Ontologies and Semantic Tools for the Management of Full-text Archaeological Documentation. Assessments from the Hala Sultan Tekke Case-study

Niccolucci, F.[1], Felicetti, A[2], Samaes, M.[3], Hermon, S.[1], Nys, K.[3]
[1] S TARC, The Cyprus Institute, Cyprus
[2] PIN, Universitá degli Studi di Firenze, Italy
[3] MARI, Vrije Universiteit Brussel, Belgium
niccolucci@cyi.ac.cy, achille.felicetti@pin.unifi.it, msamaes@vub.ac.be, s.hermon@cyi.ac.cy, Karin.Nys@vub.ac.be

The present paper describes a documentation system developed for the management of free text information produced or used by archaeologists during their activities. This class of information comprises a wide range of documents, usually very difficult to process using traditional forms and relational databases. Our system provides a way of preserving the integrity of the original documents without sacrificing the efficiency in information retrieval. The development and the testing of our system is focused on the management of documentation produced during the excavation at Hala Sultan Tekke, an archaeological site located in south-eastern Cyprus.

Keywords: Archaeological Documentation, Semantic Tools, Ontologies.

1. Introduction

Free text information, produced or used during the archaeological excavation activity, is a class of information that comprises a wide range of documents, from the texts of the ancient sources to the diaries and notes collected during the excavation activity, usually very difficult to process using traditional forms and relational databases (CRESCIOLI *et al.*, 2002; HODDER , 2002).

The system described in this paper tries to approach this problem from a totally different point of view, putting the main focus 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 the efficiency in information retrieval. Performance and usability of the system is guaranteed by state-of-the-art tools and technologies, while data interoperability and integration is ensured by the use of well known standards, like CIDOC-CRM and RDF for data encoding (HOLMEN and ULEBERG, 1996).

The management system is the fruit of the collaboration between the computer experts of PIN (Italy), STARC at the Cyprus Institute, and the archaeologists of the Mediterranean Archaeological Research Institute (MARI) of the Vrije Universiteit Brussel (Belgium). The framework will also be released as part of the 3D-COFORM European project. The final application will be for a general purpose. At present, it is applied to a case-study concerning a 35 year-long Late Bronze Age excavation in Cyprus (the Hala Sultan Tekke excavation documentation). Furthermore, it will also be used in another project about medieval archaeology and architecture in the Aegean Islands.

2. System Overview

The system is composed of 3 different tools combined together to create an archive of semantically annotated text documents that can query the archive in many ways:

1. SAD, a featured container for the semantic and textual information;

2. AnnoMAD, a web Graphical User Interface for the editing and annotation of documents and for the creation of semantic relations;

3. An extended version of the Quantum GIS open source application, an advanced GIS application used for the management of the geographical information related with the archaeological excavation, extended with some plugins for the semantic encoding of spatial data (D'ANDREA *et al.*, 2009; FELICETTI, 2006).

2.1. The core: SAD

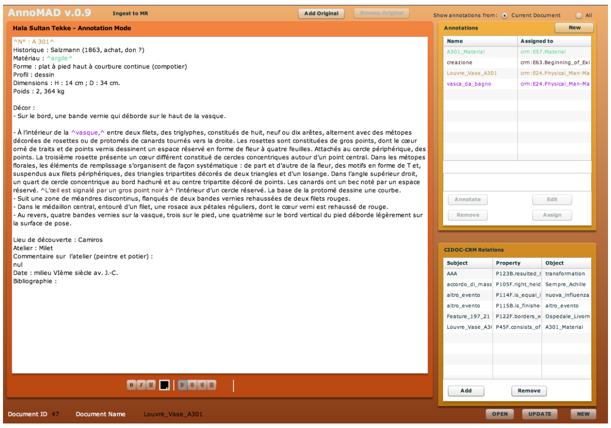
The whole framework is based on a Semantic Archive (SAD) built to store the digital version of all the documents and the related semantic annotations. SAD is based on MAD, a toolset developed at PIN during the EPOCH project and has already been tested and used in many different applications. SAD is the semantic framework of MAD, entirely written in PHP and which takes advantage of ARC2 RDF classes for PHP, an open source set of PHP libraries specifically oriented towards the management of semantic information and query features. SAD made it possible to store and manage every kind of HTML/XML based information, including RDF and GML spatial data. Information stored in the archive is available for every kind of operation or request. The Semantic Archive also supports many popular communication protocols such as REST, SOAP, JSON, and provides SPARQL query features to retrieve information in a semantic way.

