Encoding Cultural Heritage
Information for the Semantic Web

Procedures for Data Integration through CIDOC-CRM Mapping

Ø. Eide*, A. Felicetti**, C. E. Ore*, A. D’Andrea*** and J. Holmen*

* Unit for Digital Documentation, University of Oslo - Norway
** PIN Scarl VASTLAB, University of Florence - Italy
*** CISA, University fo Naples “L’Orientale” - Italy
The BABEL tower....

In CH circles the BABEL tower is a *Metaphor* about lack of communication or incommutability for too many languages and different cultures.
As regard archaeological documentation (forms, reports, etc..) we have different streams:

- National Rules
- Archaeological Backgrounds
- Best Practices
- Local Experiences
- Standard
- Legal Obligations

EPOCH Conference on OPEN DIGITAL CULTURAL HERITAGE SYSTEMS - Rome 25th feb 2008
A paradoxical situation could be:

Io sono un preistorico e scavo seguendo le superfici...

I’m a classical archaeologist. I dig following inscriptions and texts....

je ne fait pas des fouilles archéologiques... mais je rassemble la documentation suivant le Standard ICOMOS....

Here is what can happen in the same archaeological area.....
The BABEL tower....

Could the technology help us in overcoming these issues....SURE!!!!
The BABEL tower....

<table>
<thead>
<tr>
<th>Human approach</th>
<th>Languages</th>
<th>Terminologies</th>
<th>Thesaurii</th>
<th>Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Legal approach</td>
<td>National Rules</td>
<td>Legal obligations</td>
<td>Standard</td>
<td>Guidelines</td>
</tr>
<tr>
<td>ITC</td>
<td>Platforms</td>
<td>Program Languages</td>
<td>Formats</td>
<td>Softwares</td>
</tr>
</tbody>
</table>

Resources on-line with scarce access, not always useful and usable for the lack of interoperability
A new challenge for the reconciliation of these disparate sources stored separately each other (we wanted also try to reconcile different disciplinary groups separated by cultures a much as content (methods, theories...))

A SEMANTIC APPROACH......
.....The integration of knowledge.... through ontologies

The use of CIDOC-CRM as a sort of inter-lingua enabling to guarantee data and schemas integration among cultural heritage information/archives/repositories
What is the CIDOC CRM?

• is not a metadata standard
• it is a Conceptual Reference Model for the analysis and design of cultural information systems
• does not define the terminology used to document these data structures

• It is addressed to explain the logic of what they “do” documentation
As part of the EPOCH a complete framework for the mapping and management of cultural heritage information in a semantic web context has been developed by

- PIN (University of Florence, Italy)
- CISA (University of Naples “L’Orientale”, Italy)
- EDD (University of Oslo, Norway)

**AMA** (Archive Mapper for Archaeology) is one of the NEWTON projects (NEW TOols Needed)
BUT!!!

mapping requires skills and knowledge which are uncommon among the cultural heritage Professionals
The aim of the **AMA project** was to develop a tool for **semi-automated mapping** of cultural heritage data to CIDOC-CRM (ISO 2006: 21127).

The reason for this investment was that such a tool can enhance **interoperability** among the different archives and datasets produced in the field of Cultural Heritage.

We created a tool 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 relied upon two concepts:

**Mapping** = the rules allowing to write and conciliate different schemas

**Template** = the instantiation of the abstract mapping between the source data structure and the mapping target (CIDOC-CRM compliant). Templates capture the semantic structure of the sources and their transformations ensuring the modularity and future maintenance of the system.
A schema for a mapping-tool

Template

Classes and Properties

CIDOC-CRM

Schema to be mapped

E65_Event

is a

CompilationEvent

P14_carried out by

E39_actor

is a

E21_person

E31_InformationCarrier

P94_was created by

P70_is documented in

E50_date

US

Site

Compiler

YearForm

EPOCH Conference on OPEN DIGITAL CULTURAL HERITAGE SYSTEMS - Rome 25th Feb 2008
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
AMATool: partners

CISA (Italy)
PIN (Italy)
UNIREL (Italy)
CIMEC (Romania)
IAA (Israel)
Oxford ArchDigital (UK)
University of Kent (UK)
Paveprime LtD (UK)
University of Oslo (Norway)
ROB (Netherlands)
VARTEC (Belgium)
Tools developed

- **AMA**
  - Current release
  - New online application
  - Future development
  - Software release after EPOCH

- **MAD**
  - Review release
  - Development after review
  - Future development
  - Software release after EPOCH

- Complementary tools
Creation of an **Open Source tool** for mapping existing archaeological datasets to CIDOC-CRM compliant structures. Enhance data **interchange** and **interoperability** between existing and future repositories [http://www.epoch-net.org/AMA/](http://www.epoch-net.org/AMA/)
A new and powerful online version of the application is under development (PHP)

Beta version available at http://ama.ilbello.com
AMA: new version features

- Possibility to **upload** both **starting schema** and **target ontology** for any kind of mapping (not only towards CIDOC-CRM)

- **Ontology to ontology** mapping capabilities

- Advanced mapping features for complex mapping definitions (i.e. creation of **new entities** and **properties** to represent **implicit elements** and **relations**) Creation of **shortcuts** to simplify the mapping process

- Generation of **mapping templates** to be applied to information stored in databases in order to get perfectly converted semantic archives

- Graphic visualization of simple and complex relations obtained during the mapping process.

