

Encoding Cultural Heritage Information for the Semantic Web. Procedures for Data Integration through CIDOC-CRM Mapping

Ø. Eide¹, A. Felicetti², C.E. Ore¹, A. D'Andrea³ and J. Holmen¹

¹Unit for Digital Documentation, University of Oslo, Norway ²PIN, University of Florence, Italy ³CISA, University of Naples "L'Orientale", Italy

Abstract

This paper describes the background and methods for the production of CIDOC-CRM compliant data sets from diverse collections of source data. The construction of such data sets is based on data in column format, typically exported for databases, as well as free text, typically created through scanning and OCR processing or transcription.

Categories and Subject Descriptors (according to ACM CCS): H.3.1 [Content Analysis and Indexing]:

1. Introduction

As part of the EPOCH Network of Excellence, several tool chains for cultural heritage have been investigated, and a complete framework for the mapping and management of cultural heritage information in a semantic web context has been developed. PIN (University of Florence, Italy), CISA (University of Naples "L'Orientale", Italy) and EDD (University of Oslo, Norway) have been central within the EPOCH context in these activities through the AMA project.

AMA (Archive Mapper for Archaeology) is one of the NEWTON projects (NEW TOols Needed). EPOCH's NEWTONs aim at plugging gaps in the digital processing of Cultural Heritage on the basis of existing tools or tools under development for Epoch's Common Infrastructure (WP 3.3). One of the main goals is to develop entire scenarios that can handle multiple steps in such processing chains.

2. General Mapping problems in cultural heritage

Access to data, interoperability and standard structure are vital to guarantee the usability and usefulness of Cultural Heritage records, which are digitally archived in a myriad of different ways and are stored separately from each other. In many ways, they are as dispersed as the material culture they refer to.

Proposed standards have not yet overcome the diffidence of culture professionals and heritage policy makers because of the difficulty of mapping existing data structures to them. Mapping requires skills and knowledge which are uncommon among the cultural heritage professionals with the most thorough knowledge of the material to be mapped. It is not only the absence of facilitating tools, but also the existence of practices, legal obligations and the lack of a clear motivation that has as yet delayed or reduced the creation of such mappings to a handful cases.

There have also existed a hope that automatic tools, such as the ones developed in natural language processing research, will solve

the integration problems. Although such tools have been and will continue to be helpful in many areas, solutions to the data integration problems have to be solved at a different level, and human interaction will continue to be necessary.

3. The AMA project

The aim of the AMA project was to develop tools for semi-automated mapping of cultural heritage data to CIDOC-CRM, ISO 21127 [CDG*05]. The reason for this investment was that such tools would enhance interoperability among the different archives and datasets produced in the field of Cultural Heritage [D'A06c]. For this purpose we created a set of tools able to extract and encode legacy information coming from diverse sources, to store and manage this information using a semantic enabled container and to make it available for query and reuse.

The tool set developed in the AMA project includes:

- A powerful mapping application for the creation of mappings from existing datasets
- A tool for mapping cultural heritage information contained in free text into a CIDOC-CRM compliant data model
- Templates describing relations between the structure of existing archives and CIDOC-CRM
- A semantic framework to store, manage and browse the encoded information providing user-friendly interfaces

All tools have been developed as part of the EPOCH and AMA projects and can be freely downloaded from the EPOCH web site with no usage restrictions. The rest of this article describes the background for the tools as well as their implementation and use.

4. Data standards

4.1. CIDOC-CRM

International standards and ontologies for data encoding are crucial to speed up interoperability and the process of integration. CIDOC-CRM, which is created to capture the richness typical of Cultural Heritage information, fully fits our needs: its classes and properties work perfectly to capture the concepts underlying database structures, providing a high level of data integration.

