Investigation on Semantics to Improve the COVAX System

Luciana Bordoni

ENEA – UDA/ADVISOR -
Via Anguillarese, 301, 00060 S. Maria di Galeria (Rome), Italy
bordoni@casaccia.enea.it

Abstract

The purpose of COVAX (Contemporary Culture Virtual Archives in XML) financed by the European Commission in IST Programme was to analyse and draw up the technical solutions required to provide access through the Internet to homogeneously-encoded document descriptions of archive, library and museum collections based on the application of XML. The project demonstrated its feasibility through a prototype containing a meaningful sample of all the different types of documents to build a global system for search and retrieval. The aim of this paper is to create in the COVAX system a new presentation of the documents. A system capable of processing markup semantics declarations can act as an interactive environment for testing conjectures and validating hypotheses. Semantics is one of the ways of improving information retrieval performances; we will explore this problem in the COVAX case study. We will investigate the possibility to derive a semantic knowledge from COVAX repositories, in order to improve the site analysis process and the query answering process.

1. Introduction

Information about a particular person or topic can be created by multiple users, served by various services and dispersed across multiple sites over the Internet. Adoption of standardized metadata vocabularies and ontologies are contributing to the realization of the next generation Web – the Semantic Web. One of the key promises of the Semantic Web (Berners-Lee; Hendler & Lassila, 2001) is that it will provide the necessary infrastructure for enabling services and applications on the Web to automatically aggregate and integrate information. Archives, museums and libraries are making enormous contributions to the amount of information on the Internet through the digitisation and online publication of their photographic, audio, film and video collections.

Within this paper we attempt to exploit all of these developments: the rapid growth in multimedia content, the standardization of content description, and the semantic web infrastructure. We think that by using automated computer processing of metadata to organize and combine information it will be possible to generate new knowledge. The Semantic Web is an activity of the W3C which aims to extend the current Web by providing tools that enable resources on the web to be defined and semantically linked in a way that facilitates automated discovery, aggregation and re-use across various applications. The Web Ontology Working Group are currently developing a Web Ontology Language (OWL), based on RDF Schema (http://www.w3.org/TR) for defining structured, Web-based ontologies which will provide richer integration and interoperability of data among descriptive communities. Ontologies are often seen as basic building blocks for the Semantic Web, as they provide a reusable piece of knowledge about a specific domain. Furthermore, ontologies have been accepted as powerful description tools, and for this reason they are appropriate for playing the role of semantic views. Currently, there is a great deal of interest in the development of ontologies to facilitate knowledge sharing in general and database integration in particular. Nowadays, XML has become an increasingly important data format for storing and interchanging data among various systems and databases on the Internet. As a new markup language that supports user-defined tags, and encourages the separation of document content from its presentation, XML is able to automate Web information processing, in particular for data exchange and interoperability which are major issues in business-to-business electronic commerce. On the other hand, to enable efficient business application development in large-scale electronic commerce environments, current XML lacks the modeling power in describing real-word data and their complex interrelationships, and thus providing the objects' necessary semantics. One current trend in the literature is to apply data models developed for semistructured and unstructured data to XML. Generic information modelling concepts promote the understanding of XML document management. The current XML technology provides lots of tools and languages, but there is almost no guidance for a precise semantic specification of the content or the logical structure of the document. New standards such as XML Schema or XLink aggravate these problems because they increase the number of syntactic alternatives how to specify the semantics of the XML document.

The remainder of the article is organized as follows. Section 2 presents the work carried out in the frame of COVAX. Section 3 characterizes the specific problems that motivate the need for a semantics in the COVAX system. Section 4 discusses conclusion and possible future work.

2. Covax system

During the past few years the development of information technologies, communication nets and mark-up languages, as well as the opening and interconnection of systems have contributed to create new possibilities for transmission, search and recovery of documents, especially of primary documents. At the same time, the potential of Internet together with the information needs

of the archives, libraries and museums community, researchers, professors and students, to access all type of documents have increased. Archives, libraries and museums are in the centre of this information treatment and dissemination process. They are strongly affected by the technological innovations in the field of support research and information.

