Integrating data from The Perseus Project and Arachne using the CIDOC CRM
An Examination from a Software Developer's Perspective

Robert Kummer, Perseus Project at Tufts University and Research Archive for Ancient Sculpture at the University of Cologne (FA), rokummer@gmail.com

Abstract: In a joint effort, The Perseus Project, a digital library hosted at Tufts University, and Arachne, the central database for archaeological objects of the German Archaeological Institute (DAI) and the Research Archive for Ancient Sculpture at the University of Cologne (FA), want to make their data accessible to a greater audience using the CIDOC CRM data model. Given the fact that the information in each of their databases is of interest to a large community of people, efforts to overcome the current lack of data integration have to be made. Aside from the philosophical implications and the mathematical background involved, the main concern of this project will be the practicability of a software implementation of all relevant concepts using basic Semantic Web technologies as described by the W3C, along with an investigation of the usability of the CIDOC CRM for a multilingual interface. The main purpose of the implementation process is to get a deeper understanding of the concepts and technologies involved when dealing with the Semantic Web, ontologies in general and the CIDOC CRM in particular. For the process of implementation it is essential that software tools team up with a methodical process, and that appropriate tools be discovered or developed and documented. Functional requirements have a tendency to evolve relatively rapidly when information systems are used by historians. Information systems in the humanities are especially confronted with the problem of constant change through the acquisition of new project partners carrying new and varied source material. As a consequence, potential integration efforts have to cope with changes of the database schemas and therefore should be flexible.

In a joint effort The Perseus Project, a digital library project hosted at Tufts University and Arachne, the central database for archaeological objects of the German Archaeological Institute (DAI) and the Research Archive for Ancient Sculpture at the University of Cologne (FA), want to make their data accessible to a greater audience using the CIDOC CRM data model. Given the fact that the information on each of the databases is of interest for a large community of people, efforts to overcome the current lack of data integration have to be made. Another area of interest is the usability of the CIDOC CRM for a multilingual interface. Besides the philosophical implications and the mathematical background, the practicability of a software implementation of all relevant concepts using basic Semantic Web technologies as described by the W3C will be the main concern. The main purpose of the implementation process is to get a deeper understanding of the concepts and technologies involved when dealing with the Semantic Web, Ontologies in general and the CIDOC CRM in particular. For the process of implementation it is essential that methodical proceedings and software tools have to team up, appropriate tools have to be discovered or developed and documented. Functional requirements have a tendency to evolve relatively rapid while information systems are used by historians. Especially information systems in the humanities are confronted with the problem of constant change through acquisition of new project partners carrying new source material with different properties. As a consequence

potential integration efforts have to cope with changes of the database schemas and therefore should be flexible.

Many different information systems with different methodical approaches can be found in the field of historical cultural research; each one is designed according to a specific scientific question and perspective. This is a productive situation and therefore should be welcomed, but the experience of using information systems for historical research could be greatly enhanced by creating a common platform for information retrieval. Scientific databases holding historical cultural material use the specialized terminology of their respective areas of research and a certain national language. The general problems involved with machine translation are well known. Moreover, terminology and conventions used can vary within one sub-domain and therefore should be restricted. Given the fact that the information in each of the databases is of interest to a large community of people, efforts to overcome the current problems with data integration have to be made. To build a software system able to do that, each database and its interface have to comply with several requirements and have to supply functionality for importing and exporting data. Against this background it seems reasonable that the CIDOC CRM, [1] delivering a set of standardized terms and properties, could serve as a foundation for heterogeneity.

In a joint effort, two parties from classics and archaeology intend to formulate a research program for achieving the goals mentioned above. These parties will be The Perseus Project [8] and Arachne [4]. The Perseus Project is a digital library hosted at Tufts University. It provides humanities resources in digital form with a focus on Classics but also provides early modern and even more recent material. Arachne is the central database for archaeological objects of the German Archaeological Institute (DAI) and the Research Archive for Ancient Sculpture at the University of Cologne (FA). DAI and FA joined their efforts in developing Arachne as a free tool for archaeological internet research.

Currently only a few implementations exist that try to bridge the gap between more than one language and several data models at the same time. To overcome the lack of research with implementing the CIDOC CRM as an intellectual concept and as a software system, we will undertake a robust implementation. Although it is often emphasized that the CIDOC CRM is an intellectual artefact which does not directly deal with implementation, exactly that is intended: To apply the CIDOC CRM to two information systems holding historical cultural material (in this case Perseus and Arachne) and to elaborate a robust implementation of a mapping agent. [2] But what exactly are we doing while implementing intellectual concepts as a software system? To learn how to implement an intellectual concept like the CIDOC CRM in the field of archaeology, we need to understand what archaeologists are doing when conducting research.

