Maria M. Nisheva-Pavlova
(Faculty of Mathematics and Informatics,
Sofia University "St. Kliment Ohridski" and
Institute of Mathematics and Informatics,
Bulgarian Academy of Sciences)

Pavel I. Pavlov,
Anna S. Devreni-Koutsouki
(Faculty of Mathematics and Informatics,
Sofia University "St. Kliment Ohridski")

ONTOLOGY-BASED ACCESS TO DIGITIZED CULTURAL HERITAGE AND ARCHIVAL COLLECTIONS

Abstract: The paper discusses several aspects of the use of ontological knowledge and some concomitant Semantic Web technologies in the development of software tools for operative access to digitized cultural heritage and archival collections. The emphasis falls on a number of general issues like semantic mark-up of content, information integration, interoperability of ontologies etc. Some domain-specific problems e.g. the scope of the ontologies that are needed for the purpose (and which ones should the heritage sector develop and which ones will be possible to borrow from other sectors) have been also analyzed. Two successful projects directed to the implementation of ontology-driven access to various types of cultural heritage repositories have been analyzed as examples of best practices in the area. The experience in building software tools for knowledge-based search in repositories of digitized manuscripts and archival materials gained at the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences and the Faculty of Mathematics and Informatics of Sofia University has been discussed as well.

Keywords: Digitization, Metadata, Ontology, Semantic Annotation

1. Introduction: Why Cultural Heritage and Archival Collections on the Semantic Web?

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.

A special characteristic of cultural collection contents is the semantic richness. Collection items have their history and are related in many ways to our environment, to the society, and to other collection items. For example, a chair may be made of wood and leather, may be of a certain style, was designed by a famous designer, was manufactured by a certain company during a time period, was used in a certain building together with other pieces of furniture, and so on. Other collection items, locations, time periods, designers, companies etc. can be related to the chair through their properties and implicitly constitute a complicated semantic network of associations. This 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.

Archival collections should be characterized as even more semantically rich than cultural ones. An archival document may concern time periods, events, persons, places etc. mentioned in completely different context in other documents in the same or in another collection and only the careful study of all related documents (which might be distributed in lots of repositories) can give an objective view to the information looked for.

Having in mind all these reasons, we consider it expedient to publish digitized heritage and archival 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 offers new possibilities when publishing museum and archival collections on the Web [3]: *collection interoperability in content* (Web languages, standards, and ontologies make it possible to make 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 can be created).

One of the major hurdles facing one in building software which uses Semantic Web technology is the lack of suitable ontologies. Languages such as OWL (W3C’s Web Ontology Semantic Markup Language for publishing and sharing ontologies) enable the rapid development of ontologies but lots of questions concerning multilingual capabilities and processing of synonyms stay still open. Thinking pragmatically, we need to consider [7]:

- Can we cost the creation of appropriate ontologies for the heritage sector?
- How can we prioritize the ontologies that are needed? In particular, which ones should the heritage sector develop and which ones will we be able to borrow from other sectors?
- What heritage-based organizations should focus on ontology creation?
- Ontologies often fail to be interoperable. What solutions are there to this problem and how can they be made to work effectively?
- Does OWL provide a suitable mechanism for ontology creation for the heritage sector?

Proof and trust are emerging as other central issues. How do we know that what our agent has discovered through its search on the Semantic Web can be trusted? Even in the case of ontologies how should we decide whose ontology to trust? This is especially important when two ontologies may conflict with each other. Similarly we are faced with the difficulties of ensuring and maintaining semantic integrity and the lack of methods for testing its presence.

2. The Best Practices

Several successful projects which provide ontology-based access to cultural heritage collections already exist. Among the most popular ones in this group we should mention the projects REACH and MUSEUMFINLAND.

