The 9th International Symposium on Virtual Reality, Archaeology and Cultural Heritage
VAST (2008)
M. Ashley, S. Hermon, A. Proenca, and K. Rodriguez-Echavarria (Editors)

Semantic Web, Digital Libraries and the future of Cultural Heritage

A. Felicetti[1] and H. Mara[2]

[1,2]PIN, University of Florence, Italy

Abstract

This paper provides an overview regarding the application of the Semantic Web oriented technologies we have developed as part of the EPOCH and AMA projects for Cultural Heritage Digital Libraries. We wanted to enhance interoperability among diverse archives and to make disperse digital information available through the web in a standard format. Our toolset includes an application for mapping existing archive schemas to ontology schemas (AMA Mapping Tool), a tool to recursively markup unstructured text documents (AMA Text Tool) and a Semantic Web Database able to store, query and return simple and complex semantic information (MAD). We used the CIDOC-CRM core ontology to define the entities we dealt with and to describe concepts and relations among them.

Categories and Subject Descriptors (according to ACM CCS): H.3.5 [Information Storage and Retrieval]: Web-based Services

1. Introduction

Digital Libraries provide access to large amounts of resources in the form of digital objects and a huge variety of tools to search, browse and use the digital content in many ways. Complex and flexible systems of metadata schema have also been developed and used to describe digital objects in collections (i.e. Dublin Core, MODS, METS).

Unfortunately every system and metadata schema has a closed structure and uses different approaches to deal with digital data. Thus every system tends to become a wonderful dream island, rich in wonders to explore and treasures to discover. Most of the time, however, if one does not possess the right map, discovery and exploration are difficult.

The European Commission is funding numerous projects related to Digital Libraries to enhance the availability, usability and the long term preservation of relevant information of any kind.

The biggest challenge will be to guarantee all these features in the future. This will occur, in our opinion, through the implementation of strong semantic layers settled on top of Digital Libraries to unify the description of objects and concepts coming from different digital archives.

Semantic Web tools allow digital library maintainers to link their schema and automatically translate their terms, expanding mutual comprehensibility. The semantic layers will be the Treasure Maps allowing users to discover Treasure Islands: any "sailor" can easily find his way to the information if the "coordinates system" of his map is designed in a uniform, logic and accurate way and if "X" clearly marks the spot.

This paper tries to outline the importance of using Semantic Web technologies to enhance the usability and interoperability of Digital Libraries. It describes the tools and the technology we developed as results of our research on standards and semantic web technology application in the framework of the EPOCH [EPO] and AMA [AMA] projects to implement semantic capabilities on digital data.

2. Semantic integration

Digital Libraries are currently very complex entities. It is often difficult to explain what they actually are, but we could think of them as big indexes designed to serve as guides for retrieving and returning digital information stored on the Web in different formats and archives.

The first problem to face is that every guide has its retrieval system and uses a metadata grammar to describe and index data so specifically that it would never work on other systems. None of these metadata systems can analyze all the information on the Web, unless we make them available through a machine understandable format using RDF [RDF].

The second relevant problem concerns the information itself: the huge variety of formats used to index data is a big obstacle to integration and must be seriously analyzed. Even if we limit our efforts solely to Cultural Heritage archives (i.e. databases of museums and collections, archaeological excavation records, reports and other unstructured data) we have to admit that information is as dispersed as the material culture it refers to.

To create a uniform conceptual layer, semantic information should be extracted from databases, HTML pages, descriptive texts, metadata tags and put into a standard format in order to capture the conceptual meanings and to create correspondences at a higher level (conceptual mapping).

These operations are facilitated today by the constant activity of the W3 Consortium in defining new standards for web information encoding, such as RDF and SPARQL, by the use of ontologies such CIDOC-CRM [CDGS], explicitly created for Cultural Heritage and by the power of the available semantic tools, able to physically extract semantic information in a semi- automatic way. [NF07]

Once the conceptual layer for both data and metadata is ready, the semantic information will be stored in a container based on RDF and ontologies. This will be the real integration place, where unified conceptual information from different digital archives will be browsed and searched as a unique Digital Library. The RDF language is robust and flexible enough to guarantee interoperability and to provide a common denominator not only among Digital Libraries, but also with other systems and services.

