IST-2002- 507382

EPOCH

Excellence in Processing Open Cultural Heritage

Network of Excellence

Information Society Technologies

D.2.4.7 (Final): Showcase 7 "Archaeological Documentation for the Semantic Web"

Due date of deliverable: 29 April 2005
Actual submission date: 28 April 2005

Start date of project: 15/03/2004
Duration: 4 Years

Ename Center

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Activity 2.4.7

Showcase **Archaeological documentation for the Semantic Web**

Franco Niccolucci

PIN, University of Florence, Italy

Archaeological documentation for the Semantic Web is aimed at showcasing the potential of using XML, XML native databases and standards for archaeological documentation.

Starting from several sources, consisting of diverse archaeological documents, it shows how these may be encoded – if available as text files – or converted – if already available as structured data for example in a database – to a common format allowing data interchange, interoperability and search.

For the showcase a set of test files has been chosen and an initial analysis on them has been performed, aiming at defining the most suitable encoding systems for each of them. Tests include the following datasets:

1. Excavation data from a recent investigation in Northern Italy (various archaeological sites in Friuli), stored in a Microsoft Access database and structured according to the official ICCD (Central Italian Institute for Cataloguing) forms; a stratigraphic methodology was used for these excavations.
2. Excavation data from recent investigations in Southern Italy (Cumae), structured according to the Syslat documentation system. Syslat is a French package, developed in the 90s on a Macintosh platform and based on Hypercard. It envisages a specific methodology for documentation, based on the stratigraphic method but introducing additional concepts for interpretation (for example, so-called ensembles and facts).
3. Reports of surveys and excavations dating to the end of the 19th Century and the beginning of the 20th Century, in Florence (italy). In this case the stratigraphic method was implicit or not used.
4. Reports from investigations carried on at the end of the 19th Century in Cumae (Southern Italy). Here the stratigraphy is implicit.
5. Bibliographical records on old reports. These include a short summary created for the purpose.
6. Museographical records from the Museum project, Norway.
7. Short texts with descriptions of interpretations taken from illustrative material on Roman Florence produced in early ‘20s.

The analysis has evidenced that document encoding heavily relies on the nature of the source. For structured data, stored into databases, extraction depends on the DBMS used and generally produces a flat structure in the DTD, while text encoding tends to be richer and more layered. Since the definition of the elements to be used for text encoding is in principle arbitrary, interoperability may be improved if it includes elements compatible with those derived from the database structure. In the test cases, such a compatibility between different cases has been assured by controlling the definitions of data structures within the working team. This procedure cannot however be assumed as a general rule, because different teams may operate on different sources, often ignoring each other’s work until their work is published. The necessity of referring to the same standard is therefore evident. The definition of a guideline for archaeological text encoding has consequently been postponed to the availability of an archaeological implementation of a standard as CIDOC-CRM.

Database conversion from Syslat (case no. 2) has required considerable skills because data were stored on an obsolete system (the software is no more supported by the producer since 1998 and runs only on old equipment, with rather old versions of the operating system). However, after the

initial tests based on a limited number of records, the extraction was successfully performed for some 3000 records, all those present in the original database.

The conversion of this dataset has evidenced an unforeseen aspect. The absence of validation and the presence of different authors in a work continuing for many campaigns and throughout several years has led to inconsistencies e.g. in abbreviations. Just to quote one, the common “US” acronym (shorthand for Stratigraphic Unit) is present also as “U.S.”, “U S”, “UU.SS.” and so on. Also typos introduce additional inconsistencies in fields where no controlled vocabulary was used.

In the present case, the few controlled vocabularies included terms in French (coming from the original French implementation), of little or no use in a context where all the texts were in Italian and the archive was planned for an audience of Italian speakers.

Therefore a clean-up tool was prepared for internal use, and it developed into a general tool for “cleaning” data by end-users. The tool analyses the dataset and guesses the nature of fields, distinguishing among lists, where a controlled vocabulary would be preferential, descriptive fields, seldom used fields and so on. The user may check variants, modify values to correct errors or introduce normalization aliases without altering the source content. The tool is in fact designed for being used by archaeologists because decision on aliases, for instance, may only be taken by those having professional competence on the content.

After cleanup, a dataset of over 2500 records was made available for demonstrations, mainly concerning the search tool. The search tool is based on eXist, an open source native XML database which has been chosen for its superior performance with respect to similar Open Source products. It must be noted that eXist outperforms several similar commercial products; at least, according to commercial descriptions of the latter. Searches on eXist are based on the Xquery language.

The search tool is web based: a simple interface has been created with some search functions and options. A more general, and customizable, tool may be easily created, for instance allowing free choice of search fields, composition of search criteria and so on. The results are displayed as a list of records; visualized ones are displayed with hypertext links to records related to the current as recorded in the data, for example using stratigraphic relations.