The objective of the REACH project [2] is to develop an ontology-based representation in order to provide an enhanced unified access to heterogeneous distributed cultural heritage digital databases. The complete system will be composed of the following subsystems: (1) a cultural heritage web portal for unified access to the information and services, (2) digitization system for the efficient digitization of artwork and collections, (3) a core ontology to describe and organize cultural heritage content, (4) multimedia content-based as well as ontology-based search engine to offer advanced choices of searching methods, (5) 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 information (metadata) originating from different sources. The integration process involves efficient mapping of the available metadata to the concepts and relations of the core ontology, so 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 (CIDOC-Conceptual Reference Model) [1] ontology, developed by CIDOC, the Museum Documentation Standards Group, was used.

The web portal will provide advanced searching capabilities to users. The requirement is that users will be able to use a variety of searching functionalities so that access to the underlying information will be easier and more effective. These functionalities namely include ontology-based search, content-based visual search and a novel hybrid ontology-visual search.

The ontology-based search will give the opportunity to the users to take advantage of the ontological data structure and look for specific information. The search can be conducted by using two different methods. With the first method, predefined concepts will be available as links in the web interface. A tree-like interface gives an illustrative example of the structure of the underlying knowledge. By using this method the user can select a concept to start the search process. As a second step the user selects the desired place and the corresponding results are displayed. This approach is useful for visitors to the web portal to easily browse through the ontology and review the content.

Using the second search method, the user has the option to type in keywords in a text field. The ontology is queried and the objects that were found to contain the keywords in their metadata are displayed in the result set. This allows the users to have access to the ontology content by not restricting their searching criteria to a single field.

MUSEUMFINLAND [3] is the most ambitious and realized 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. The idea is to offer to the user a semantic browsing and searching facility in the combined collection knowledge base. This facility is implemented by a 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:

(1) The Artifacts (ObjectTypes) ontology is a taxonomy of tangible collection objects, such as pottery, cloths, weapons, etc. All artifact exhibits in the system belong to some class in this ontology. The taxonomy was extended with properties available from an underlying thesaurus MASA [4].

(2) The Materials ontology is a taxonomy of the artifact materials, such as steel, silk, tree, etc. The classes are based on MASA.

(3) The Actors ontology defines classes of agents, such as persons, companies etc., and individuals as instances of these classes.

(4) The Situations ontology is a taxonomy that includes intangible happenings, situations, events, and processes that take place in the society, such as farming, feasts, sports, war, etc. The classes are based on MASA.

(5) The Locations ontology represents areas and places on the Earth. It contains classes such as Continent, Country, County, City, Farm etc. The main content in the ontology consists of its individual location instances (e.g., Helsinki or Finland) and their mutual meronymy relations (e.g., Helsinki is a part of Finland).

(6) The Times ontology is a meronymy of various predefined historical periods. First, there are categories representing special eras of interest such as the Middle Ages and the time of WorldWar II. Second, there is a linear breakdown hierarchy of

centuries and decades. The properties of time concepts are a human readable label of period and the beginning and end year of the time interval.

(7) The Collections ontology is a taxonomy that classifies the collections included in the portal under the museums hosting them. The properties of the taxonomy indicate the name and the hosting museum of the collection.

All taxonomy classes in MUSEUMFINLAND are instances of metaclasses for which properties such as the creator, description, date of creation, etc. can be specified.

Ontogator will provide the user with the following semantics-based facilities:

- *View-based filtering.* Ontogator shows the multiple ontologies used in annotating collection data. By selecting ontological classes from these hierarchies, the user can express the search profile easily in the right terminology.
- *Topic-based navigation.* Ontogator supports topic-based navigation according to the underlying idea of Topic Maps. The creation of semantic links between topics of interest is based on 1) the collection domain ontologies (classes and their relations) and 2) on actual collection data (instance data). The links give the user contextual and pragmatic information about the objects in the collection.
- *Ontological search engine for Finnish.* A search engine is being developed for generating hit lists in the same fashion as search engines on the WWW. However, the discussed engine will understand and make use of the semantic relationships between keywords.