3. Creating the Semantic Layer

3.1. Mapping

For every adventure, for every exploration, or any kind of quest for treasures, what is really needed is a map showing us how to reach the island and where to find the buried gold. Similarly the creation of logical maps is one of the most important activities towards data integration, even if the mapping process requires uncommon skills and knowledge among those who are supposed to do the job.

The mapping process consists of defining the elements of a given data schema (starting schema) using the entities provided by another data schema (target schema) in order to establish a direct correspondence between the two different element sets. Usually the mapping occurs between a closed or personal data schema and a more general and widely used standard.

The standards presented in the past were not widely accepted by the community. Culture professionals and heritage policy makers found it difficult to map the existing data structures to them, which effectively impeded the preservation of existing archives. In fact, it is not the absence of a facilitating tool, but the existence of practices, legal obligations and the lack of a clear motivation that has delayed or reduced the creation of such mappings to a handful cases . Many mapping templates already exist among common ontologies and metadata standards used for Digital Libraries (i.e. Dublin Core and FRBR metadata schema are already mapped to CIDOC-CRM ontology) and many others will be interconnected to create graphs of concepts that can leverage the semantic use of digital content [DL07].

In addition, the namespace mechanism provided by the RDF syntax can be used to extend existing ontologies by incorporating concepts coming from other schema to describe, for instance, 3D objects or geographic entities [FL07]. In the following RDF example, a GML polygon object has been used to extend the CIDOC-CRM information object describing a place:

```
<crm:E53.Place rdf:about="US1020">
 <crm:P67B.is_referred_to_by>
  <crm:E73.Information_Object
rdf:about="gmlModel_US1020">
   <gml:Polygon srsName="osgb:BNG">
    <gml:outerBoundaryIs>
     <gml:LinearRing>
      <gml:coordinates>
         278534.100,187424.700
         278529.250,187430.900
         278528.700,187431.650
         278527.250,187433.600
      </gml:coordinates>
     </gml:LinearRing>
    </gml:outerBoundaryIs>
   </gml:Polygon>
  </crm:E73.Information_Object>
 </crm:E53.Place>
</rdf:RDF>
```

3.2. The AMA Mapping Tool

To make the mapping process easy and accessible, we created the AMA Mapping Tool, a flexible tool developed as part of the AMA Project to facilitate the mapping of different archaeological and museum collection data models (with various structured, as well as non-structured data, i.e. text description) to a common standard based on CIDOC-CRM ontology.

The tool is based on the concept of "template", this being the instantiation of the abstract mapping between the source data structure and the standard one. Templates capture the semantic structure, as well as the intellectual content, of the sources and their transformations.