2.2. AnnoMAD

AnnoMAD is a complete and easy to use web interface written using the Adobe Flex framework and developed for the annotation and the semantic enrichment of archaeological texts, printed or type-written sources and manuscripts. The main purpose of the AnnoMAD interface is the creation of semantic archives using the annotation paradigm. The tool interface was developed using Adobe Flex, a versatile software development kit for the development and deployment of cross-platform and cross-browser rich internet applications using the Adobe Flash browser plugin. Thanks to this, the application can run in any kind of browser without any of the compatibility bugs that usually affect Javascript and Ajax applications. AnnoMAD is also designed to support a wide range of communication protocols, and can be easily deployed in many standard client/server scenarios. AnnoMAD is perfectly integrated with SAD. Client/server communication is guaranteed by the SOAP protocol through which it is possible to manage the data stored into the Semantic Archive (SAD) and to perform query and retrieval operations. SOAP is also used for receiving and showing query results formatted according to the user's preferences.

Anyway, AnnoMAD is not solely binded to SAD, but can also be used as a client for obtaining and retrieving semantic information to/from different containers and databases, including Digital Libraries. Some successful tests have been carried out by interfacing AnnoMAD with the RDF Metadata Repository, developed for the 3D-COFORM project, to send and retrieve annotated texts with their semantic layer of information, using standard communication protocols. Within the framework of the same project, we are also extending our application by integrating it with the Multilingual Framework, a 3D-COFORM set of libraries and services able to implement interface localization and language-sensitive access to metadata, entities and relations, and to support multilingual querying and annotation.Basically AnnoMAD is intended to allow users to put a conceptual layer on top of the formatted text, either during the editing of the text itself or after the document has been finalized: the user selects a document or a portion of text and assigns it an annotation. An *annotation*, in our view, is a piece of semantic information that can be attached to a whole document or a portion of it, in order to create a semantic description and to specify its meaning in a formal way.Semantic relations with other pieces of annotated texts can also be established using the properties provided by the chosen ontology. Furthermore it is possible to assign an existing annotation to different pieces of text or documents (NICCOLUCCI *et al.*, 2009).

The current version of AnnoMAD also provides some features for the definition of global annotations for the most common entities and concepts. This particular class of annotations is created as a set of instances of specific ontology entities at a global level, rather than starting from a specific document. Once defined, they remain available for the whole system and can be reused many times and in many documents. This category of annotations may include, for instance, common entities, like the ones related to the archaeologists involved in the excavation (actors) or to the periods for artifact dating (e.g. "Bronze Age", "Iron Age" and so on).At the end of the annotation process, the edited documents, the related annotations and relations are stored into the semantic archive and are immediately ready to be browsed and queried. A wide set of query features is actually provided by the AnnoMAD interface itself. It includes simple queries, that is just a free text search in the whole archive, and semantic queries, for which it is possible to specify relevant entities and relationships. Another important feature is the internal faceted browsing interface that allows users to navigate and slice the semantic information stored in the SAD container (Figure 1).



^N° : A 301^
Historique : Salzmann (1863, achat, don ?)
Matériau : ^argile^
Forme : plat à pied haut à courbure continue (compotier)
Profil : dessin
Dimensions : H : 14 cm ; D : 34 cm.
Poids : 2, 364 kg

Décor :
- Sur le bord, une bande vernie qui déborde sur le haut de la vasque.

- À l'intérieur de la ^vasque,^ entre deux filets, des triglyphes, constitués de huit, neuf ou dix arêtes, alternent avec des métopes décorées de rosettes ou de protomés de canards tournés vers la droite. Les rosettes sont constituées de gros points, dont le cœur orné de traits et de points vernis dessinent un espace réservé en forme de fleur à quatre feuilles. Attachés au cercle périphérique, des points. La troisième rosette présente un cœur différent constitué de cercles concentriques autour d'un point central. Dans les métopes florales, les éléments de remplissage s'organisent de façon systématique : de part et d'autre de la fleur, des motifs en forme de T et, suspendus aux filets périphériques, des triangles tripartites décorés de deux triangles et d'un losange. Dans l'angle supérieur droit, un quart de cercle concentrique au bord hachuré et au centre tripartite décoré de points. Les canards ont un bec noté par un espace réservé. ^L'œil est signalé par un gros point noir à^ l'intérieur d'un cercle réservé. La base de la protomé dessine une courbe.
- Suit une zone de méandres discontinus, flanqués de deux bandes vernies rehaussées de deux filets rouges.
- Dans le médaillon central, entouré d'un filet, une rosace aux pétales réguliers, dont le cœur verni est rehaussé de rouge.
- Au revers, quatre bandes vernies sur la vasque, trois sur le pied, une quatrième sur le bord vertical du pied déborde légèrement sur la surface de pose.