3. The Bulgarian Experience

Bulgarian institutions are at the very beginning of the development of tools providing ontology-based access to digitized cultural heritage and archival collections. Some first results in this direction [6] have been obtained at the Digital Humanities Department of the Institute of Mathematics and Informatics at Bulgarian Academy of Sciences in collaboration with specialists from the Computer Informatics Department of the Faculty of Mathematics and Informatics at Sofia University “St. Kliment Ohridski”.

We elaborated a methodology for development of tools for knowledge-based search in repositories of digitized manuscripts. It is designated to assist the search activities in collections that may enlist XML documents which should be catalogue descriptions or marked-up full texts of mediaeval manuscripts. Our methodology is directed to the development of software environments that will be able to deal with complex user queries and answer questions such as “When are written manuscripts in which natural calamities or irregularities are mentioned?” or “Where are stored manuscripts in which significant social events are mentioned?”.

Currently we lay aside the problems connected with the processing of questions formulated in natural language and concentrate on queries containing conjunctions and disjunctions of key words and phrases.

As a result of the processing of a user query, a set of documents (manuscript descriptions and/or texts of manuscripts) containing words and phrases semantically related to these used in the query should be retrieved and properly visualized. The scope of the queries should not be predefined, but it is necessary to have a clear idea about their area(s) in order to provide and describe the corresponding domain knowledge.

The emphasis in the suggested methodology falls on the following main topics:

- Development of proper ontologies describing the conceptual knowledge relevant to the chosen domain(s). These ontologies define sets of concepts with their basic properties and the relationships between them. The concepts should be defined in many languages.
- Development of proper intelligent agents for search and processing purposes that are able to retrieve and filter documents by their semantic properties.

The main idea of our methodology is to provide the search engine with the necessary knowledge describing the semantic relationships between concepts in a wide range of domains. This knowledge can be represented as a set of appropriate ontologies. The ontologies used in our experimental implementation have mostly the form of concept hierarchies. They describe sets of domain concepts with the class–subclass relation between them and thus introduce the specific terminology of interest for various types of users. These ontologies are utilized by the search engine to augment the user queries with words and phrases denoting more particular concepts than the ones used in the original search requests. Some suitable dictionaries of synonyms should be used for similar purposes as well.

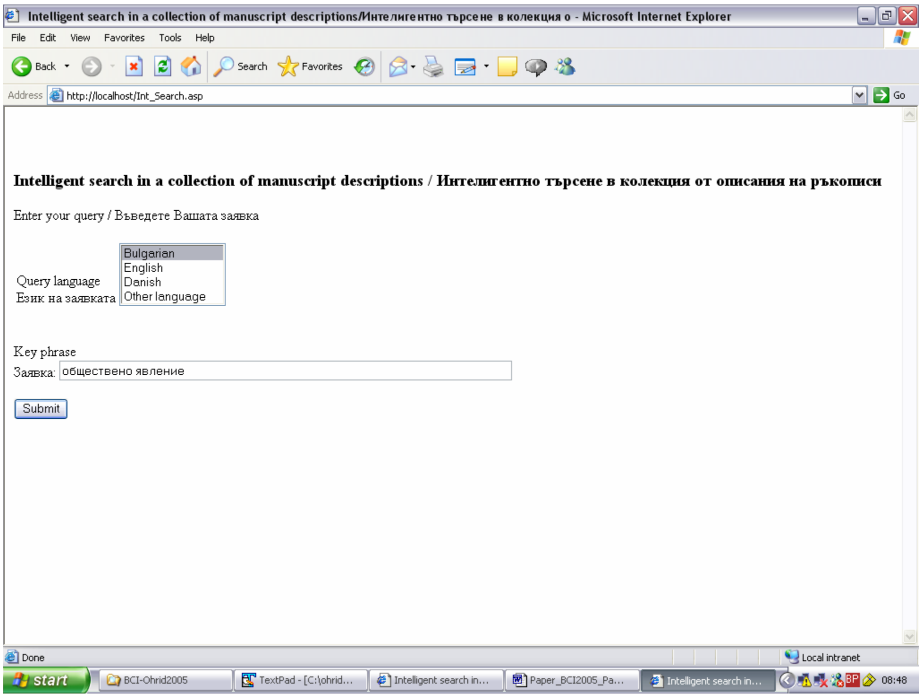
An experimental software tool that implements the discussed methodology with some restrictions imposed on the user queries has been recently under development.