The AMA Mapping Tool comes with a web application

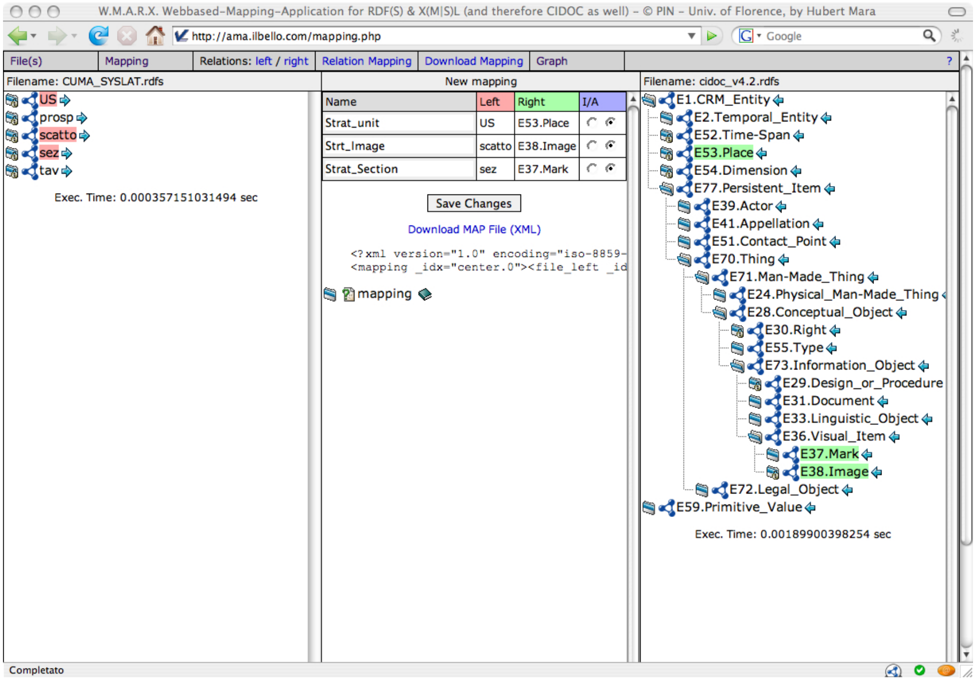

Figure 1: *The AMA Mapping Tool web interface in action to create mapping templates.*

developed according to the open source principle (see Figure 1). We chose to implement the tool in PHP5 using the DOM extension for parsing XML.

A common request made by our EPOCH partners was to develop in the widely used PHP4, but for future sustainability we decided against it, as its end of life was announced on July, 13th 2007. Even though the capabilities of XML and XSD already allow a vast number of possibilities for database schema, we focused on adding support to RDF-Schema (RDFS).

Therefore we choose to use the RAP - RDF API for PHP - of the Freie Universitat Berlin, Germany [RAP], which is available under the GNU Lesser General Public License (GPL). Another reason to choose RAP, besides GPL and functionality, is the long- term existence and regular update policy of its authors. As PHP and RAP are used on (web-)server-side you need only a HTML-browser to access it - no Java(script) or any other plug-in, Flash or ActiveX is required.

To use the AMA tool you must upload a source and a target schema in any XML format, while RDFS is recommended, at least for the target schema (e.g. CIDOC- CRM), due to its semantic nature and its easier mapping capabilities. The next step is mapping pairs of classes (1:1).

If it is necessary to map classes on 1:n or n:m ratio,this can be achieved by giving the same name to sets of pairs. Furthermore you can specify if a mapping is an alias (A) or an inheritance (I).

The third and final step of the mapping is the definition of the relations between classes in the target schema using the mapped class properties. This can be achieved either by choosing the property/range and then the object/class or vice versa: choose the object/class and then the property/range connecting them.

The application also provides the possibility to add new classes, in case the existing ones are not sufficient to define a relation. You can download your mapping schema as XML containing the mapping and the relations of classes. In addition a graphical representation of mapping and relations can be generated and directly browsed on-line.

The final result of the AMA mapping process is the creation of a mapping file to be used as a template for the data conversion process, directly in the original databases or along with conversion tools like D2R. Semantic data extracted from archives will be ready for integration with other CIDOC-CRM compliant archives and for use in Semantic Web contexts.

3.3. Getting semantic information from databases

D2R is a tool providing a powerful set of features for real time conversion and query without modify legacy RDBMS. The D2R tool can be used for publishing our relational databases on the Semantic Web, enabling RDF and HTML browsers to navigate the content of the database and allowing applications to query the database. SPARQL is the query language used.

Then a customizable mapping framework called D2RQ is used to map database content into a set of virtual RDF datasets that can be browsed and searched by semantic web applications. Requests from the Web are rewritten into SQL queries via the mapping. This on-the-fly translation allows to publish RDF from large live databases and eliminates the need for replicating the data into a dedicated RDF triple store.

D2R Server's Linked Data interface makes RDF descriptions of individual resources available over the HTTP protocol. An RDF description can be retrieved simply by accessing the resource's URI (or URL) over the Web. Using a Semantic Web browser like Tabulator or BrownSauce you can follow links from one resource to the next, surfing the Web of Data.

D2R Server uses the D2RQ Mapping Language to map the content of a relational database to RDF. A D2RQ mapping specifies how resources are identified and which properties are used to describe the resources.