Lieu de découverte : Camiros
Atelier : Milet
Commentaire sur l'atelier (peintre et potier) :
nul
Date : milieu VIème siècle av. J.-C.
Bibliographie :

Name	Assigned to
A301_Material	crm:E57.Material
creazione	crm:E63.Beginning_of_Exi
Louvre_Vase_A301	crm:E24.Physical_Man-Ma
vasca_da_bagno	crm:E24.Physical_Man-Ma

Subject	Property	Object
AAA	P123B.resulted_i	transformation
accordo_di_mass	P105F.right_held	Sempre_Achille
altro_evento	P114F.is_equal_i	nuova_influenza
altro_evento	P115B.is_finishe	altro_evento
Feature_197_21	P122F.borders_w	Ospedale_Livorn
Louvre_Vase_A3(	P45F.consists_of	A301_Material

Document ID 47 Document Name Louvre_Vase_A301

Figure 1: *The AnnoMAD user interface.*

2.3. The extended Quantum GIS

The third component of our framework is a set of plugins for Quantum GIS, a popular GIS software that is currently also used for the creation of the Geographic Information System for the archaeological site of Hala Sultan Tekke. Quantum GIS is a very easy to use and rather powerful multi-platform Open Source GIS, developed by the Open Source Geospatial Foundation (OSGeo) to visualize, manage, edit, analyze data, and compose printable maps. Quantum GIS has an easy to use user interface and an extremely flexible modular architecture, composed of a set of core functions and plugins that allows a high degree of personalization (see Figure 2). The Python language can be used to create new plugins or to extend existing ones.

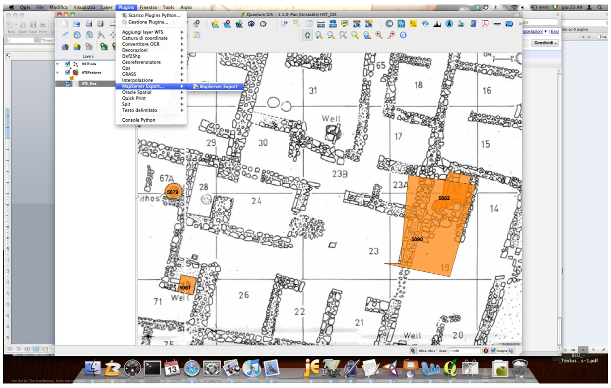


Figure 2: *Quantum GIS in action with the Hala Sultan Tekke geographic information.*

3. Ontologies

For encoding the texts and the geographic data and for creating the annotations and their relationships, we used the entities and the properties provided by the CIDOC-CRM ontology that perfectly fits our needs. CIDOC-CRM has specifically been created for the encoding of Cultural Heritage documentation.

CIDOC-CRM is an extensive and very complex ontology. To make the annotation process more intuitive for the archaeologists supposedly are going to use the application, we created a domain ontology defining domain specific entities mapped on a subset of CIDOC-CRM. Even though our system allows users to use entities and relations "as a" from CIDOC-CRM core. The core ontology was agreed upon by the archaeologists and the applications developers at the very beginning of the development process. Many extensions of the domain ontology belonging to the case under study are, however, always possible to add later on, during the encoding process.

4. The Hala Sultan Tekke excavation site

The development and the testing of our system is mainly focused on the management of the wide and complex documentation produced during the excavation at Hala Sultan Tekke. This archaeological site is located in south-eastern Cyprus and has been extensively excavated from 1971 to 2005 by Prof. Paul Åström. Since 2001, Prof. Karin Nys has been the Assistant Director and from 2009 she is in charge of all the archaeological research. The excavation in the central Area 8 exposed a large Late Bronze Age harbor town. Other excavated areas revealed additional parts of the town, as well as tombs and wells. The town plan consists of seven building units arranged along a wide street that runs from North to South. The finds testify the daily life, the various trade relations and the involvement in the copper industry. The site is one of the Cypriot urban polities that actively participated in the Eastern Mediterranean exchange network. The oldest remains date back to the end of the Middle Bronze Age (ca. 1600 BCE). The greater part of the excavation however, reveals the last phases of the settlement, just before its final abandonment at the end of the Late Bronze Age (ca. 1110 BCE) (ÅSTRÖM *et al.*, 1975).