A typical user query in the discussed software tool may contain a word or a phrase of interest for the person who formulates the query. The goal is to find all documents in the collection containing the originally given word/phrase or words/phrases that are semantically related with it and then to display properly the corresponding elements of the found documents.

Figure 1 shows a screenshot displaying an example user query in the discussed software tool. The user is asked to determine the language in which the query is formulated and then to type the phrase representing the query.

After the user query is entered, it is processed consecutively by the corresponding intelligent software agents.

As a result of the user query processing, the corresponding XML elements of all documents in the collection containing words or phrases semantically related to the one given by the user are properly visualized.


Intelligent search in a collection of manuscript descriptions/Интелигентно търсене в колекция о - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Go
Done
Local intranet
start
BCI-Ohrid2005
TextPad - [C:\ohrid...
Intelligent search in...
Paper_BCI2005_Pa...
Intelligent search in...
08:48

Address http://localhost/Int_Search.asp

Intelligent search in a collection of manuscript descriptions / Интелигентно търсене в колекция от описания на ръкописи

Enter your query / Въведете Вашата заявка

Query language
Език на заявката: Bulgarian | English | Danish | Other language

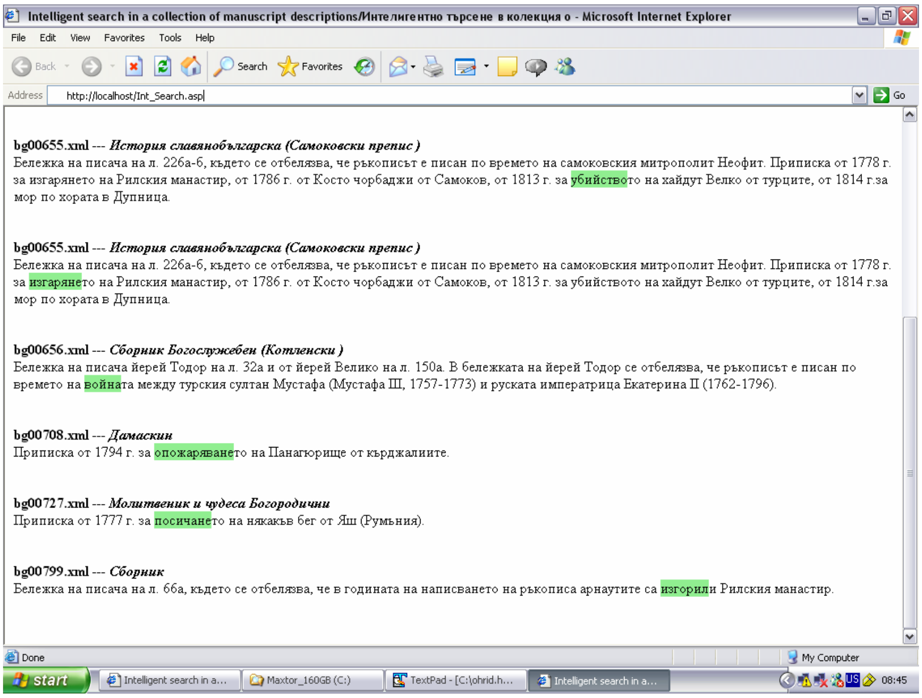
Key phrase
Заявка: общественo явление

Submit

Fig. 1. An example user query

Our experiments have been carried out with an existing collection of approximately 800 descriptions of mediaeval Bulgarian manuscripts. These manuscripts are mostly with religious content and some multiform information could be found mainly in the XML element “additions” of their catalogue descriptions (this element is used to record and discuss any written or drawn additional text found in a manuscript, such as marginalia, scribblings, etc. which the cataloguer considers of interest or importance). Because of that we decided to restrict the search in our collection and to perform it only in the “additions” elements of the XML documents.

