A STAR is Born: Some Emerging Semantic Technologies for Archaeological Resources

Keith MAY[1]– Ceri BINDING[2] – Doug TUDHOPE[2]

[1] English Heritage, Strategy Dept, Fort Cumberland, Portsmouth, UK
[2] Glamorgan University, Faculty of Advanced Technology
[1]Keith.May@english-heritage.org.uk
[2]{cbinding; dstudhope}@glam.ac.uk

Abstract

Following work on data modelling of the varied archaeological systems at English Heritage's Centre for Archaeology (CfA), a Conceptual Reference Model (CRM) has been produced using extensions of the higher level CIDOC-CRM ontology. Some initial results of the modelling, using the CIDOC-CRM, were presented at CAA2004 in Prato (Cripps and May 2004 CAA forthcoming). Since then work has progressed on refining the modelling and the graphical diagram of the CfA's archaeological processes to create a number of further extensions of the CIDOC-CRM to explicitly reflect the greater details (granularity) of CfA's archaeological information domain. This paper presents the latest archaeological ontological modelling at English Heritage (EH), and discusses some of the methods used to implement the modelling by mapping the ontological model to specific data fields in the CfA's legacy systems, existing databases and planned future systems. This latest EH modelling and mapping work has been conducted in collaboration with Glamorgan University's Faculty of Advanced Technology, under the remit of the AHRC funded STAR project (Semantic Technologies for Archaeological Resources), and includes the development of the archaeological CIDOC-CRM extensions in RDF format, along with the investigation of the emerging SKOS W3C standard for development of online semantic terminology and knowledge organisation tools and has most recently resulted in a prototype CRM "semantic" browser.

Keywords

Ontology, CIDOC-CRM, CRM-EH, RDF, Semantic Web, STAR

1. Introduction

Much of the background to the ontological modelling of the English Heritage (EH) archaeological domain has been presented at a previous CAA conference in 2004 (Cripps and May 2004, forthcoming) and further publications and outputs are available from the CIDOC CRM website (Cripps *et al.* 2004, May, 2006). One of the principle drivers for the ontological modelling was the requirement to develop a new information system for the EH archaeological teams that better reflected the inter-relationships between their data sets that had become quite isolated as information islands in an "archaeological information archipelago". Another was to be able to re-integrate existing or older legacy data sets together with data that would be entered into the new information system in the future. It was decided to try and model not just existing data, but also try to produce a model that better reflected how those disparate databases might be integrated into a newly designed system. The English Heritage CRM will be referred to in this article and future publications as the "CRM-EH" to distinguish it from the CIDOC CRM ontology which it is based upon.

2. Why do it? Joining up projects and planning for a new Information System

Considerable efforts have been expended at English Heritage, especially in recent years as systems become older, in trying to integrate the data from various archaeological projects and their associated activities. For example an excavation may produce a project database of contexts, finds, plans, photos and text-based reports, but there may also be the original geophysical survey of the site, or part of the site or a related site, the environmental samples and soil samples derived for scientific analysis, the human and animal bone remains that form a separate area of study. Each of these activities – often along with

many others depending upon the size, period and complexity of the archaeological site – will produce databases and associated data that all relate to each other in various ways. However due to the way the data is held in relational databases it is often quite difficult to construct relational queries (using database query language such as SQL) that reflect the complexities of such relationships (e.g. "can you find all the samples with Spelt seeds from Corn Dryers that were associated with 2nd century contexts and which also contained Barley grains").

It was in order to start being able to explore this type of more complex, or semantically defined, query that the CRM modelling was undertaken. To do so required the use of some form of over-arching relational language that would be better able to express the complexities of the relationships between our data and more explicitly reflect the kinds of complex semantics that such queries required. This led to the idea of adopting an ontology, although it was not obvious whether any such ontology existed for archaeology. This led to the consideration of the CIDOC-CRM (now formally ISO 21127:2006), which was known to be a developing standard for event based modelling of Cultural Heritage information.