The excavation campaigns were documented in a full-text, mainly hand written archive, including feature forms, find forms, pottery forms, field notes, artifact drawings and maps, as well as analogue photos. The archaeological material is, for the most part, kept in the storerooms of the Larnaca District Museum. Despite the publication of preliminary reports in 12 volumes, much of the post-excavation analysis of, for example, stratigraphy, architecture, pottery and other artifacts has yet to be performed. Our primary concern was to process the Hala Sultan Tekke excavation archive in an optimal way. This involves using a tool that must store and correlate digital versions of the archive's content, and that is easily accessible for every user. The traditional methods in combination with a relational database could not fulfill the requirements dealing with the complexity of the excavations dataset (NICCOLUCCI and D'ANDREA, 2006).

5. Development process and user approach

Our system is mainly intended to be used by archaeologists and Cultural Heritage experts, who can understand the meaning of the original source content and therefore solve any interpretation problem. Text encoding may take place at the same time as digitization or at a later stage. In order to achieve all database requirements, it is necessary that the computer system designer and the archaeologist have a constructive dialogue to modify and improve the system. A beta-version of the Hala Sultan Tekke database was provided by the computer system designers and discussed with the archaeologists. Through a constant dialogue between the user (archaeologist) and the designer (IT-specialist), it was possible to fix bugs, system weaknesses and conceptual problems and to insert extra or new requirements. The implementation of solutions and new features will lead to an improved version and to the development of a general-purpose tool that will be used for every kind of textual documentation, not only archaeological, using any ontology required by users.

Conclusions and further work

The final goal of our work is the release of a general-purpose tool indifferent to non homogeneity of sources, but flexible enough to manage textual sources that can also be very different from one another, and encompass archaeological records, historical sources, and direct observations. The wide use of standards at any level of the encoding operations will guarantee interoperability, even with data coming from different contexts, like geographic information and structured data from databases and Digital Libraries. Even though the current version of the system supports only CIDOC-CRM based ontologies, the final version will guarantee support for every standard encoded ontology, allowing users to import and use other conceptual schemas.

References

ÅSTRÖM P. *et.al.* Hala Sultan Tekke 1-12. (*Studies in Mediterranean Archaeology* 45: 1-12), Göteborg, Jonsered, Sävedalen, 1975-2007.

CRESCIOLI M., D'ANDREA A., NICCOLUCCI F., 2002: XML Encoding of Archaeological Unstructured Data. In Archaeological Informatics: Pushing the Envelope. *Proceedings of CAA2001* , Burenhult G., (Ed.), Archaeopress, Oxford, pp. 267–275.

D' ANDREA A., FELICETTI A., LORENZINI M., PERLINGIERI, 2009: C. Spatial and non-spatial archaeological data integration using MAD. In Layers of Perception. *Proceedings of CAA2007* , Posluschny A., Lambers K., Herzog I., (Eds.), Archaeolingua, Budapest.

D' ANDREA A., MARCHESE G. , ZOPPI T., 2006: Ontological Modelling for Archaeological Data. In *VAST: 7th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage* , Ioannides M., Arnold D., Niccolucci F., Mania K., (Eds.), Eurographics Association, pp. 211–218.

FELICETTI A. MAD, 2006: Managing Archaeological Data. In *The evolution of Information and Communication Technology in Cultural Heritage* , Ioannides M., Arnold D., Niccolucci F., Mania K., (Eds.), Archaeolingua, Budapest, pp. 124–131.

FELICETTI A., LORENZINI M., 2007: Open Source and Open Standards for Using Integrated Geographic Data on the Web. In *VAST: 8th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage* , Arnold D., Niccolucci F., Chalmers A., (Eds.), Eurographics Association, pp. 63–70.

HODDER, I, 2002: *The Archaeological Process: An Introduction*. Wiley-Blackwell, Oxford.

HOLMEN, J. and ULEBERG, E. SGML-encoding of archaeological texts. http://www.dokpro.uio.no/engelsk/text/getting_most_out_of_it.html.

NICCOLUCCI F., D'ANDREA A., 2006, An Ontology for 3D Cultural Objects. In *VAST: 7th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage* (, Ioannides M., Arnold D., Niccolucci F., Mania K., (Eds.), Eurographics Association, pp. 203–210.

NICCOLUCCI F., FELICETTI A., HERMON S., NYS K., 2009, Managing Full-text Excavation Data with Semantic Tools. In *VAST: 10th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage* , Debattista K., Perlingieri C., Pitzalis D., Spina S., (Eds.), Eurographics Association, pp. 125–132.