Figure 2 shows a screenshot displaying part of the search results for the key phrase “social phenomenon” in this collection.


Intelligent search in a collection of manuscript descriptions/Интелигентно търсене в колекция о - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost/Int_Search.asp
Go
Done
My Computer
start
Intelligent search in a...
Maxtor_160GB (C:)
TextPad - [C:\ohrid.h...
Intelligent search in a...
08:45

bg00655.xml --- ***История славянобългарска (Самоковски препис)***
Бележка на писача на л. 226а-б, където се отбелязва, че ръкописът е писан по времето на самоковския митрополит Неофит. Приписка от 1778 г. за изгарянето на Рилския манастир, от 1786 г. от Косто чорбаджи от Самоков, от 1813 г. за убийството на хайдут Велко от турците, от 1814 г.за мор по хората в Дупница.

bg00655.xml --- ***История славянобългарска (Самоковски препис)***
Бележка на писача на л. 226а-б, където се отбелязва, че ръкописът е писан по времето на самоковския митрополит Неофит. Приписка от 1778 г. за изгарянето на Рилския манастир, от 1786 г. от Косто чорбаджи от Самоков, от 1813 г. за убийството на хайдут Велко от турците, от 1814 г.за мор по хората в Дупница.

bg00656.xml --- ***Сборник Богослужебен (Котленски)***
Бележка на писача йерей Тодор на л. 32а и от йерей Велико на л. 150а. В бележката на йерей Тодор се отбелязва, че ръкописът е писан по времето на войната между турския султан Мустафа (Мустафа III, 1757-1773) и руската императрица Екатерина II (1762-1796).

bg00708.xml --- ***Дамаскин***
Приписка от 1794 г. за опожаряването на Панагюрище от кърджалиите.

bg00727.xml --- ***Молитвеник и чудеса Богородични***
Приписка от 1777 г. за посичането на някакъв бег от Яш (Румъния).

bg00799.xml --- ***Сборник***
Бележка на писача на л. 66а, където се отбелязва, че в годината на написването на ръкописа арнаутите са изгорили Рилския манастир.

Fig. 2. Some search results

The discussed methodology is quite general and is at the root of two ongoing projects directed to digitization and on-line access to archival collections.

The first of these projects is directed to the digitization of a collection of archival documents from the period of the organization of the Sofia Municipal Government (1878 – 1879) and the development of a website presenting this collection [5]. This collection consists of approximately 980 original hand-written documents concerning the establishment of civic authorities of Sofia, building the administrative system, the order and law authorities, communal health services and educational system etc. around and after the end of the Russo-Turkish war (1877 – 1878). Because of these reasons we consider it expedient to include in the digitized version of the collection not only digital images of the chosen archival documents but also structured electronic transcriptions of their full texts and proper descriptions of the collection as a whole as well as descriptions of its parts (known as archival units) and all particular documents in it.

Our final goal is to give the user the opportunity to switch between two types of interface to the collection. The first one is based on the principles of the “standard” archivist’s view to an archival collection. The user can browse the properly visualized metadata describing the collection at various levels and the different kinds of representations of the documents within the collection. Short historical data accompany this type of interface to the collection.