EPOCH Conference on *OPEN DIGITAL CULTURAL HERITAGE SYSTEMS* - Rome 25th Feb 2008
Inplementation of a **high level mapping language** to describe the mappings (under development by Martin Doerr's team in Crete)

Implementation and archiving of **mapping templates** to be used as mapping starting points

Integration with MAD and other tools for automatic data extraction and storing in a semantic container
AMA releases after EPOCH

- 2 versions: Online Application and a stand-alone application
- GPL licenses
- Supported formats: all XML compliant documents, including semantic grammars (RDF and OWL), and schema formats (XSD/RDFS)
The most central classes and properties:
TEI guidelines

A set of guidelines for XML encoding of culture heritage texts. There are three primary functions of the TEI guidelines:

• guidance for individual or local practice in text creation and data capture;
• support of data interchange;
• support of application-independent local processing.
• Free text documents contain useful information
• This information is more useful in a computerised environment if it can be extracted into a well defined format
• This is impossible to do correctly by automatic tools
• This is very time consuming to do manually
• Therefore: A semi-automatic approach
• Developed by EDD, University of Oslo
• Based on 14 years experience on XML content markup of texts
• Tool for semi-automatic semantic enrichment of XML markup
• Creates CIDOC-CRM compliant markup
• Information stored in XML
What to encode in a text

- Reference to entities (e.g. person names)
- Co-reference in the text (e.g. "he" refers to the same person as the name on the previous line)
- Relations between entities (e.g. family relations between persons)
• In the tool, a document can be searched for patterns
• The result list is a KWIC concordance
• All or some of the results can be chosen for tagging
• Iterative process — semantic bootstrapping
Application designed to manage **structured** and **unstructured** archaeological excavation datasets encoded in **XML**

Developed using **Open Source technologies** and entirely based on **XML** and **W3C standards**

The core: **eXist Native XML Database** ([http://exist-db.org](http://exist-db.org))

- Fully **XPath/XQuery** and **SPARQL** aware
- Featured by dynamic **XSLT** transformation and presentation of documents and query results
- **CIDOC-CRM ontology** for semantic data encoding
Data structure of MAD

XML documents indexed and stored in a file-system-like structure of folders and subfolders

- XPath and XQuery to query XML documents
MAD: Testing the application

- First release of MAD tested for the EPOCH Partners’ XML database (http://partners.epoch.net.org:8080/exist/partners/index.xml)

- ... and for an XML database of relevant european and italian projects (http://www.epoch.net.org:8080/exist/projects/index.xml)
MAD: The SAD extension

- Implementation of the RDF language
- Full SPARQL and RQL query languages support
- Semantic Browser: intuitive set of interfaces to navigate semantic models Full CIDOC-CRM compliancy

Presented at the last EPOCH review in July 2007
ARCHAEOLOGICAL EXCAVATION DATASET OF CUMA CONTAINING INFORMATION ON STRATIGRAPHICAL UNITS AND OTHER RELATED RESOURCES CONVERTED FROM SYSLAT (HYPERCARD) AND MANAGED USING MAD

The MAD framework used for the creation of an online application for the complete management of coins collections for the COINS Project - http://www.coins-project.eu

Some features of MAD will be used for the creation of a wide XML repository of ancient iberian pottery (University of Jaen - Spain)
MAD: The future

- Full integration of **GML geospatial features** within the CIDOC-CRM framework
- XML nativ support of MAD can be used in the field of digital preservation for setting up annotations repositories and creating co-reference resolution services
- Possible COLLADA/X3D support for 3D objects represented in XML-based formats
- AJAX integration and XForms support for easy web applications development and complex XML information deployment and integration
- **Interfaces development** to query complex semantic data in a visual and user-friendly way
MAD releases

- 2 versions: **Online Application** and **Downloadable Version** to be used as a **stand-alone** application or integrated in a **client-server** context (i.e. using distributed archives)

- **GPL License**

- Supported formats: all XML compliant documents, including semantic grammars (RDF and OWL), graphic formats (SVG), geographic formats (GML) and 3D formats (COLLADA/X3D)