CIDOC-CRM is a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information. It was developed by interdisciplinary teams of experts, coming from fields such as computer science, archaeology, museum documentation, history of arts, natural history, library science, physics and philosophy, under the aegis of the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). The harmonisation of CIDOC-CRM and IFLA's FRBR [FRB98] is completed and have been submitted to IFLA for comments. The EAD has already been mapped to CIDOC-CRM [TD01]. This shows that CIDOC-CRM falls well in with a larger memory institutional framework.

CIDOC-CRM is defined in an object oriented formalism which allow for a compact definition with abstraction and generalisation. The model is event centric, that is, actors, places and objects are connected via events. CIDOC-CRM is a core ontology in the sense that the model does not have classes for all particulars like for example the Art and Architecture Thesaurus with thousands of concepts. CIDOC-CRM has little more than 80 classes and 130 properties.

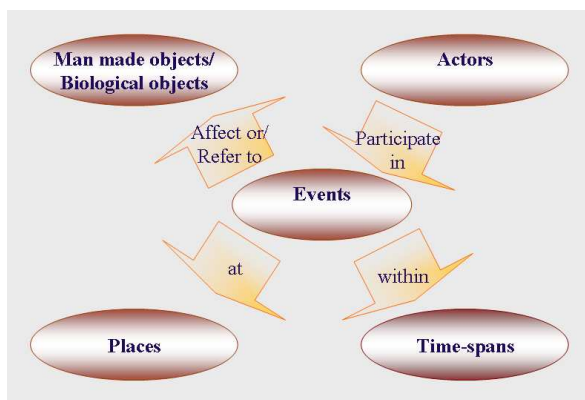


Figure 1: The most central classes and properties for data interchange in CIDOC-CRM.

4.2. TEI

TEI is a consortium of institutions and individuals from all over the world. The TEI is also a set of guidelines for the encoding of textual material, and it is a set of computer readable files. The guidelines and the computer readable files specify a set of rules documents have to adhere to in order to be accepted as TEI documents.

One of the main goals of TEI is to capture a wide range of intellectual work. It is used extensively in areas such as edition philology, but it is quite possible to create a detailed encoding scheme for e.g. archaeological grey documents and include it as a local extension of TEI, creating documents very different from most other TEI documents [EH06].

According to the TEI guidelines [TEI08, section iv.i], there are three primary functions of the guidelines:

- guidance for individual or local practice in text creation and data capture;
- support of data interchange;
- support of application-independent local processing.

TEI is also an important discussion forum, or a set of fora, both through meetings and on-line media, as well as through articles and books. Few encoding problems will be faced that has not been discussed somewhere in the guidelines or in the literature surrounding TEI.

5. From database to CIDOC-CRM

The AMA project was aimed at overcoming the mapping problems through the implementation of a flexible mapping tool in order 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 [D'A06b]. The necessary information was extracted from such mappings to convert individual datasets to a common, CIDOC-CRM compliant structure [DMZ06].

5.1. AMA Mapping Tool

The design of the tool design had to start from an accurate knowledge of the material it handles, that is, the source data structure or, at least, their common matrix as generated by national regulations and most common practices. It also had to be tested on a wide sample of such material.

In principle, the tool had to operate on any preliminary, fragmentary or old archaeological format or museum data (both structured and non-structured), and modern datasets, mapping their structure into a single, standardised CRM-compliant system, easily accessible for investigation by means of a web-based interface. Since CIDOC-CRM deals well with museum collections, while work still had to be undertaken for archaeological excavation data, monuments investigation and landscape analysis, the project also considered these aspects.

The tool is based on the concept of “template”, this being the instantiation of the abstract mapping between the source data structure and the mapping target. A template documents the source data structure and the mapping, using formal XML documents, and automatically supports the conversion of actual datasets when desired. Templates capture the semantic structure, as well as the intellectual content, of the sources and their transformations. They ensure the modularity and future maintenance of the system, in case new or extended versions of the standard are available [D'A06a].