In contrast to enabling both databases to deal with the CIDOC CRM data, it seems to be easier to build a mapping agent for each of them than to change the structure of all participating information systems. This mapping agent has to be aware of both database schemas to be able to translate data to a shared common format. Because it has been argued that the belief in easily building a mapping agent is naïve [9] we will focus on that point in order to find ways to overcome the current problems. After exporting the data and mapping it to the CIDOC CRM data model the exported and integrated data will be stored in a central repository providing basic query functionality for the time being. At a later stage this repository should offer facilities for complex

database queries in more than one language with acceptable performance. Since a large corpus of possible queries will be supported, massive problems with complexity and performance are expected. More knowledge about how historians formulate their queries could help to reduce this complexity.

In order to put these ideas into action, standards for representing and querying archaeological domain knowledge have to be found. For experimentation both project partners will provide an XML dump as raw material of their data model and content which at a later stage can be done by periodic data harvesting. [3] It is the first goal to map the structure and the data to a markup format that both parties will agree to and which will be stored in a central ontology driven repository server for further development (e.g. RDF) [5]. The data storage within the central repository has the advantage that the databases do not have to change their implementation. The repository will provide basic query functionality through a semantic layer using an abstract language independent data model like the CIDOC CRM (e.g. SPARQL). Moreover, both parties have to agree on the level of semantic and structural granularity of the provided data.

To answer the questions essential for the process of implementation, methodical proceedings and software tools have to team up. A survey of existing software that is able to deal with the standards mentioned above has already been carried out, with the following results. Protégé, a free ontology editor and knowledge base framework, helped with exploring and analysing the data models against the CIDOC CRM ontology and experimenting with retrieval techniques. [PROTÉGÉ] Jena is a framework for building Semantic Web applications. The Jena API developed by HP Labs in connection with Schemagen is one component for a very simple mapping that dumps the exported data to RDF/XML using the RDF API. The Jena framework furthermore supports reading and writing RDF in several formats: an OWL API, in-memory and persistent storage, and a SPARQL query engine. [JENA] Both projects are free of charge and seem to have an active development team and a strong user community, promising further development and enhancements.

After having developed a prototype of a mapping component using the tools and standards mentioned above with an extremely simple data model there is much to be said for a CRMish implementation relying on standard technologies as proposed by the W3C rather than for a full CRMized implementation. Although the high complexity offered by new data models enables us to realize new ideas, neither the intellectual considerations nor the available software tools have long been discussed in the humanities, which leads to a lack of experience in applying the relevant methods and discovering adequate standards. Due to this it is difficult to estimate efforts and costs to formulate realistic project goals for organizations wanting to implement the CIDOC CRM. Ways to reduce, hide or manage complexity have to be found.

The internal data model is presented to the user by a graphical user interface. Here new GUI concepts have to be developed to present internal complexity to users in a manner that they can handle. In order to archive this, visualization is indispensable. Using a data model which offers a potentially huge complexity therefore poses a great demand on the usability of an information system to be put across to the user in a way he can understand and handle. As an example the Protégé Project made some efforts to automatically build user interfaces from the underlying data model. Ontobroker, a deductive, object oriented database system, combines classic search interfaces with visualization techniques. [ONTOBROKER]

What is meant by saying that the mapping agent has to be flexible? In software technology one often means that a piece of software has to be modular, adaptable and maintainable. It is interesting that all three points are dealing with complexity. These points become problematic when dealing with information systems working with historical cultural data. In this context an information system lays the foundation for being able to handle complex and often semi-structured data from the field of history. In addition, functional requirements have a tendency to evolve relatively rapid while information systems are used by historians. As the understanding of the subject increases, new questions and requirements come up. A flexible information system therefore must be able to advance at the same pace as an information system evolves an aspect to be considered in the design phase already. [0]

Due to a strong commercial influence, relational databases are widely used for historical cultural knowledge. Relational databases do not support rich semantic modelling of data which urges the software developer to model semantics on higher levels of the information system. Therefore it has to be taken into consideration that each information system changes on several layers. A change in one of these layers potentially causes the need to adapt the mapping mechanism and regarding only the data model is not enough. Beneath the internal data model (storage and internal representation of factual knowledge) the application logic (first layer of interpretation) retrieves and recombines the data for the graphical user interface (second layer of interpretation, interpretation of data model) including layouts that display the information to the user in multiple views and the user's implicit knowledge (third layer of interpretation) about the information system, including implicit conventions, have to be taken into account. The implementation of the CIDOC CRM impacts all the mentioned layers and a mapping agent has to be aware of all these layers because it needs to preserve all implicit layers of meaning. This is reminiscent of the principles of composition that have been formulated by Frege. That principle has to be formalized to make a mapping agent reusable: "The meaning of a complex expression is determined by its *structure* and the meanings of its constituents." [10] The resulting question is: How can a mapping component be built in such a way that it can adapt to changes easily?

As mentioned above, the semantics of a database application is spread over several layers. It is at least questionable that complex and highly structured data contained in one information system (context I) can be transferred to another information system (context II) without loss of (structure!) information without preserving the process of composition. Most current databases in the humanities don't meet halfway because of their proprietary semantic modelling without relying on standards. In the future it could be reasonable to build awareness of a shared data model like the CIDOC CRM into the original database application. Since it appears too complex to map the whole data structure to a shared data model it is important to determine those parts of the data model that are most important and valuable for integration. Furthermore it is not practicable to map each detail and therefore a reasonable level of detail has to be determined.

