies, Vol. 43 (1995), No. 5-6, pp. 907–928.
4. S. Ross, Position Paper: Towards a Semantic Web for Heritage Resources. In: DigiCULT Thematic Issue 3, ISBN 3-902448-00-8, 2003, pp. 7–11.
5. M. Lowndes, An Introduction to the Semantic Web for Museums. International Conference for Culture and Heritage On-line: Museums and the Web 2006 (Albuquerque, New Mexico, March 22-25, 2006). http://www.archimuse.com/mw2006/papers/lowndes/lowndes.html (vis. on July 20, 2014).
6. E. Hyvönen et al., MUSEUMFINLAND – Finnish Museums on the SemanticWeb. Journal of Web Semantics, Elsevier, Vol. 3, No. 2-3 (2005).
7. V. de Boer et al., Amsterdam Museum Linked Open Data. Semantic Web, Vol. 4, No. 3 (2013), pp. 237-243.
8. G. McKenna et al., Digitisation: Standards Landscape for European Museums, Archives, Libraries. ATHENA Project, http://www.athenaeurope.org/index.php?en/198/athena-booklets (visited on July 20, 2014).
9. P. Le Bœuf et al., Using an Ontology for Interoperability and Browsing of Museum, Library and Archive Information. International Council of Museums 14th Triennial Meeting (The Hague, Netherlands, September 12-16, 2005).
10. O. Signore, Ontology Driven Access to Museum Information. Annual Conference of CIDOC Documentation and Users CIDOC 2005 (Zagreb, Croatia, May 24-27, 2005).
11. C. Doulaverakis, Y. Kompatsiaris, M. Strintzis, Ontology-Based Access to Multimedia Cultural Heritage Collections – The REACH Project. Proceedings of the International Conference on Computer as a Tool EUROCON 2005, IEEE, 2005, pp. 151-154.
12. M. Doerr, The CIDOC CRM: An Ontological Approach to Semantic Interoperability of Metadata. AI Magazine, Vol. 24 (2003), No. 3, pp. 75–92.
13. Europeana Data Model Documentation, http://pro.europeana.eu/edm-documentation (visited on July 20, 2014).
14. E. Pavlova, System for Sentiment Analysis of Online Text Documents. MSc Thesis, Sofia University “St. Kliment Ohridski” – FMI, 2014.