The AMA Mapping tool developed at PIN University of Florence and CISA University of Naples comes with an on-line interface written in PHP through which the mapping template can be defined and exported for further use. Features include the possibility to upload the XML starting schema (i.e. the XML description of the user's databases or an existing ontology), to perform the mapping operation using the simple and intuitive on-line interface and to export the high level mapping files to be used in further data conversion and extraction. The tool also provides the possibility to upload a different target ontology (CIDOC-CRM is the predefined one) and an

advanced set of mapping features for complex mapping definitions, such as for the creation of new entities and properties to represent implicit elements and relations, the creation of shortcuts to simplify the mapping process and a graphic visualisation of simple and complex relations obtained during the mapping process.

High level templates coming out of AMA can be directly imported and used by the MAD framework on information stored in legacy databases in order to get perfectly mapped semantic archives, as described below [AMA].

6. From text to CIDOC-CRM

The objectives of the AMA NEWTON is to create tools to assist the construction of mappings from archaeological archive material, reports, catalogues and databases to CIDOC-CRM. An important step in this process is the extraction of information from free text documents into a well defined format with a predefined mapping to the CIDOC-CRM or to a subset thereof.

There are several strategies to perform such extraction of information from running texts:

1. manual extraction by reading and keying the information of interest into a predefined form
2. encoding the information in the actual text by the use of XML
3. using statistical based topic and summary extracting software

The first strategy has been and still is the most widespread. To have a specialist reading and keying the core data into a registration form seems faster measured by the number of pages processed per day than encoding the content of the text by the use of XML. However, the job requires a scholar/specialist and it is very hard to do the proof reading and virtually impossible to check the correctness of the data at a later stage. It will also be too costly to repeat this process with another focus [JHOong].

The XML-encoding process is time-consuming but can be done by non-specialists working under the supervision of domain experts. It is also easy to track from which part of the original textual documents the information has been taken. Based on the XML-markup it is also possible to create mappings to different database formats, store the text in a free text retrieval system, or convert it to HTML or PDF for publication.

The third method is best suited for finding documents in large document collections and not for identifying highly granulated facts. It is however an interesting field of research.

The Unit for Digital Documentation (EDD) has 14 years of experience with large scale XML (previously SGML) content markup of archaeological texts as well as other categories of text [HOE04]. This is very time consuming and costly to do manually. To speed up the process many simple markup programs have been written for ad hoc use. However, it is hard to write a program that is capable to do non trivial tagging correctly. This is even more difficult than writing program for automatic topic indexing of large document collections, as the latter programs only gives a statistically based indication of the content of the documents. It may be impossible to create software to fully automatise the process; it is certainly out of scope for the AMA project to try to write such software. In the AMA project we are developing a semiautomatic tool based on well known language technological methods for speeding up and improving the encoding process.

6.1. The work process

An XML document conforming with the recommendation of TEI, extended with the necessary elements for marking up archaeologically interesting content, would be the ultimate goal for the mark up of archaeological documents and reports of which large collections exist in many countries. The encoding process can be divided in the following steps:

1. Create an electronic text from the original documents, if necessary — in some cases, the text already exist in digital form
2. Supply the electronic text with a TEI-header with bibliographical information about the original text and the current electronic text (obligatory)
3. Give the text a structural mark up i.e. identify and encode chapters, paragraphs, pagination etc. (optional and ad libitum)
4. Identify and mark up the archaeological information, that is, persons, places, excavations and other events, artifacts etc. The mark up should be done according to a predefined XML grammar designed for type of text in question. Such a grammar, also called a tag set, can be expressed as e.g. a DTD or an XML schema. This could be the extended TEI described above. In any case, the XML grammar must be based on a data model with a mapping to CIDOC-CRM or a subset thereof. “Based on” means that the tags in the text can be used as information anchors from which one can define an extraction from the text into e.g. a database defined according to the data model.

The data model described in step 4 is similar to the template described in section 5.1. The data model is mapped to CIDOC CRM and this serves as the abstract mapping between the source data structure, i.e. the predefined XML grammar, and the mapping target.