The second type of provided on-line access to the collection may be described as the semantics oriented one. A set of access tools realizing various types of document search and retrieval (chronological, oriented to the kinds of documents within the collection, subject oriented etc.) has been under development for the purpose. Most of these tools use the values of the corresponding elements of the structured electronic transcriptions of archival documents. In particular, the subject oriented search is based on the use of the semantic annotation of the documents. The semantic annotation consists of appropriate words and phrases (chosen form an especially created ontology) that describe the content of the document. When the user defines his query, the corresponding access tool augments it by words and phrases semantically related to these used in the original query and some synonyms of the main terms from an appropriate dictionary. Then the obtained query is processed in a standard way. A tool for search in the full texts of the document transcriptions is provided as well.

The second project in the area of ontology-based access to archival collections is aimed at building an electronic archive of documents issued by the Bulgarian Ministry of Education in the 40ies and 50ies of the 20th century and stored in archival funds 177K and 798K within the State Archival Fund of the General Department of Archives at the Council of Ministers of the Republic of Bulgaria. This archive contains digital images of various types of documents of the educational institutions and the governmental bodies (official documentation, letters, certificates, notes and other working materials, photographs, newspapers etc.) concerning the organization and development of the educational system in Bulgaria. The digital copies of more than 1500 documents are accompanied with proper descriptions containing corresponding types of metadata (depending on the types of the original documents and the methods of their creation). A special kind of metadata is the semantic annotation of each particular document. It consists of concepts from an especially developed domain ontology covering the structure of the Bulgarian educational system and educational administrative documentation. This ontology is created having in mind the specific professional interests of the expected typical users of the electronic archive. It includes more than 100 concepts (classes) with the most important relationships between them. The emphasis falls on the parts of the ontology related to the evaluation and to the efficiency of the educational process. This “educational” ontology will play the key role in the development of various software tools for semantics oriented browsing, search of information and document retrieval.

4. Conclusion

The first experimental results in application of Semantic Web technologies to digital preservation and providing access to cultural heritage and archival 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 hope that in the near future Semantic Web people could handle heritage in ways that accurately reflect the community needs, and not always just the wishes of the institutions that own content. This will make cultural and scientific heritage far more accessible to those people who want to know but it requires leadership and opportunity from Governments as well as large scale, collaborative efforts from the international community.

Acknowledgements. This work has been partly funded by the Sofia University SRF under Contract No. 106/2007.

References

[1] Doerr, M. *The CIDOC CRM: An Ontological Approach to Semantic Interoperability of Metadata*. AI Magazine, Vol. 24 (2003), No. 3, 75–92.

[2] Doulaverakis, C., Y. Kompatsiaris, M. Strintzis. *Ontology-Based Access to Multimedia Cultural Heritage Collections – The REACH Project*. Paper available at http://www.cost292.org/pubs/eurocon05-/Charalampos_Doulaverakis.pdf, visited on October 20, 2007.

[3] Hyvönen, E. et al. *MuseumFinland - Finnish Museums on the Semantic Web*. Journal of Web Semantics, Vol. 3, No. 2, 2005.

[4] Leskinen, R. L. (Ed.). *Museoalan asiasanasto*. Museovirasto, Helsinki, Finland, 1997.

[5] Nisheva-Pavlova, M., P. Pavlov, N. Markov, M. Nedeva. *Digitisation and Access to Archival Collections: A Case Study of the Sofia Municipal Government (1878 – 1879)*. Proceedings of the 11th International Conference on Electronic Publishing (Vienna, Austria, 13–15 June 2007), ISBN 978-3-85437-292-9, 2007, 277–284.

[6] Pavlov, P., M. Nisheva-Pavlova. *Knowledge-Based Search in Collections of Digitized Manuscripts: First Results*. Proceedings of the 10th International Conference on Electronic Publishing (Bansko, Bulgaria, 14–16 June 2006), ISBN 978-954-16-0040-5, 2006, 27–36.

[7] Ross, S. Position *Paper: Towards a Semantic Web for Heritage Resources*. In: DigiCULT Thematic Issue 3, ISBN 3-902448-00-8, 2003, 7–11.