Digital collections of semantically annotated cultural heritage texts

Achille Felicetti
Italy

Free text is a class of information that comprises a wide range of cultural heritage documents, usually very difficult to process with traditional forms and relational databases. To this class belong various types of documents, (e.g. texts from ancient sources, comments or technical notes produced by cultural heritage experts, excavation diaries written during archaeological excavation activity) showing different schemas and multiple ways of content organisation and presentation.

The creation of coherent archives to store the content of these documents in a meaningful way implies a totally different point of view, where the main focus is on the text and its meaning, rather than on the structure of its *container* (e.g. the tables of a database or the fields of a form). This document-centric approach provides a way of preserving the integrity of the original documents without sacrificing efficient information retrieval.

In the past many attempts have been made to solve the problem of encoding free texts in a standard way and many philosophical and practical approaches have been tested. The appearance of XML at the end of the Nineties gave new hope to the advocates of documentation systems more respectful of the richness of information. Initially proposed for historical sources, it was suggested that XML encoding of cultural heritage texts could provide a way of dealing with legacy data, mostly still on paper, in manuscript or printed form. This was proposed for 19th century archaeological records and implemented in the Norwegian Museum Project and in the EU ARENA project.

Such pioneering work was appropriate for the sources it dealt with, but had limits in efficiency and did not consider interoperability as a priority goal. Attempts to use XML-native DBMS did not go well. Above all, such systems did not enable the creation of relations among the elements used for encoding and gave no solution for cross- and co-reference.

The introduction of CIDOC-CRM as an ISO 21127 standard for cultural heritage documentation opened up new ways of dealing with cultural heritage data in an interoperable way: the flexibility and extensibility of the model simplify the process of adapting the schema to the content of the document and the mapping process between different schemas (see Figure 1).

At present, the technology is mature enough to allow the development of systems for the complete management of this type of document and the creation of semantic digital archives which can guarantee, above all:

Achille Felicetti

Digital collections of semantically annotated cultural heritage texts

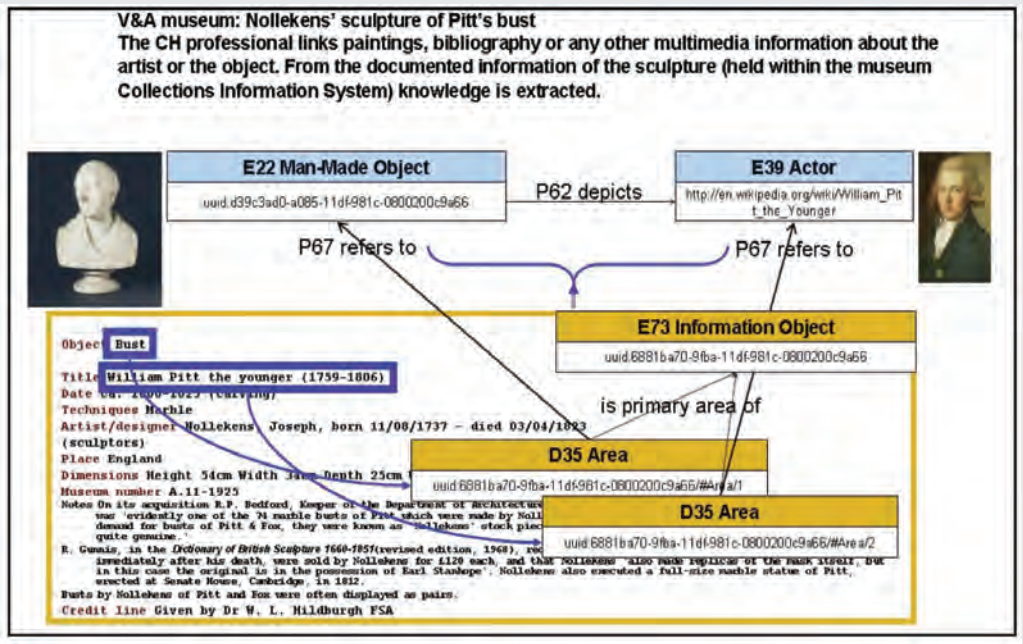


Fig. 1

• Preservation of the source integrity, making no changes, extraction or summarization in the original text.
• Efficiency in retrieval, using tools with good performance in searching.
• Interoperability, relying on CIDOC-CRM for encoding.
• Interaction with semantically-enabled tools based on RDF.
• Encoding and management by non-technical users availing themselves of user-friendly interfaces