The steps 1 and 2 is not in the scope of the AMA text encoding tool. Step 3 may be in the scope, but is not central. Step 4 is the most intellectually demanding process and definitely the most time consuming part. The AMA text tool assist the encoder to speed up this process. In our large scale digitisation projects in Norway the actual mark up work has been done by unskilled low cost persons on work retraining schemas. The mark up has later been refined by using an ad hoc KWIC concordance program and other ad hoc text manipulation tools written in Perl 5. The functionality described below is based on many years experience in using such ad hoc tools.

6.2. The functionality of the AMA text tool

The tool is based on techniques from computational linguistics. A Key Word In Context (KWIC) concordance is a concordance where there is one line per hit and the hits are listed with a piece of the context at the actual hit point. Such a list may look like this (more context is usually included, but the lines are shortened to fit the column width of this paper):

of the second	angel	, these verses composed in
scroll of the	angel	on the left side of the
relating the	angel	statue to the tomb of Henry
backside of the	angels	confirm that they originally

The main feature of the AMA text tool is a KWIC concordance tool directly connected to the electronic text(s). The KWIC concordance tool is used to find a word, a phrase or a pattern of words and possibly XML-markup already in the text. The user can then study

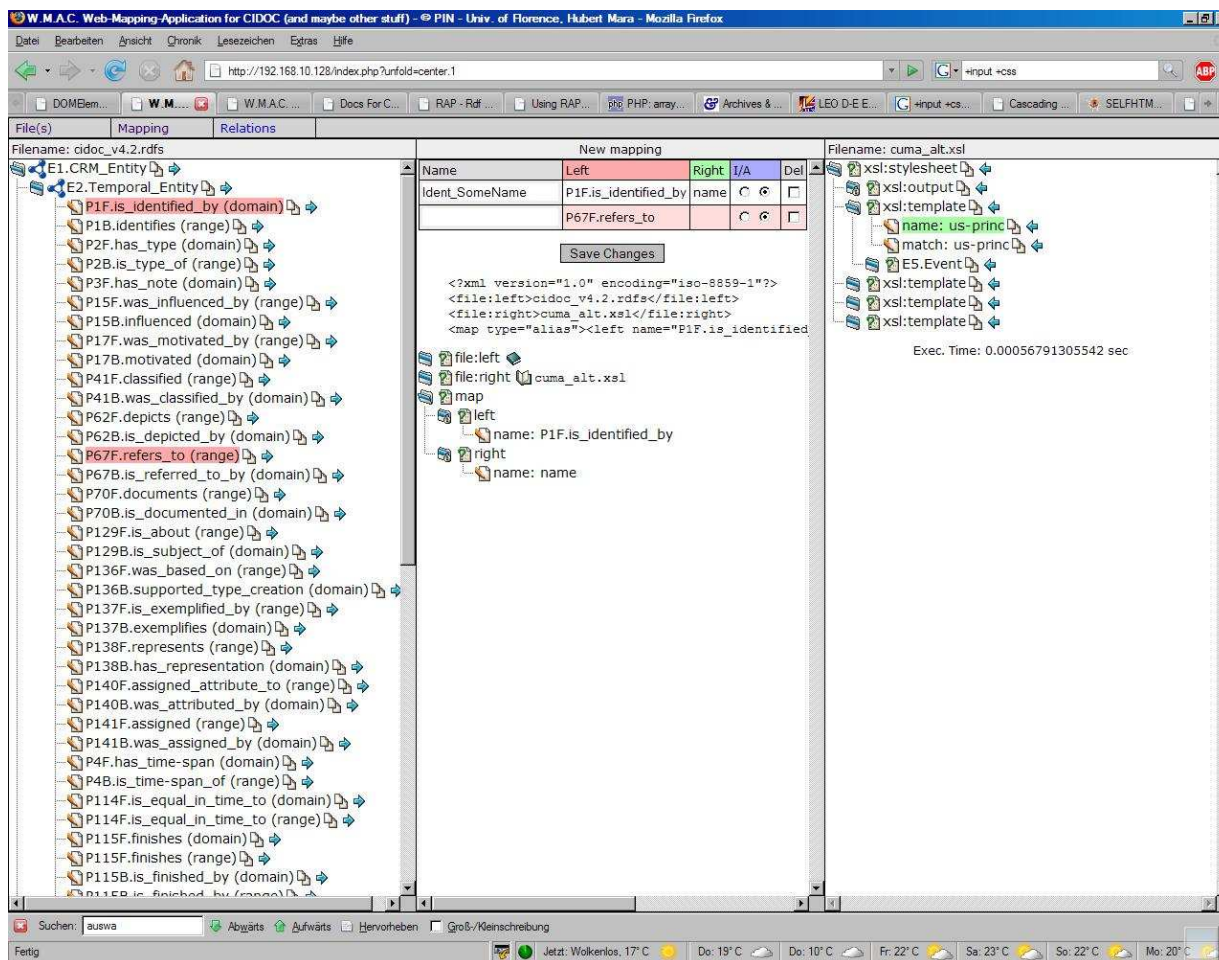


Figure 2: A snapshot of the new AMA online mapping tool.

the text and mark or tick which occurrences in the KWIC concordance he/she want to give a certain mark up (tag). The system then inserts the mark up in the text file(s).

This semi-automatic “search and replace” feature is implemented in a way that makes it possible for the user to create and include new algorithms both for text retrieval and for text mark up. In this way the tool features can be widened by the users when needed, more complex language specific algorithms can be added and new features of electronic text processing can be integrated.

To make the system more user friendly, the ability to parse and use an XML grammar will be included in a later version. Then it can suggest for the user which elements are allowed to be inserted on a certain point in the text. An XML grammar checker (a validator) will also be connected for use from within the program in a future version.