3. Why the CIDOC-CRM?

Rather than attempting the somewhat daunting task of trying to invent an ontology for archaeology from scratch it seemed much more viable to try to adopt, or extend, the existing CIDOC CRM ontology which has been evolved by many people's work since 1996 (Crofts *et al.* 2008), in order to model many aspects of the cultural heritage domain. On examining the CIDOC CRM, and after some quite intense introductory workshops, it became clear that many of the higher level concepts used in the CIDOC ontology were applicable and very relevant to the EH archaeological domain. Initially the modelling of the EH archaeological domain was carried out using simply the existing CIDOC CRM ontology (ref Cripps *et al.* 2004) but after consultation with CIDOC CRM-SIG it was agreed that the archaeological entities should be treated as extensions of the relevant CIDOC CRM entities.

4. Integration issues – old data and new requirements

To test the possible implementation of the modelling a number of data sets were chosen in order to map the entities in the ontological model to specific data fields in example archaeological data sets. The choice of initial data sets to map to the CRM-EH was very much based on 'test-bed' requirements to enable proto-typing of various implementation issues that those datasets represented rather than primarily for the archaeological questions that their content might answer, although an attempt was made to also choose datasets that broadly covered a range of archaeological periods. The initial datasets chosen were Raunds Roman Archaeological Database (RRAD) along with Raunds prehistoric data, Raunds environmental sampling data and the Silchester LEAP data. These datasets were selected to cover a variety of issues that the STAR project wanted to address. Each dataset was originally created in different types of database software and accordingly had different underlying - though conceptually related – data structures. RRAD was built in MS Access; Raunds environmental data is held in a DBF MS Excel format; Raunds prehistoric data was held in the English Heritage legacy system known as Delilah which outputs in a comma-delimited ASCII format; Silchester Leap data has been published online and is available from the ADS website (http://ads.ahds.ac.uk/catalogue/archive/silchester_ahrc_2007/index.cfm) but was actually supplied to STAR in it's MYSQL data format from the IADB. As well as being from quite differing database origins these data sets were also from different stages in the project management process which archaeological projects tend to follow (English Heritage 2007): Raunds prehistoric data was the excavation data as archived after work on the site was completed; Raunds environmental data derived from the specialist environmental assessment work carried out by staff of the former Ancient Monuments Lab at English Heritage (Campbell forthcoming); RRAD is at the Analysis stage following on from the recommendations in the Assessment stage work; Silchester Leap data was integrated with a 'fully' published and peer reviewed journal article in internet Archaeology (Clarke 2007) . These characteristics of the four initial datasets are summarized in *Table 1*.

For the purposes of being able to query across all of these datasets simultaneously they were mapped to the CRM-EH and their data and the relationships

	Database Type	Archaeological Period	MoRPHE Project stages
Raunds Prehistoric	EH Delilah - CSV	Neolithic & Bronze Age	Execution - Excavation
Raunds Environmental data	MS Excel - DBF	RO, IA, BA, NE, *et al*	Execution - Assessment
Raunds Roman (RRAD)	MS Access - MDB	Roman & Iron Age	Execution - Analysis
Silchester LEAP data	MySQL - MYD	Roman & Late Iron Age	Execution - Publication

Table 1. Summary of main attributes of initial test-bed data sets for STAR prototype CRM browser.

between data entities were exported to an RDF triple store (see 6.2 below)

5. Modelling versus mapping

The following three sections will briefly show the difference between the uses of the terms "modelling" and "mapping" and give some details of the methods used by EH for ontological modelling using the CIDOC CRM and in mapping to specific datasets using the CRM-EH.

5.1. CRM modelling methods

The most common approach to date for working with the CIDOC CRM is to take a well defined data model – generally extracted from existing database structures – and map specific data items directly to CRM entities. Unfortunately, not all the archaeological systems in use within English Heritage had suitable design documentation to enable this and many 'systems' were either not actually on any computer or relied heavily on manual input. As a consequence, a different approach was adopted.

