Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM

Ceri Binding[1], Keith May[2], Douglas Tudhope[1]

[1] University of Glamorgan, Pontypridd, UK
{cbinding, dstudhope} @glam.ac.uk

[2] English Heritage, Portsmouth, UK
keith.may@english-heritage.org.uk

Abstract. Findings from a data mapping and extraction exercise undertaken as part of the STAR project are described and related to recent work in the area. The exercise was undertaken in conjunction with English Heritage and encompassed five differently structured relational databases containing various results of archaeological excavations. The aim of the exercise was to demonstrate the potential benefits in cross searching data expressed as RDF and conforming to a common overarching conceptual data structure schema - the English Heritage Centre for Archaeology ontological model (CRM-EH), an extension of the CIDOC Conceptual Reference Model (CRM). A semi-automatic mapping/extraction tool proved an essential component. The viability of the approach is demonstrated by web services and a client application on an integrated data and concept network.

Keywords: knowledge organization systems, mapping, CIDOC CRM, core ontology, semantic interoperability, semi-automatic mapping tool, thesaurus, terminology services

1 Introduction

Increasingly within archaeology, the Web is used for the dissemination of datasets. This contributes to the growing amount of information on the 'deep web', which a recent Bright Planet study [1] estimated to be 400-550 times larger than the commonly defined World Wide Web. However Google and other web search engines are ill equipped to retrieve information from the richly structured databases that are key resources for humanities scholars. Cultural heritage and memory institutions generally are seeking to expose databases and repositories of digitised items previously confined to specialists, to a wider academic and general audience.

The work described here draws on work carried out for DELOS WP5 activities on Semantic Interoperability [2] and the STAR (Semantic Technologies for Archaeology Resources) project [3]. The work is in collaboration with English

Heritage (EH), building on their extension of the CIDOC CRM core ontology [4] for the archaeological domain (CRM-EH). The aim of the research is to investigate the utility of mapping different datasets to a common overarching ontology, where the datasets are indexed by domain thesauri and other vocabularies. The rationale is to promote effective search across multiple different databases and their associated controlled vocabularies.

The specialisation of the CRM schema for the archaeological excavation and analysis processes undertaken by English Heritage had only existed previously on paper (**Fig. 1**). Working with May, an initial implementation of the CRM-EH environmental archaeology extension was produced by Glamorgan as a modular RDF extension referencing the published (v4.2) RDFS implementation of the CRM [5]. In addition other useful modular extensions were produced; one in particular to specify inverse relationships between existing CRM properties – information that was not explicit in the existing published RDFS implementation[2] but would be used extensively within STAR.

This exercise raised various practical issues including modelling of literal properties, specification of unique identifiers, property sub-classes and mapping to controlled vocabularies.

2 Extending the CIDOC CRM for the Archaeology Domain

Within archaeology, the CIDOC Conceptual Reference Model (CRM) is emerging as a core ontology [6]. The CRM is the result of 10 years effort by the CIDOC Documentation Standards Working Group and has become an ISO Standard (ISO 21127:2006). It encompasses cultural heritage generally and is envisaged as 'semantic glue' mediating between different sources and types of information. Thus it has particular relevance for archaeological cross domain research.

EH plays a leading role both nationally and internationally in dissemination of standards, and its staff are known for work in digital archiving [7]. The existing situation is one of fragmented datasets and applications, employing different schema and terminology systems. The initial work on the CRM-EH was prompted by a need to model the archaeological processes and concepts in use by the (EH) archaeological teams, to inform future systems design and to aid in the potential integration of archaeological information in interoperable web based research initiatives. The initial picture showed the archaeological systems as a rather disparate grouping, or 'archipelago', of diverse, specialised, but rather isolated and independent information systems and databases. In many cases, due to their age, these systems do not have very clear mechanisms to enable the sharing of data either between the different data islands within EH or with the outside world. Whereas conventional entity-relationship

[2] An OWL version of CIDOC CRM (v4.2) was published as this work was nearing completion, however being a translation of the existing RDFS implementation it did not contain the owl:inverseOf relationships required for use within STAR. A later version (v4.2.4) was subsequently made available incorporating these relationships but it references a different base namespace and uses different property naming conventions to the earlier RDFS & OWL versions.

modelling work had proved quite successful in revealing gaps between existing systems, it did not readily enable the modelling of likely solutions, i.e. how the information held in different systems could be shared. Due to this need for an integrative metadata framework, EH have built a supplementary ontology (CRM-EH), representing the broader archaeological processes in considerable detail and complexity by extending the basic CIDOC CRM standard.

The CRM-EH comprises 125 extension sub-classes and 4 extension sub-properties. It is based on the archaeological notion of a *context*, modelled as a place, from which the constituent *context stuff* has been removed by a series of archaeological *events*. It includes entities to describe stratigraphic relationships and phasing information, finds recording and environmental sampling [8], [9], [10].