Another development will be to include the possibility of displaying two or several KWIC concordances. This is very useful when one need to interconnect elements by the use of ‘id’ and ‘idref’ attributes.

The program is written in Java and is free software, available for anyone from the EPOCH web site.

7. Storage and retrieval of mapped data

7.1. MAD

MAD (Managing Archaeological Data) is a framework originally designed to manage structured and unstructured archaeological excavation datasets encoded using XML syntax, including free text documents marked up in XML, as described above. During the development process we added more features transforming the first release into a multipurpose engine able to store and manage ontology encoded information, i.e. data structured in CIDOC-CRM compliant form. The framework can be used to browse and query such data in many powerful ways and to transform and supply semantic data on demand. The whole framework is developed using Open Source software and standard XML and W3C technology.

The first release of the tool was able to manage XML documents in a native way. It was built around eXist XML database [eXi], a Java tool easy to integrate into XML processing applications. It is able to store and index XML documents in a file-system-like structure of folders and sub-folders called collections. This allows a very simple mechanism to organise and browse XML content. The system is able to index structured documents, i.e. XML documents with a database-like structure, and also unstructured ones in the form of

tagged free texts such as excavation diaries. This includes XML documents created using the tools described above.

This gives interesting options for the future. It is important to keep the references between the CIDOC-CRM compliant models and the XML documents they are based on. As eXist is commonly used for storage of XML documents used in cultural heritage through its popularity in the TEI community, see e.g. [PGWR*05], integration between TEI and CIDOC-CRM, which has been researched for some years [EO07] will be easier to implement.

eXist comes with a set of XPath/XQuery interfaces to perform queries over the XML collections and retrieve relevant information in a fast and simple way [XQu05]. A complete transformation framework based on XSL Transformation and a pipeline mechanism allows on-the-fly mapping, transformation, serialisation and presentation of XML content in many other formats, including HTML and RDF [eXi].