The resulting data will be held in a central repository committed to the CIDOC CRM data model. This repository will contain many lean but highly structured records. Each record links to the original data source for further information. The analytical strength of the CIDOC CRM data model is most effective when the form of query operations is not limited. In general a query is fast if data is requested in a way that is supported by the data structure. Therefore questions should be restricted or precompiled for those which are used often by a crawler application. How can the conflict between open scientific questions resulting in complex and non predictable query operations and performance be solved?

References

[0] Onno Boonstra, Leen Breure and Peter Doorn (2004): Past, present and future of historical information science, in: Historical Social Research / Historische Sozialforschung, Vol. 29 (2004), No. 2.

[1] Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, Matthew Stiff (2005): Definition of the CIDOC Conceptual Reference Model.
URL: http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.pdf

[2] Martin Doerr (2001): Mapping format for data structures of the CIDOC CRM, July 2001.
URL: http://cidoc.ics.forth.gr/docs/mappingdtd.pdf

[3] Reinhard Foertsch (2003): Internet-Datenbanken. Jahrestagung des DArV am 20. - 22. Juni 2003 in Köln zu "Archäologie und Medien"
URL: http://mar-fileserver.klassarchaeologie.uni-koeln.de/lehre/dozenten/praesentationen/Internet_Datenbanken.htm

[4] Reinhard Foertsch (2006): ARACHNE - Datenbank und kulturelle Archive des Forschungsarchivs fuer Antike Plastik Koeln und des Deutschen Archaeologischen Instituts.
URL: http://arachne.uni-koeln.de/

[5] T. R. Gruber (1993): Toward principles for the design of ontologies used for knowledge sharing. Presented at the Padua workshop on Formal Ontology, March 1993, to appear in an edited collection by Nicola Guarino.
URL: ftp://ftp.ksl.stanford.edu/pub/KSL_Reports/KSL-93-04.ps.gz

[6] ICS-FORTH (2005): Partial Definition of the CIDOC Conceptual Reference Model version 4.2 in RDF, June 2005.
URL: http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs

[7] Carl Lagoze, Jane Hunter (2001): "The ABC Ontology and Model", DC-2001, International Conference on Dublin Core and Metadata, Tokyo, October 2001.
URL: http://metadata.net/harmony/lagoze_hunter_dc2001.pdf

[8] David A. Smith, Jeffrey A. Rydberg-Cox, and Gregory R. Crane. The Perseus Project: A digital library for the humanities. Literary and Linguistic Computing, 15(1):15-25, 2000.
Abstract URL: http://www.oup.co.uk/litlin/hdb/Volume_15/Issue_01/150015.sgm.abs.html

[9] Regine Stein et al. (2005): Das CIDOC Conceptual Reference Model: Eine Hilfe für den Datenaustausch? Deutscher Museumsbund Fachgruppe Dokumentation, Arbeitsgruppe Datenaustausch Berlin.
URL: http://www.museumsbund.de/cms/fileadmin/fg_doku/publikationen/CIDOC_CRM-Datenaustausch.pdf

[10] Zoltán Gendler Szabó (2005): Compositionality, in: Zalta, Edward N. (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2005 Edition).
URL: http://plato.stanford.edu/entries/compositionality/

Web Standards

T. Berners-Lee, R. Fielding, L. Masinter (2005): Uniform Resource Identifier (URI): Generic Syntax
URL: http://www.gbiv.com/protocols/uri/rfc/rfc3986.html

UNI (2005): What is Unicode?
URL: http://www.unicode.org/standard/WhatIsUnicode.html

W3C (2001): Semantic Web.
URL: http://www.w3.org/2001/sw/

W3C (2004): Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation 10 February 2004.
URL: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

W3C (2004): RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation 10 February 2004.
URL: http://www.w3.org/TR/2004/REC-rdf-schema-20040210/

W3C (2004): OWL Web Ontology Language. Use Cases and Requirements. W3C Recommendation 10 February 2004
URL: http://www.w3.org/TR/2004/REC-webont-req-20040210/

W3C (2004): Extensible Markup Language (XML) 1.0 (Third Edition). W3C Recommendation 04 February 2004.
URL: http://www.w3.org/TR/2004/REC-xml-20040204

W3C (2006): SPARQL Protocol for RDF. W3C Working Draft 25 January 2006.
URL: http://www.w3.org/TR/2006/WD-rdf-sparql-protocol-20060125/

Software

Chris Bizer, Daniel Westphal (2006): Developers Guide to Semantic Web Toolkits for different Programming Languages.
URL: http://www.wiwiss.fu-berlin.de/suhl/bizer/toolkits/

Jena Semantic Web Framework.
URL: http://jena.sourceforge.net/

Jena schemagen HOWTO

URL: http://jena.sourceforge.net/how-to/schemagen.html

Ontobroker
URL: http://ontobroker.semanticweb.org/

The Protégé Ontology Editor and Knowledge Acquisition System
URL: http://protege.stanford.edu/