A typical system for the creation and management of a digital archive of texts should provide different tools combined together, in particular:

• A featured container for the semantic and textual information, providing at least a triple store for the RDF and CIDOC-CRM encoded information and a document container where the digital document could be stored, like FedoraCommons;
• An annotation tool for the editing and annotation of documents and for the creation of semantic relations;
• Other tools and interfaces for the implementation of interoperability between the textual information and other kinds of data (images, 3D models, geographical information and so on):

The annotation tool should be the core of the whole system and should provide all the necessary features to capture meaningful information from a text and to store both the original text and the extracted knowledge into the semantic container.

An important addition to the text annotation tool consists of a thesaurus and a gazetteer, used to normalize the terminology and the semantic structure of the text.

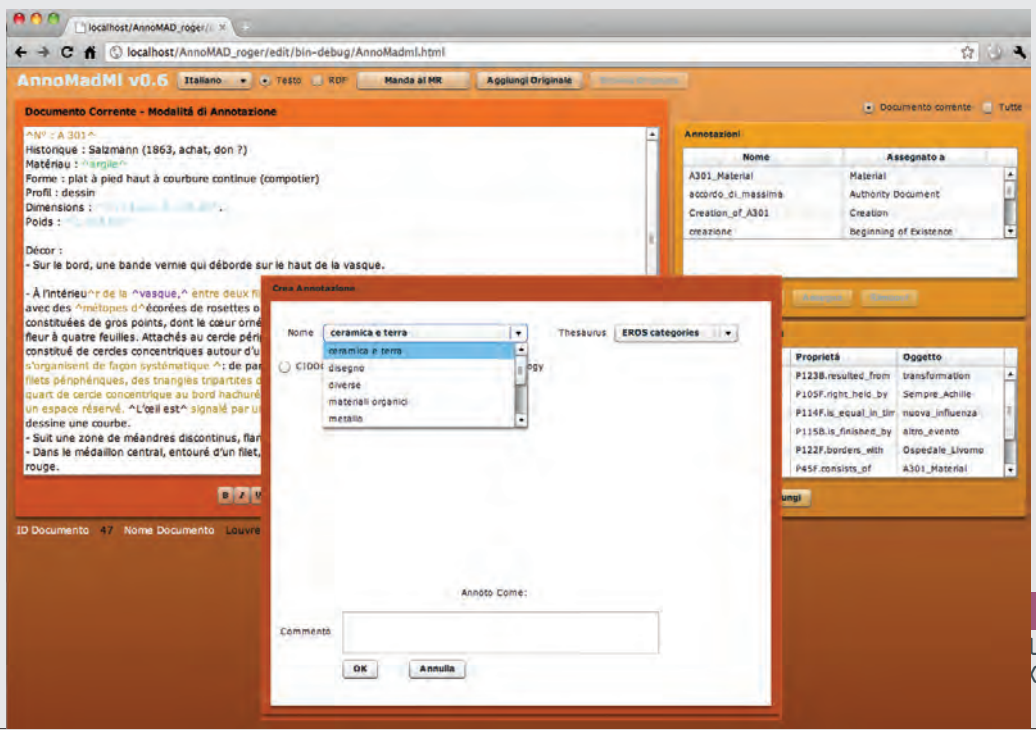
The semantic container should guarantee the creation of a coherent archive, and should make all the information available for every kind of operation or request by replying to many requests. It should support many communication protocols such as REST, SOAP, and JSON, and it should provide SPARQL query features to retrieve information in a semantic way.