The main advantage in using D2RQ is the possibility to customize the mapping by using elements from widely accepted ontologies. The mapping file can be edited with any text editor or automatically generated by modifying the AMA Tool middleware mapping files [D2R].

3.4. Unstructured documents: AMA TextTool

Another important step towards the process of data integration in Cultural Heritage concerns the encoding of free texts made available for processing by semantic engines.

Dealing with this kind of document means dealing with unstructured information that would be impossible to extract and put in a semantic format using fully automatic procedures. Most of the data extraction or encoding is usually carried out manually by reading and keying or marking up text information using XML tags.

However semi-automatic tools can assist the users during the encoding process and simplify their work by providing control mechanisms able to validate the markup and manage co-references in the text.

The Unit for Digital Documentation (EDD) of the University of Oslo, also involved in the AMA Project, has developed the AMA TextTool, a semi-automatic tool aimed at speeding up and improving the encoding process of archaeological texts in CIDOC-CRM. The tool is written in Java to be cross platform and is based on concepts and techniques from computational linguistics.

The AMA TextTool implements a KWIC (Key Word In Context) concordance tool directly connected to the electronic text(s) used to find a word, a phrase or a pattern of words and possibly XML-markup already in the text.

Users can then analyze the text and mark which occurrences in the KWIC concordance they want to tag. The system then inserts the mark up in the text file(s). This semi-automatic "search and replace" feature makes it possible for the user to create and include new algorithms, both for text retrieval and for text mark up, and new features of electronic text processing.

The AMA TextTool includes functions to enrich texts with a TEI-header, providing bibliographical information about the original text and the current electronic one, and structural markups to identify, for instance, chapters and paragraphs.

The structural information can then be extended by adding the semantic layer with the identification and markup of the conceptual elements present in the texts (actors, objects, places, events). The latter operation can be accomplished using the set of tags and elements provided by ontologies like CIDOC-CRM.

After the tagging process is complete, the documents are ready to use in different Digital Libraries, both for their structural and semantic encoding. It is vital in this case to preserve the textual nature of this kind of document, avoiding the need to extract relevant pieces of information to be stored in relational databases. This operation would cause the destruction of the original document flow.

HTML and TEI are suitable formats to make them accessible and easy to share over the Web. What can be extracted instead is the conceptual information to be used for the description of the documents' semantic maps and indexes. Conceptual descriptions can then be written using the RDF syntax and shared with other semantic information coming from databases and linked to the original documents stored in Digital Libraries using the URI mechanism [EFO*08].

4. MAD: a Semantic Web Database

4.1. Overview

A powerful container is needed to manage the complexity arising from the data encoded using ontologies and the RDF syntax. It must deal with semantic data, just as relational databases deal with data stored in tables and records and with their relationships.

A Semantic Database is needed for this, a big container where semantic information is stored and maintained to provide users with the necessary tools for their semantic search. For this purpose we created MAD (Managing Archaeological Data), a framework originally designed to manage structured and unstructured archaeological excavation datasets encoded using XML syntax, including free text documents marked up in XML.