The purpose of COVAX (Contemporary Culture Virtual Archives in XML)[1] financed by the European Commission in IST Programme (http://www.covax.org/) was to analyse and draw up the technical solutions required to provide access through the Internet to homogeneously-encoded document descriptions of archive, library and museum collections based on the application of XML. The project demonstrated its feasibility through a prototype containing a meaningful sample of all the different types of documents to build a global system for search and retrieval (Bordoni, 2002). It is based on the assumption that in libraries, archives and museums an enormous number of descriptions could be made available over the Internet by converting. existing records or by creating new ones to specific XML DTD's. The objective was to study the systems and databases to be integrated into COVAX. For each database to be integrated, the following information was collected: nature of the contents being stored, geographical and institutional scope, media (text, video, etc), computer systems and environments (hardware, software, standards supported, openness, etc). The aim of this phase was to establish the way of performing the data integration, to assess the common representation schema and prepare conversion or adaptation procedures. It can be said that the issue with more influence along the project was the existence of five different types of MARC records (IBERMARC, CATMARC, UNIMARC, UKMARC, LIBRISMARC).

Regarding the content scope it is interesting to underline the synergy between partner' collections, because project' partners come from a very different background: multimedia producers, university libraries, research institutions, libraries consortium. This has produced a wide range of different documents, with different levels helping to produce a complete picture of contents.

COVAX partners have implemented two different database models: ad hoc XML databases, or existing non-XML repositories. In the latter case, information is retrieved from the original database and transform into XML format before presenting it to users. To summarize, COVAX is not only incorporating XML as a basic standard but also integrating other standards, and adapting them to XML.

COVAX partners have implemented XML repositories using two software packages, Tamino from Software AG, a COVA X technical partner and TeXtML from IXIAsoft. Sites have been established in Graz, London, Madrid, Rome and Salzburg.

[1] The consortium was formed by the following institutions: Residencia de Estudiantes (RE, Spain); Software AG España, S.A. (Spain); Angewandte Informationstechnik mbH. (Austria); Universitat Oberta de Catalunya (Spain); University of Karlskrona/Ronneby (Sweden); Salzburg Research mbH (Austria); Biblioteca de Menéndez Pelayo (Spain); London and South Eastern Library Region, replaced by South Bank University (United Kingdom) and Ente per le Nuove tecnologie, l'Energia e l'Ambiente (Italy).

3. Covax: a semantic perspective

Ontologies are a key enabling technology for the Semantic Web. They interweave human understanding of symbols with their machine processability. Ontologies were developed in artificial intelligence to facilitate knowledge sharing and re-use. Since the early 1990s, ontologies have become a popular research topic. They have been studied by several artificial intelligence research communities, including knowledge engineering, natural-language processing and knowledge representation. More recently, the use of ontologies has also become widespread in fields such as intelligent information integration, cooperative information systems, information retrieval, electronic commerce, and knowledge management. The reason ontologies are becoming popular is largely due to what they promise: *a shared and common understanding of a domain that can be communicated between people and application systems.*

Developing ontologies is central to the vision of Semantic Web-based knowledge management. However, because of the size of ontologies, their complexity, their formal underpinning and the necessity to come towards a shared understanding within a group of people when defining an ontology, ontology construction is still far from being a well-understood process. In recent years, research has aimed at paving the way for the construction of ontologies by ontology development environments. Different directions have been taken to support the engineering of ontologies:

1. Several seminal proposals for guiding the ontology development process by engineering methodologies have been described which influenced the ontology development environments.
2. Inferencing mechanisms for large ontologies have been developed and implemented also to support ontology engineering.
3. The need to achieve consensus about an ontology was reflected by collaborative environments for ontology engineering.

Ontologies glue together two essential aspects that help to bring the web to its full potential:

?? Ontologies define formal semantics for information, consequently allowing information processing by a computer.

?? Ontologies define real-world semantics, which makes it possible to link machine processable content with meaning for humans based on consensual terminologies.

XML provides an homogeneous and normalised environment to formalise and code heterogeneous documents (coming from archives, libraries or museums) in structures of information. The construction of common methods for search and retrieval documents and databases of varied typology is then facilitated. Another important

factor is that XML allows to convert isolated electronic resources, existent in archives, libraries and museums, in a network of distributed informative resources, which can be extended beyond their framework. The disadvantage of "flat", format-based representation languages such as HTML, is that they rarely combine information about content (the text a wtiter wants to disseminate) and layout (the format in which this is to be done). Information access to users is provided independently of their location or structural characteristics. Some users refer to XML as "semantic markup" contrary to HTML. Others say "XML is just syntax- no semantics". XML has merely the potential to improve the document semantics through markup because with XML as language, it is possible to introduce new tags that represent some meta information. Like any other specification language, it can be used to express the semantics of documents and their components to a certain extent. An XML document is stored in a database (or somewhere else) and retrieved as the "same" document back again. This is important for XML applications that need to retrieve exactly the document with exactly the same layout which includes things in XML like CDATA sections, character entities, comments, and processing instructions. DTDs or XML Schemas are ill-suited to express the semantics of the content of the document elements. The reason is that they had been designed for serialization of data as a prerequisite for data exchange among systems. In that case the interacting systems are themselves responsible for the enforcement of the data integrity constrains that cannot be expressed with XML.