As yet, the search tool is available only as a standalone, but it is planned to make it available on the web after IPR on the source data are cleared. Possibly, an intermediate solution requiring registration of web users will be adopted.

The showcase has demonstrated that archaeological data conversion is possible and in fact simple; that it can be performed also on “vintage” datasets, thus preserving them from technological obsolescence and eventual impossibility of access; that conversion allows “cleanup” of datasets, an operation which is very often indispensable; and that searches may be effective and have a good performance.

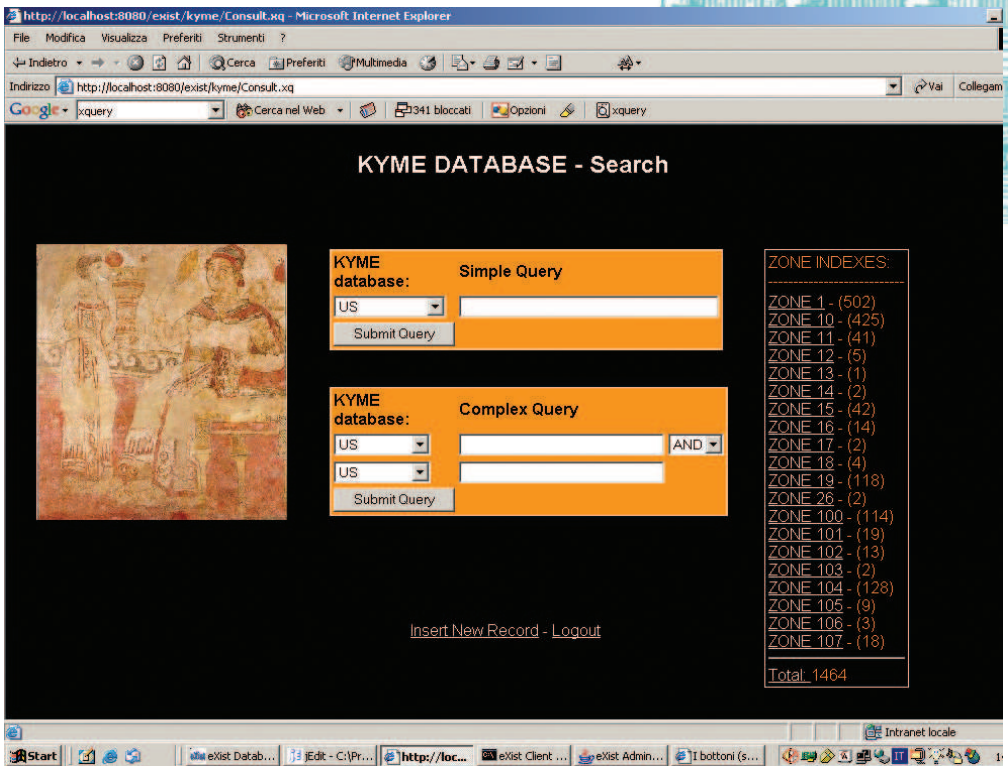
Another interesting feature was the compatibility with jnet, a tool developed within showcase “Tool for Stratigraphic Data Recording”. This allows using the stratigraphic data concerning relations to feed the present records into jnet and obtain the Harris matrix of the excavation.

The showcase has finally evidenced, as already mentioned, the necessity of the adoption of common documentation frameworks, as those provided by standards. While compliancy to standards is a matter of individual choice for research work, as archaeological text encoding, and their adoption may be fostered by appropriate guidelines and be motivated by the advantages of the use of standards, for instance interoperability, tool availability, and so on, international standards may encounter difficulties whenever national regulations are in force. In such cases, the availability of a mapping system of national “standards” to international ones may be a key factor to comply with local regulations and maintain the interoperability provided by international standards as CIDOC-CRM.

Archaeological Documentation for the Semantic Web

Different data sources Archaeological documentation is definitely going digital, but this trend may not be able to solve the problems arising when it is desired to perform a cross-archive search. What is in theory made possible by the support of IT, namely the possibility of managing effectively huge and diverse data archives, is often frustrated by the different structure such archives were given by their creators. This showcase aims at showing that such integration is in fact possible, with an already available technology which substantially improves the way digital archaeological data have been handled as yet. The showcase also will consider existing paper documentation and will show how it can be integrated with digital archives.

Standards The showcase aims at conjugating standardization – that is the compliancy with the CIDOC-CRM, the standard for museum collections developed by the ICOM documentation committee and already accepted as an ISO draft – with easiness of use and flexibility. However, it is planned to 'tokenize' the proposed encoding in order to guarantee cross-archive interoperability leaving to researchers the choice of record details, tailoring them to the specific researcher's needs. Multi-lingual issues are also addressed: when existing archives are in different languages, the system enables cross-searches trhough standardized description of the archive content.