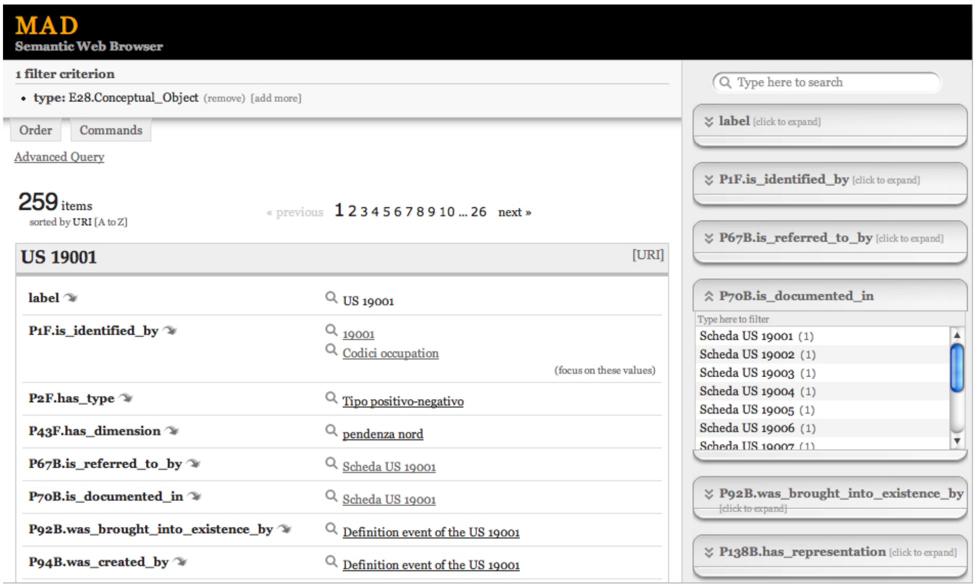

Figure 2: *MAD Semantic Browser Interface with faceted browsing capabilities showing entities and properties of the CIDOC-CRM*

The latest release of MAD comes with a multipurpose semantic engine able to store and manage ontology encoded information, i.e. data structured in CIDOC-CRM compliant format, a semantic query set of interfaces based on SPARQL and RQL query languages and a Firefox plug-in implementing a semantic browser for RDF graphs . MAD can be used to store, browse and query semantic data in many powerful ways, but also to transform and supply semantic information on demand. The whole framework has been developed as part of an EPOCH activity for creation of information management systems for the Semantic Web and is entirely based on Open Source software, XML standards and W3C technology [Fel06].

4.2. The MAD Semantic Web Database

MAD is built around a powerful Java native XML Database providing technology to store and index XML documents in a file-system-like structure of folders and subfolders (collections).

This container can be used to store information coming from mapped and digital metadata and content, along with annotations and tag sets created by users, mapping templates, schema, ontologies and everything can be expressed using the RDF language.

Users can browse and query this integrated archive to get semantic information on Cultural Heritage objects described herein, or references to remote digital objects stored elsewhere (i.e. URIs and URresources). In this sense MAD acts as a Semantic Digital Library.

4.3. Semantic Queries in MAD

In order to query RDF data we are using the two most important languages available at present: SPARQL [SPA] and RQL [KAC*02], two semantic query languages designed for retrieving information from RDF graphs. These languages provide a clear, easy query syntax and the ability to obtain information from encoded documents without knowing its explicit syntactical structure (i.e. elements and properties names).

RQL is used due to its ability to combine schema and data querying using advanced pattern-matching facilities. SPARQL and RQL combined have been used for the creation of a group of semantic query interfaces able to parse the RDF documents in different ways.

The ability of RQL to query ontology structures make the retrieval of classes, subclasses and properties from the models very simple allowing the building of structure-based semantic queries.

Classes and properties can be clicked to visualize subclasses and sub-properties and to define the elements that will participate to the query definition. While clicking on the different elements, an RQL query is constructed and then submitted, to be evaluated by the Semantic engine (Fig. 2).

A human readable version of the query is also shown to make it understandable. An RQL query serialized on the CIDOC-CRM structure, for instance, may appear like the following:

```
select $class0
```

```
from  {instance : $class0} @p {value : $class1}
where $class0 in subClas-
sOf( kyme:E28.Conceptual_Object)
and @p = kyme:P70B.is_documented_in
and $class1 in subClas-
sOand value like "*1022*"
```

It is the corresponding machine readable version of the clearer spoken language request: "I am looking for a Conceptual Object(E28) which is documented in(P70B) the Document(E31) containing 1022".

4.4. MAD: The Semantic Web Browser plug-in

The richness of the RDF graph model in which data are distributed, makes it often difficult for users to get effective and meaningful data retrieval when only a simple or complex query interface is used, particularly when the graph grows in dimensions and complexity . Sometimes it would be easier and faster to browse the multidimensional graph structure allowing users to chose a starting point and move along different paths through the graph to reach the desired data. To allow this kind of data navigation we have developed a Mozilla Firefox plug-in based on SIMILE technology [SIM].