7.2. MAD: The SAD extension

The second release of the tool was enriched with a set of features devoted to the management of RDF documents encoded using CIDOC-CRM, the ontology we have chosen for modelling our data. RDF-encoded ontological information required a new query framework implementing SPARQL and RQL, query languages specifically designed for retrieving information from semantic data. The SAD (Semantic MAD) extension was created mainly for this purpose [RDF04a].

Since RDF can also be represented in XML, we tried first of all to use XQuery for the location and retrieval of information from semantic documents. However we noticed that XQuery was insufficient for querying RDF graphs, even when encoded in XML [RDF04b]. An RDF graph is actually a set of triples, each consisting of a subject, an object, and a property relationship between them. These triples can come from a variety of sources. For instance, they may come directly from an RDF document or an ontology instance, they may be inferred from other RDF triples or be the RDF expression of data stored in other formats, such as XML or relational databases.

W3C has defined a new query language called SPARQL specifically designed for getting information from RDF graphs. The language consists of three specifications:

1. the *query language* itself, that is the core of the whole language
2. the *query results XML format*, a simple format easy to process with common XML tools such as XSLT
3. the *data access protocol* used to define simple HTTP and SOAP requests for remotely querying RDF databases and any data repository that can be mapped to the RDF model

The XML results format is used to generate responses from services that implement this protocol [SPA06]. SPARQL features include facilities to:

- extract information in the form of URIs, nodes and literals
- extract RDF subgraphs
- construct new RDF graphs based on information in the queried graphs

As a data access language, it can easily be used for both local and remote access combining the ability of database queries to pull data

from huge databases with the possibility to write queries in an application that can extract relevant data stored across the entire World Wide Web. The SAD extension implemented most of these features using APIs provided by the SESAME framework to build semantic query capabilities on top of the native XML database used for storing data.

Another important feature included in the SAD extension is the semantic browser that allows users to navigate through the complexity of semantic relationships.

The richness of the RDF graph model in which semantic data are distributed will often make it difficult for users to get an effective and meaningful data retrieval, if only a simple or complex query interface is used, in particular when the graph grows in dimensions and complexity. Sometimes it would be simpler and faster to browse the multidimensional structure of the graph, allowing users to choose a starting point and to move along different paths through the graph to reach the desired data.

To allow this kind of data navigation we implemented a semantic browser based on the *faceted browsing* user interface paradigm which, unlike a simple hierarchical scheme, gives users the ability to find items based on more than one dimension. A facet is a particular metadata field that is considered important for the dataset that users are browsing. The tool can be configured to prioritise which facets are “open” when the page loads and in what order. Once the facets are selected for a specific dataset, the browser starts processing the dataset and extracts a list of facets, their values, and the number of times each facet value occurs in the dataset. Then it is possible to add or remove restrictions in order to focus on more specific or more general parts of the model.

A “free text” restriction that reduces the browsed dataset to all items that contain the searched string in their properties’ values is also possible by entering a search string in an input box.

7.3. Geographic information in MAD

During the last year of EPOCH’s activity, we created some experimental geographic functions to integrate spatial archaeological information for the management of unstructured documents, such as excavation diaries and reports, in a spatial context. The system, once fully implemented, will allow the creation and distribution of rich geospatial relationships across the Web and the use of geographic data in a Semantic Web scenario. Tests have been carried out using the Geographic Markup Language (GML) to encode geographic data related to archaeological records and to store them in our container. Data serialised by the MAD system can be directly transformed in SVG or visualised using map server web applications. The flexibility of GML features will also allow the implementation of complex query-on-map functions to visually query and generate dynamic maps. MAD can also host and serialise KML archaeological files to be used in Google Earth and Google Maps applications.

The data integration of both geographic and non geographic data stored in a single container and managed using the same interfaces and the native web compliance provided by MAD will make archaeological information ready to be queried, updated and exchanged over the web, promoting the semantic evolution of geospatial web services.