The initial intention was to take the results from the Review of Existing Systems produced as part of the assessment stage of the Revelation Project (May, S 2004), supported by a first round of interviews with members of staff, in order to gather enough information to produce a series of draft models. These models could then be taken round to CfA staff in an iterative process, refining and enhancing them to capture additional detail and check for misinterpretations. It soon became apparent that for this process to work, both interviewer and interviewee needed to be familiar enough with the CRM for them to discuss their work in terms of CRM constructs. Accordingly, the initial interviews were used to collect notes and produce draft diagrams without using CRM constructs.

The next step was the compilation of an overall model built using Universal Modeling Language (UML) diagrams to present CfA concepts using CRM entities and properties in a graphical form. This allowed for ease of understanding of the model as it developed, a graphical representation being much easier to work with than a list of mapping statements. The event-driven nature of the CRM also enabled the modeling to be more explicit about gaps in our representation of the Archaeological information domain. If an object exists, be it an archaeological physical object (colloquially described by the concept "find") or its associated documentation (an Information Object in CRM terms), they must be the product of some event that brought them into existence, either a creation event in the past that produced the physical find, or a creation event in the present (or more accurately, the more recent past) where an archaeologist documented the discovery of the find during some archaeological process (e.g. excavation or 'finds-recording' or analysis). Hence it follows that if there are objects without associated events, there must be events missing from the model.

5.2. CRM-EH modelling diagram

The need to graphically model the detailed inter-relationships of the Conceptual Reference Model for Archaeology (i.e. literally to "see the bigger picture") led to the creation of a diagrammatic representation of all the (circa) 125 CRM archaeological entities and their associated properties – including about 10 CRM archaeological properties extensions - and after a considerable amount of iterative drafting and re-drafting of the working versions between the various domain experts from different archaeological teams, it was finally given a comprehensive layout overhaul by one of the Graphics Design experts in the archaeological illustrations team at Fort Cumberland to remove as many of the overlapping lines between relationships (rather like de-bugging a stratigraphic matrix) to try and make the diagram as visually comprehensible as possible in terms of it's layout. Even so the diagrammatic view of the model

is complex enough that it remains difficult to display and to publish in a non-digital format.

http://cidoc.ics.forth.gr/docs/Ontological_ Modelling_Project_Report_ Sep2004.pdf (see CD ROM)

Fig 1. Ontological Model of Centre for Archaeology Information Domain (V9)

Because the diagram is complex, and in order to capture the semantic details that each archaeological extension represents, a supporting text based Word document containing the individual scope notes, in the form of a series of tables, was produced for each entity. The scope notes describe significant points about the precise meanings of particular entities and also for the main entities the tables show the main relationships that hold between the entities, as inherited from the CIDOC CRM. A central feature of this document is a statement outlining each concept as used in the model, similar in form to CRM scope notes, and like the CIDOC CRM these scope notes were later included in the RDF descriptions field for the RDF version of the CRM-EH model available from the STAR website.

5.3. Mapping

The first stage of the modelling diagram and associated descriptions was completed in 2004 and disseminated as the results of the CfA's Ontological Modelling Project (Cripps *et al.* 2004). The next stage of work would be an attempt to implement (or indeed 'road test') the CRM-EH modelling, but to do this required a direct mapping of the entities in the CRM-EH extended model to specific data fields in a suitable selection of associated archaeological databases that the CfA model purports to represent (i.e. actual items of data in various context recording systems). This is one of the primary aims of the STAR project.

Rather like the early stages of the modelling project, a short assessment was made of available software for carrying out the process of mapping an entity (i.e. matching a field in a database table to the appropriate entity in the CRM-EH model, or vice versa).

In some extreme cases the mapping resulted in the need to create "surrogate data". This was because the model included entities where we had modelled our future information requirements, i.e. information that we decided we would want to hold in the newly designed system, but which we currently, or in former legacy systems, had not been recording in the database e.g. the event - documented as metadata - which created a site photograph whereas CfA in the past only usually kept a record of the photo itself and, more recently, a description of what it depicts).