The plug-in turns the Firefox browser into a semantic browser able to open, search and save RDF datasets. Browsing is based on the faceted browsing UI paradigm. A facet in this view is a particular metadata element considered important for the dataset browsed. Users can select a starting point which they consider relevant for their search.

The browser extracts a list of facets, their values, and the number of times each facet value occurs in the dataset. Then it's possible to add or remove restrictions in order to focus on more specific or more general slices of the model.

A "free text" restriction can be also added to reduce the browsed dataset to all items that contain the required string (Fig. 3). The interface was also configured to interact with the MAD container: all the semantic information stored therein can be browsed in order to retrieve relevant semantic objects and references to external resources. RDF resources on the Web can also be visualized by providing their URLs and saving them in the MAD Semantic Database.

We have tested the MAD framework to build an on-line version of the archaeological dataset recorded during the excavation of the ancient city of Cuma, containing information on stratigraphical units and other related resources.

We are also using MAD with the AMA toolset to create an on-line application for the complete semantic management of coin collections for the COINS Project [COI]. All this information is stored in RDF format and is ready to be queried, integrated and shared in the Semantic Web framework [MAD].

5. The Arrigo VII funerary monument semantic index

The Arrigo VII Mausoleum is a multimedia application created by the CNR of Pisa for the Museum of the Cathedral of Pisa. It reconstructs the history and the form of the funerary complex of the Holy Roman Emperor Henry (Arrigo in Italian) VII of Luxembourg, who died near Siena in 1313 and was buried in Pisa [BBC*04].

The Arrigo application has been included in the Interactive Salon, a touring exhibition created for the EPOCH Project and used for the creation of a demo shown during the EPOCH final event in Rome [ARF].

The application combines 3D reconstructions of the monuments, 3D models of the statues still present in the Cathedral of Pisa, multimedia content including sounds and movies and descriptions, either in free text or hypertext formats.

This case study was created to demonstrate the possibility of creating connections among digital objects, free texts and semantic information and was built in collaboration with the University of Leuven and the University of Graz.

To build a test Digital Library, 3D objects were extracted from the application, digitally annotated and stored in the Fedora container of the University of Leuven. In the same container we placed texts containing the descriptions of the various elements, marked up using the AMA TextTool to generate a set of TEI encoded HTML documents and the RDF CIDOC-CRM encoded information describing the content.

All the RDF from digital annotations of 3D objects and from tagged texts were stored in the MAD Semantic Database to build a semantic index (semantic map) of our digital archive. Semantic information and digital objects were identified using the URI mechanism and linked via specific URLs.

Once the Digital Library and the Semantic Index are created, it is possible to retrieve digital objects by simply using the semantic interfaces provided by MAD to query the RDF information.

It is possible, for instance, to find an object (i.e. a statue) made of marble and composed of different parts (hands, shoulders, head and so on), identified by an appellation and positioned in a specific place (i.e. a corner of the Cathedral of Pisa). The result of this query would be the RDF description of this entity (the statue), its URI and the URL pointer to the digital object stored in our Digital Library.

6. Conclusion and future work

Right from the start, the Semantic Web encountered a lot of resistance from developers and users, mainly for the lack of knowledge in this field and for the absence of applications that put this vision into practice. During the last years Semantic Web technologies have become mature enough to be used in a fruitful way and we are now able to take advantage of the power provided by them. The Arrigo case study demonstrated how simple the integration can be. The tools we developed proved to be powerful and flexible enough to be used for enhancing interoperability in different context, either in pure Semantic Web implementations or in Digital Libraries scenarios.

Thanks to the CIDOC-CRM encoding of Arrigo's semantic information, the integration with information coming from other archives, already stored in MAD, was natural and immediate, as it is very simple to integrate elements when they are represented at a conceptual level (physical objects, places, actors and so on) [SFPH]. The door towards Digital Libraries integration can now be fully opened.

Future work will focus on enhancing interoperability among our tools to guarantee a more fully integrated and usable system. Currently, for instance, the mapping files created by AMA needs further processing operations to be used for real data conversion. For these reasons we are going to provide MAD with a mapping framework able to connect the MAD engine directly to legacy databases and to extract information automatically, according to the high level mapping files created using the AMA Mapping Tool. Other important improvements will concern the interfaces allowing users to perform more intuitive queries on semantic data, thus reducing the gap among query languages and natural languages.