The search page for the Cumae archive

Applications The tools used to develop the showcase are freely downloadable on the Internet amd are based on public domain standards. It is rather easy to convert existing databases as shown by the application to the excavation database of Cumae, a Classical site in Southern Italy and to a set of excavation databases and excavation diaries from a Medieval site. Digitization and encoding of ineteenth and early twentieth century reports of archaeological collections and sporadic finds in Florence has been successfully tested as well, while the encoding of data on collections of Norwegian archaeological museums, is being mapped to this new system. The case studies total more than 10.000 records.

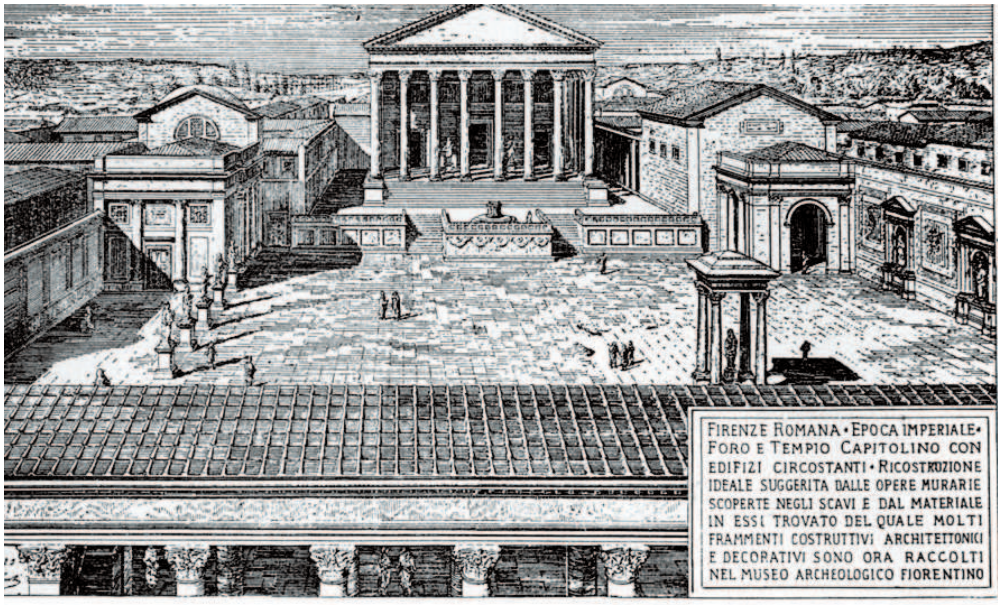
The management of images and/or drawings is still under development; in some cases it has been tested using SVG, an XML web-compliant vector format for drawings.

As far as excavation data are concerned, a system is being developed at Kent University to produce automatically the Harris matrix from the stratigraphic information encoded into the records.

The Semantic Web This buzz-word is used here to indicate the perspective assumed by the proposed methodology: archives may be accessed through the web by means of a browser and they are created in such a way that an intelligent web agent may search them and find relevant information according to user specified criteria.

As yet, the system allows cross-archive searches so that, for instance, data concerning ceramics may be searched in dif-

ferent excavation archives (e.g. referring to different campaigns), museum collections and old antiquarian reports. Such searches may take place remotely over the Internet on distributed archives.



Roman Florence in the reconstruction of Corinto Corinti (early XX century), one of the old archaeological reports considered as a case study

Technical details Archives are encoded using an XML CIDOC-CRM compliant structure. Existing databases may be easily converted with no loss of information. The search engine is based on eXist, an Open Source native XML DBMS. Data are organized in collections (corresponding to individual archives) with a hierarchical structure, and each collection may be searched separately or at any chosen aggregation level: e.g. all collections pertaining to a site may be grouped together in a super-collection while archives maintain their individuality, with a directory-like structure.

Partners The system is being developed by several partners at different locations to test it under as many diverse conditions as possible. Other contributors are in the meanwhile proposing additional case studies. The core partners are:

- PIN, Italy
- University of Oslo, Museum Project, Norway
- University of Kent, UK
- Paveprime, UK
- University of Naples "L'Orientale" – CISA, Italy

Interested?

Are you interested in this showcase? Do you think that this approach can help you in creating effective Cultural Heritage presentation projects or can be integrated in new research projects? Please contact Prof. Franco Niccolucci (niccolucci@unifi.it) of PIN at +39 0574 602513.

EPOCH is a Network of Excellence on Intelligent Cultural Heritage within the IST (Information Society Technologies) section of the Sixth Framework Programme of the European Commission. EPOCH showcases demonstrate innovative solutions and technological integration for target application areas in the Cultural Heritage domain. As they are created with real world content, they stimulate creative thinking about the use of the technologies in Cultural Heritage, and are used to validate new technological approaches with key stakeholders in the Cultural Heritage domain. For more details, visit the project web site:

www.epoch-net.org

EPOCH is funded by the European Commission under the Community's Sixth Framework Programme, contract no. 507382. However, this leaflet reflects only the authors' views and the European Community is not liable for any use that may be made of the information contained herein.