6. Tools, technologies and into the future

The following section will briefly give an overview of some tools used to manage the development of the CRM-EH, with some of the pros and cons encountered, along with some suggestions based on 'lessons learned'.

6.1. Protégé and ontology modelling software in general

Until work began on the STAR project in January 2007, the extensions of the CIDOC CRM schema for the archaeological excavation and analysis process at EH (Cripps *et al.*) only existed "on paper" either as a PDF image of the modeling diagram (colloquially referred to as "the Flying Spaghetti Monster") or as text based descriptions of the semantic 'meanings' behind the entities and properties used in the modeling (in actuality the diagram was a PDF and the descriptions are kept as a Word.doc). Working with Tudhope, Binding and Zafiriu at Glamorgan, an initial prototype implementation of the CRM-EH environmental archaeology section in RDF was produced by Glamorgan using Protégé. A preliminary version was presented at the Cluster meeting at ECDL 2006 in Alicante and feedback from there informed further development of the CRM-EH. The implementation work was reviewed and updated by EH and the implementation was subsequently revised and extended to include all the available entities of the CRM-EH in RDF. EH were given the capability to model and design using Protégé, themselves, although in practice trying to keep the CRM-EH modeling up-to-date and available to non-IT expert archaeological domain users using ontology modeling software such as Protégé or Altova SemanticWorks has proved a relatively problematic "overhead" for this type of project.

This initial prototype implementation of the CRM-EH raised various issues concerning the practical implementation of the CRM for a working application, including literal properties, identifiers,

extensions of properties and mapping to thesauri in RDF. These issues are discussed further the ECDL paper (Binding *et al.* forthcoming).

6.2. SKOS thesauri browser

The STAR project has now developed a pilot set of semantic web services, using the SKOS Core data model for thesauri and related knowledge organization systems. SKOS is a formal RDF/XML representation standard for the large family of vocabularies and concept structures, with an informal semantics designed for information retrieval purposes. This offers a lightweight, cost effective approach for annotation, search and browsing oriented applications that do not require first order logic.

The SKOS services provide term look up in vocabularies known to the STAR system (e.g. The EH National Monuments Thesaurus and MDA Objects Thesaurus (see FISH web site), allowing browsing and semantic concept expansion. The SKOS Thesauri browser service, with extensions for concept expansion, is based on a subset of the SWAD Europe SKOS API (follow JavaDoc link to get API). The service currently consists of 7 function calls, which could be integrated into a textual or metadata based search system. For further technical details see the STAR website - http://hypermedia.research.glam.ac.uk/kos/STAR.

The services currently provide term look up across the thesauri held in the system, along with browsing and semantic concept expansion within a chosen thesaurus. This allows search to be augmented by SKOS-based vocabulary and semantic resources (assuming the services are used in conjunction with a search system). Users may browse a concept space to explore and become familiar with specialist terminology or as part of a broader application. A query is often expressed at a different level of generalisation from document content or metadata, or may employ a slightly different semantic perspective. In combination with a search system, the services allow queries to be expanded (automatically or interactively) by synonyms or by expansion over the SKOS semantic relationships. Expansion is based on a measure of 'semantic closeness'.

6.3. CRM browser

Based upon the implementation of the CRM-EH in RDF which is integrated with the existing CIDOC CRM ontology in RDF and other ontologies, the STAR project has been able to develop a CRM Browser API which enables a degree of 'semantic' searching and browsing across the four different archaeological datasets that have been mapped to the CRM-EH. The data and relationships between them, as defined by the CRM-EH model, have been exported to an RDF Triple Store running on the Glamorgan server. Work will continue on refining and enhancing the browser interfaces based upon user feedback evaluations and trials.

7. Conclusions and proposals for further work

To date only four archaeological datasets have been mapped to the CRM-EH (see section 4 above). Further work now needs to be done to test how the STAR browser and supporting web services respond to larger quantities of data, and this will require exporting more data to the RDF Triple store. An immediate advantage resulting from the careful selection of initial datasets is that there are considerable amounts of other project data sets in both the Raunds, and IADB database structures that should map relatively easily – indeed any datasets that share the same data structures (and field names) should map relatively automatically and could therefore be imported into the STAR RDF triple store (or a triple store that STAR can search across). But it may prove useful in assessing the 'cost-benefits' of the current mapping methodology to attempt further CRM-EH mappings of other database structures to enable a richer test-bed of data for running more complex semantic queries on.