References

[AMA] AMA, Archive Mapper for Archaeology. http://www.epoch.eu/AMA/.

[ARF] Arrigo in fedora: Epoch project common infrastructure tool chain demo. http://partners.epoch--net.org/common_infrastructure/wiki/index.php/Arrigo_in_Fedora.

[BBC*04] BARACCHINI C., BROGI A., CALLIERI M., CAPITANI L., CIGNONI P., FASANO A., MONTANI C., NENCI C., NOVELLO R., PINGI P., PONCHIO F., SCOPIGNO R.: Digital reconstruction of the Arrigo VII funerary complex. *In: Dieter W. Fellner, Stephen N. Spencer (Eds.): VAST 2004: The 5th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage. Brussels and Ename Center, Oudenaarde, Belgium, December 7-10* (2004).

[CDGS] CROFT S. N., DOERR M., GILL T., STEAD S.: Definition of the CIDOC Conceptual Reference Model. http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.doc.

[COI] COINS: Combat On-line Illegal Numismatic Sales. http://www.coins-project.eu.

[D2R] D2R Server. Publishing relational databases on the Semantic Web. http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/.

[DL07] DOERR M., LEBOEUF P.: Modelling Intellectual Processes: the FRBR CRM Harmonization. *First DELOS Conference on Digital Libraries, February 2007 Tirrenia, Pisa, Italy* (2007).

[EFO*08] EIDE O., FELICETTI A., ORE C., D'ANDREA A., HOLMEN J.: Encoding Cultural Heritage Information for the Semantic Web. Procedures for Data Integration through CIDOC - CRM Mapping. *In press* (2008).

[EPO] EPOCH - European Network of Excellence in Open Cultural Heritage. http://www.epoch.eu.

[Fel06] FELICETTI A.: MAD: Managing Archaeological Data. *In: M. Ioannides, D. Arnold, F. Niccolucci, K. Mania (eds.): The e-volution of Information Communication Technology in Cultural Heritage. The 7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST 2006), Nicosia, Cyprus, 30 October - 4 November* (2006).

[FL07] FELICETTI A., LORENZINI M.: Open Source and Open Standards for using integrated geographic data on the web. *In: D. Arnold, A. Chalmers, F. Niccolucci (eds.) Future Technologies to Empower Heritage Professionals. The 8th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST 2007). Brighton 26-30 November* (2007).

[KAC*02] KARVOUNARAKIS G., ALEXAKI S., CHRISTOPHIDES V., PLEXOUSAKIS D., SCHOLL M.: RQL: A Declarative Query Language for RDF. *The Eleventh International World Wide Web Conference (WWW'02), Honolulu, Hawaii, USA, May 7-11* (2002).

[MAD] MAD: Managing Archaeological Data. http://www.epoch.eu/MAD/.

[NF07] NICCOLUCCI F., FELICETTI A.: Digital Libraries and Virtual Museums. *In: V. Cappellini, J. Hemsley (eds.) Electronic Imaging and the Visual Arts. EVA 2007 Florence* (2007).

[RAP] RAP: RDF API for PHP. http://www4.wiwiss.fu-berlin.de/bizer/rdfapi/.

[RDF] W3C Semantic Web Activity, Resource Descpriition Framework. http://www.w3.org/RDF/.

[SFPH] SUGIMOTO G., FELICETTI A., PERLINGIERI C., HERMON S.: CIDOC-CRM Spider. Stonehenge as a case study of semantic data integration, year = 2007. *In: D. Arnold, A. Chalmers, F. Niccolucci (eds.) Future Technologies to Empower Heritage Professionals. The 8th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST 2007). Brighton 26-30 November*.

[SIM] The SIMILE Project. http://simile.mit.edu/.

[SPA] SPARQL Query Language for RDF. http://www.w3.org/TR/rdf-sparql-query/.

All web references have been visited on 15-07-2008.