Many new technologies have been developed in recent years to augment the usefulness of conventional structured markup. The aim of this paper is to create in the COVAX system a new presentation of the documents. A system capable of processing markup semantics declarations can act as an interactive environment for testing conjectures and validating hypotheses. Markup semantics are modeled computationally by applying knowledge representation to the problem of making the abstract structures, relationships and properties explicit. Encoded markup semantics, as investigated in the BECHAMEL project, (Renear, et al. 2002) promises important contributions. Advances in Semantic Web projects can contribute to develop such a semantics. The goal is to have the possibility to represent abstractions, relationships, and constraints. Expressive representations can support the development of better document processing systems.

Although XML Document Type Definitions provide a mechanism for specifying, in machine-readable form, the syntax of an XML markup language, there is no comparable mechanism for specifying the semantics of an XML vocabulary. Semantics is one of the ways of improving information retrieval performances; we will explore this problem in the COVAX case study. The baseline hypothesis is that the ontology-based solution will make it easier for the users to locate the information they seek and will also make it easier to share knowledge with others in the organization.

We will investigate the possibility to derive a semantic knowledge from COVAX repositories, in order to improve the site analysis process and the query answering process. The first step in this scenario is to identify relevant information entity types in the data sources. Entities are real-world concepts such as "products" or "documents", information about which is stored in – or can be derived from – the data sources. Then the taxonomical terms are determined by which the entities are classified, based on the characteristics of the available content. The classification terms may correspond to types from the domain ontology, property values, ranges of property values, combinations of property values, or relations between instances. The sequence of taxonomical terms is determined by defining generally applicable paths by which users can navigate through the taxonomy. Multiple navigation specifications are possible for each set of entities. Cluster maps can also be used for graphical navigation. Maps gradually present deeper levels of the ontology and show the current class and its super- and sub-classes.

This case study is to be seen as an exploration study. For this we would need more information retrieval tools, both standard ones and ontology-based semantic access tools. However, this exploratory case study will give a good indication as to what directions we should aim for a semantic tool.

4. Conclusions

Originally, an ontology should reflect the "truth" of a certain aspect of reality. It was the holy task of a philosopher to find such truth. Today, ontologies are used as a means of exchanging meaning between different agents. They can only provide this if they reflect an inter-subjectual consensus. By definition, they can only be the result of a social process. Thus, ontologies are as much a pre-requisite for consensus and information sharing as they are the results of them. For this reason, ontologies cannot be understood as a static model. An ontology is as much required for the exchange of meaning as the exchange of meaning may influence and modify an ontology. Consequently, evolving ontologies describe a process rather than a static model. Evolving over time is an essential requirement for useful ontologies. In the past, IT for knowledge management has focused on the management of knowledge containers using text documents as the main repository and source of knowledge. Web technology, especially ontologies and machine–processable relational meta-data, pave the way to enhanced KM solutions.

Although semantic approaches have shown the benefits of ontologies, there still exist a large number of open research issues that have to be addressed in order to make semantic web technologies fully effective. Important next steps remain to be done to expand this case study towards a service for the Semantic Web.

5. References

Berners-Lee, T. Hendler, J. & Lassila, O. (2001). The Semantib Web. Scientific American.

Bordoni, L. (2002). COVAX: A Contemporary Culture Virtual Archive in XML. In Proceedings of the 6th European Conference ECDL 2002, Rome, Italy, 661-662.

Feng, L. Chang, E. & Dillon, T. (2002). A Semantic Network-Based Design Methodology for XML Documents. ACM Transactions on Information Systems, 20 (4), 390-421.

Fensel, D. (2001). Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce. Berlin: Springer-Verlag.

Guarino, N. & Welty, C. (2000). Identity, unity, and individuality: towards a formal toolkit for ontological analysis. In Proceedings of ECAI-2000, August.

Mena, E. & Illarramendi, A. (2001). Ontology-based query processing for global information systems. Kluwer Academic Publishers.

Newell, A. (1982). The Knowledge Level. Artificial Intelligence, 18(1), 87-127.

Renear, A. Dubin, D. Sperberg-McQueen, C.M. & Huitfeldt, C. (2002). Towards a Semantics for XML Markup. In Proceedings of the DocEng'02. McLean, Virginia USA, 119-126.

Semantic Web Community Portal; http.//www.semanticWeb.org/index.html (current 26 Feb. 2004).