The development of MAD for the management of both spatial and non spatial archaeological data is indeed the first step towards

The screenshot shows the MAD Semantic Web Browser interface. At the top left, it displays 'MAD Semantic Web Browser' and '1 filter criterion' with a list containing 'type: E28.Conceptual_Object'. Below this, there are tabs for 'Order' and 'Commands', and a link for 'Advanced Query'. The main content area shows '259 items sorted by URI [A to Z]' with a pagination bar. A table displays the results for 'US 19001' with columns for property names and values. The right sidebar contains a search bar and several expandable filter criteria, including 'label', 'P1F.is_identified_by', 'P67B.is_referred_to_by', 'P70B.is_documented_in' (which is expanded to show a list of 'Scheda US' items), 'P92B.was_brought_into_existence_by', and 'P138B.has_representation'.

Figure 3: A snapshot of the semantic web browser of MAD.

the full implementation of the geospatial semantic web of culture heritage information, a future where the World Wide Web will be machine-readable and fully integrated, allowing the return of both spatial and non-spatial resources to semantic queries.

Integration of spatial and non-spatial information can be also created in MAD during the query process. The power of the MAD query system can return complex sets of XML fragments by recursively executing chains of queries for the generation of aggregated data as shown in the following fragment:

```
<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>
```

This shows the possibility to create RDFised GML code embedding fragments of GML inside RDF. This capability allows RDF

documents to use GML statements for the description of spatial features to be visualised by semantic-enabled browsers with geographic extensions, like the W3C Tabulator [Doe01]. This is in line with the way the TEI guidelines prescribe the inclusion of GML information [TEI08, section 13.3.4.1].

7.4. Mapping support in MAD

MAD was conceived from its very beginning as a repository with a query and a transformation engine on top to provide users with a generic semantic server for their applications. Users simply store their documents, ontologies and models into the MAD database and all this information will immediately become available to be browsed, queried, transformed and exchanged. But the encoding process for data to be stored in MAD is often a complex task for users to accomplish, particularly for those who have to deal with legacy data stored in diverse archives. Even the mapping files created by AMA need further processing operations to be used for data conversion. For this reasons we are going to provide MAD with a mapping framework able to connect the MAD engine with legacy databases and to extract information stored in such databases in an automatic way according to the high level mapping files created using the AMA Mapping Tool. For this purpose we are going to write a new set of Java features based on D2R Server, a tool for publishing relational databases on the Semantic Web working with, and D2R MAP, a simple rather powerful database-to-RDF mapping language [D2R]. D2R Server works with Oracle, MySQL, PostgreSQL and any SQL-92 compatible database. Microsoft Access is partially supported. D2R MAP files (containing specific mapping information) can be easily generated by MAD from the high level mapping files created with AMA.

8. Results and further development

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 and for the creation of an on-line application for the complete management of coins collections for the COINS Project [COI].

The XML native support provided by MAD can be used in the field of digital preservation for setting up annotation repositories and creating co-reference resolution services. It can also assist in the process of linking database type information together with text type information.

In the future, it will be necessary for memory institutions, such as museums, to integrate their data with data from other institutions, crossing thematic, administrative, historical as well as language and cultural borders. To be able to do so, mapping is vital. Such mapping will need clear, educated human minds. But by continuing the development of tools to assist them, we will be able to make the most out of their effort.