The STAR project is currently only resourced to build an online demonstrator, and cannot guarantee permanent maintenance of the server demonstrator, but the project team are hoping to publish the server software as open source by the end of the STAR project, and if successful the technology should be applicable to a range of online resources planned for future implementation.

Acknowledgements

The views expressed here are principally our own, but they are based on the work of various domain experts at English Heritage. We would particularly like to thank Paul Cripps for his hard work on the initial development of the archaeological ontological model.

References

Binding, Ceri, May Keith and Tudhope Doug (forthcoming). Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM. *Proceedings of ECDL 2008*.

Campbell, G. (forthcoming). Raunds Area Project: Environmental assessment report. EH Research Reports series.

Clarke, Amanda, Mike G Fulford, Mike Rains and Klare Tootell (2007). Silchester Roman Town Insula IX: The development of a roman property c. AD 40-50 - c. AD 250. *Internet Archaeology* 21 *http://intarch.ac.uk/journal/issue21/silchester_index.html (visited 27-01-10)*.

Cripps, Paul, Anne Greenhalgh, Dave Fellows, Keith May and David-Earle Robinson (2004). Ontological Modelling of the work of the Centre for Archaeology. *CIDOC CRM Technical Paper*. pdf file (207 Kb), The CRM Diagram, pdf file (65 Kb).

Cripps, Paul and Keith May (forthcoming). To OO or not to OO? Revelations from ontological modelling of an archaeological information system. *CAA Prato 2004*.

Crofts, Nick, Martin Doerr, Tony Gill, Steve Stead and Matthew Stiff (eds) (2008). Definition of the CIDOC Conceptual Reference Model. (*visited 21-05-08*) doc file (2.5 Mb), pdf file (688 kb).

English Heritage (2007) *Management of Research Projects in the Historic Environment*. http://www.english-heritage.org.uk/MoRPHE (*visited 27-01-10)*.

May, Keith (2006). Integrating Cultural and Scientific Heritage: Archaeological Ontological Modelling for the Field and the Lab. *CIDOC CRM SIG Workshop*, Heraklion (http://cidoc.ics.forth.gr/workshops/heraklion_october_2006/may.pdf *(visited 27-01-10)*.

May, Keith (2007). Report on English Heritage Archaeological Application of CRM. *CIDOC CRM SIG Workshop*, Edinburgh.

May, Sarah *et al.* (2004). *Revelation: Phase 1 Assessment*. English Heritage Research Report 78/2004.

Miles, Alistair, Brian Matthews, and Michael Wilson (2005). SKOS Core: Simple Knowledge Organisation for the Web. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, 5–13.

Tudhope, Doug, Traugott Koch and Rachel Heery (2006). Terminology Services and Technology: JISC State of the art review http://www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_technology_review_sep_06.pdf *(visited 27-01-10)*.

Web references (*all last visited 27-01-10)*

STAR project website - http://hypermedia.research.glam.ac.uk/kos/STAR

CIDOC-CRM website - http://cidoc.ics.forth.gr/index.html (now formally ISO 21127:2006)

Silchester Internet Archaeology publication - http://intarch.ac.uk/journal/issue21/silchester_index.html

Silchester LEAP archive data - http://ads.ahds.ac.uk/catalogue/archive/silchester_ahrc_2007/index.cfm

MoRPHE web page - http://www.english-heritage.org.uk/MoRPHE

SKOS Core - http://www.w3.org/2004/02/skos/

Forum for Information Standards in Heritage (FISH) - http://www.fish-forum.info/

SKOS API - http://www.w3.org/2001/sw/Europe/reports/thes/skosapi.html


Appendix A:
Ontological Model of
Centre for Archaeology
Information Domain (V9)