The whole system should also provide standard mechanisms for importing external resources (for instance OAI-PMH metadata regarding text stored in other digital libraries) and export the whole archive or parts of it in RDF, Dublin Core, OAI-PMH, and other standard formats by using mapping mechanisms, in order to provide information to other institutions and projects such as Europeana.

At the VAST-LAB (University of Florence) we have been investigating this issue for many years and have developed many tools to try and solve the difficult problems related to text encoding. We are currently collaborating in the European 3D-COFORM Project, developing tools and technologies to aid in the management of metadata regarding 3D models and for their integration with the semantic information extracted from annotated texts.

One of the tools we are developing in 3D-COFORM is AnnoMAD, a free-text annotation tool. AnnoMAD is already under testing (in its beta version) by a team of archaeologists in Cyprus to create digital archives of textual archaeological documentation, and will be released as part of the 3D-COFORM integrated toolset (see Figure 2).

Fig. 2



Achille Felicetti

Digital collections of semantically annotated cultural heritage texts

The tool is intended to manage all the non-structured raw text information that is usually impossible to manage using structured tools, like RDBMS. It will offer the possibility to superimpose CIDOC-CRM entities on pieces of free text according to specific criteria in order to create a layer of CIDOC-CRM encoded semantic metadata referring to the original document.

AnnoMAD, in its final release, will allow people working in the cultural heritage field to define a semantic description for free-text documents, a category of data that is usually very difficult to manage and whose content is sometimes impossible to retrieve. By using this tool, the cultural heritage expert in charge of this kind of document will have the possibility to associate semantic entities to portions of text in order to describe the hidden meaning of the text in a machine-readable way. At the end of the process, information will be stored and made available via a metadata repository, along with all the other similar semantic information stored therein.

The repository we will use with AnnoMAD is under development within the same 3D-COFORM framework by the FORTH institute in Crete. It will be released as a semantic repository infrastructure to manage and store semantic information and digital content, including texts in various formats (html, txt, doc, pdf, etc.) and to implement interoperability among all data and metadata.

References

Crescioli Marco, D'andrea Andrea, Niccolucci Franco, *XML Encoding of Archaeological Unstructured Data*. In: *Archaeological Informatics: Pushing the Envelope. Proceedings of CAA2001*. Editor G. Burenhult Oxford: Archaeopress, 2002 p. 267–275.

Felicetti Achille, *MAD: Managing Archaeological Data*. In: *The evolution of Information and Communication Technology in Cultural Heritage*. Ed. Ioannides M. et al. . Budapest: Archaeolingua, 2006, p. 124–131.

Felicetti Achille et al.: *AnnoMAD: A Semantic Framework for the Management and the Integration of Full-text Excavation Data and Geographic Information*. In: *The 11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST*. Ed. A. Artusi, Paris, Eurographic Symposium Proceedings, 2010 p. 123-130.

Felicetti Achille, Mara Hubert: *Semantic Web, Digital Libraries and the future of Cultural Heritage*. In: *VAST: 9th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage*. Ed. Ashley M. et al., Aire-la-Ville, Eurographics Association, 2008 p. 117–123.

Holmen Jon, Uleberg Espen: *SGML-encoding of archaeological texts*. Paper presented at the IAAC' 96 IASI: Romania. 1996. http://www.dokpro.uio.no/engelsk/text/ http://www.dokpro.uio.no/engelsk/text/ getting_most_out_of_it.html. [viewed 30 March 2011]

Niccolucci Franco: *XML and the future of humanities computing*. *ACM* Applied Computing Review 2002 vol. 10, is. 1, p. 43–47.

Niccolucci Franco at al.: *Managing Full-text Excavation Data with Semantic Tools. In: VAST: 10th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage*. Ed. K. Debattista et al., Paris, France, Eurographics Association, 2009, p. 125–132.