References

- [AMA] Ama, archive mapper for archaeology. URL: <http://www.epoch.eu/AMA/>.
- [CDG*05] CROFTS N., DOERR M., GILL T., STEAD S., M. S. (Eds.): *Definition of the CIDOC Conceptual Reference Model*. 2005. URL: http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.doc.
- [COI] The coins project. URL: <http://www.coins-project.eu>.
- [D2R] D2r server. publishing relational databases on the semantic web. URL: <http://www4.wiwi.fu-berlin.de/bizer/d2r-server/>.
- [D'A06a] D'ANDREA A.: Documentazione archeologica, standard e trattamento informatico. metodi informatici per la ricerca archeologica. In *Archaeolingua* (Budapest, 2006).
- [D'A06b] D'ANDREA A.: Ipotesi di ontologia applicata ai modelli descrittivi/interpretativi dello scavo archeologico. In *Le Ontologie in Campo Umanistico. Archeologia, Architettura e Beni Culturali* (Firenze, 2006).
- [D'A06c] D'ANDREA A.: A preliminary ontology-based model applied to the description/interpretation of the archaeological excavation. In *Proceedings of the First International Workshop on "Ontology Based Modelling in The Humanities"* (2006), von Hahn W., Vertan C., (Eds.), University of Hamburg, pp. 38–46.
- [DMZ06] D'ANDREA A., MARCHESE G., ZOPPI T.: Ontological modelling for archaeological data. In *The 7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage, VAST2006* (2006), Ioannides M., Arnold D., Niccolucci F., F. M., (Eds.), pp. 211–218.
- [Doe01] DOERR M.: *A comparison of the OpenGIS Abstract Specification with the CIDOC CRM 3.2*. Tech. rep., FORTH, 2001.
- [EH06] EIDE O., HOLMEN J.: Reading gray literature as texts. semantic mark-up of museum acquisition catalogues, 2006. Paper presented at CIDOC 2006, Gothenburg. URL: http://www.edd.uio.no/artiklar/teknikk_informatikk/CIDOC2006/EIDE_HOLMEN_Reading_Gray_Literature.pdf.
- [EO07] EIDE O., ORE C.-E. S.: From tei to a cidoc-crm conforming model. towards a better integration between text collections and other sources of cultural historical documentation, 2007. Poster presented at the Digital Humanities 2007 conference. URL: <http://www.edd.uio.no/artiklar/teknikkoding.html>.
- [eXi] exist open source native xml database. URL: <http://exist-db.org/>.
- [FRB98] *Functional Requirement for Bibliographic Records. Final Report*. Tech. rep., IFLA, 1998. URL: <http://www.ifla.org/VII/s13/frbr/frbr.pdf>.
- [HOE04] HOLMEN J., ORE C.-E., EIDE O.: Documenting two histories at once: Digging into archaeology. In *Enter the Past. The E-way into the Four Dimensions of Cultural Heritage* (2004), BAR International Series 1227, BAR, pp. 221–224.
- [JHO0ng] JORDAL E., HOLMEN J., OLSEN S. A., ORE C.-E.: From xml-tagged acquisition catalogues to an event-based relational database. In *Proceedings from CAA 2004* (forthcoming), BAR.
- [PGWR*05] PETTER C., GROVE-WHITE E., ROBERTS L., ROSE S., POSGATE J., SHOICHET J.: The robert graves diary (1935-39): a tei application using an xml database (exist). In *ACH / ALLC 2005. Conference Abstracts (2nd Edition)* (2005), University of Victoria. Humanities Computing and Media Centre, pp. 164–166.
- [RDF04a] Resource description framework (rdf): Concepts and abstract syntax, recommendation, 2004. URL: <http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/>.
- [RDF04b] Rdf/xml syntax specification (revised), recommendation, 2004. URL: <http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/>.
- [SPA06] Sparql query language for rdf, candidate recommendation, 2006. URL: <http://www.w3.org/TR/2006/CR-rdf-sparql-query-20060406/>.
- [TD01] THEODORIDOU M., DOERR M.: *Mapping of the Encoded Archival Description DTD Element Set to the CIDOC CRM*. Tech. rep., FORTH, 2001. URL: <http://cidoc.ics.forth.gr/docs/ead.pdf>.
- [TEI08] *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 1.0.1. Last updated on 3rd. February 2008*. TEI Consortium, 2008. <http://www.tei-c.org/Guidelines/P5/>.
- [XQu05] Xquery 1.0: An xml query language, candidate recommendation, 2005. URL: <http://www.w3.org/TR/2005/CR-xquery-20051103/>.