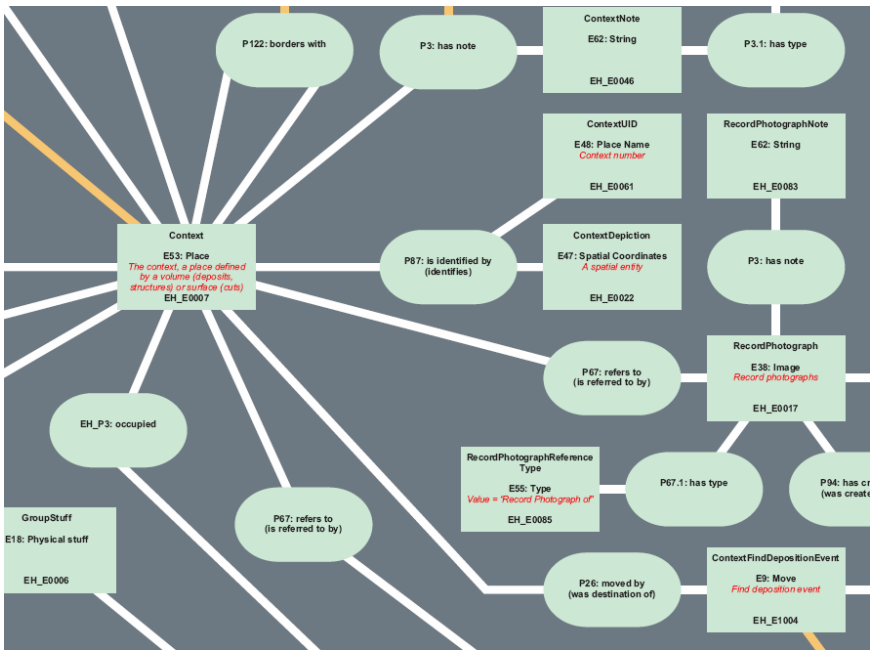

Fig. 1. Extract of English Heritage CRM-EH ontological model [8]

The intention is that a common ontology of shared meanings will provide greater semantic depth and potential for cross-domain searching by researchers within and beyond the archaeological sector. However work to date has focussed on modelling. The potential of this CRM-based extension ontology for assisting archaeological search and information extraction has not been investigated and this is one aim of the STAR research project.

3 Data Mapping

The EPOCH Network of Excellence AMA project is working on an open source tool that is intended to assist with mapping different archaeological and cultural heritage datasets to the CIDOC CRM, as a common format for interoperability [11].

Collaborative USA/German research conducted an exercise intellectually mapping the Perseus Project art and archaeology database to the CIDOC CRM and this mapping work is ongoing with the German Arachne archaeology database [12], [13]. This work discusses the potential for interoperability in a common underlying framework but highlights (in the application data considered) the need for data cleansing, common identifiers and semi-automated mapping tool assistance. They also discuss the need to explicitly model events in application workflows where that is implicit, in order to conform to the event-based CRM. The possibility of combining SKOS with the CRM is observed.

The BRICKS FP6 IP project [14], [15] stress that mappings from one dataset to another or to a common framework require intellectual work by domain experts. Their approach employed spreadsheets to intellectually define mappings from two different archaeological databases to the CIDOC CRM. These are semi-automatically transformed to XSL style sheets, which transform the data to the desired representation. BRICKS' experience in mapping different cultural heritage datasets to the CIDOC CRM encountered difficulties with the abstractness of the concepts resulting in consistency problems for the mapping work. This resulted in different mappings for the same underlying semantics and in different data objects being mapped to the same CRM entity. They pointed out a need for additional technical specifications for implementation modelling purposes. The abstractness of the CRM and the lengthy relationship chains arising from the event-based model also raised issues for designing appropriate user interfaces.

These various issues arising from detailed data mapping exercises also surfaced in the mapping/extraction phase of the STAR project and are explored below.

3.1 STAR Data Mapping Exercise

Five databases were identified as initial candidates for use within the STAR project:

- Raunds Roman Analytical Database (RRAD)
- Raunds Prehistoric Database (RPRE)
- Integrated Archaeological Database (IADB)
- Silchester Roman Database (LEAP)
- Stanwick sampling data (STAN)[3]

Each database was structured according to its own unique schema. Data coverage for the areas of archaeological activity represented by the CRM-EH ontological model

[3] The Stanwick sampling data actually represented part of the RRAD database, so the two databases were merged to enable easier subsequent data extraction.

varied considerably. By far the largest database was RRAD, however all databases contained rich information that will be of interest for the purposes of the STAR project. A design decision was taken to export the databases to a common structure, representing the information selected to be exposed for STAR Project purposes, as RDF triples[4].

The creation of initial mappings between database columns and RDF entities was a manual exercise undertaken with the benefit of domain knowledge from English Heritage. A spreadsheet of table/column names and their corresponding CRM-EH entities was produced by EH for the RRAD database. Although incomplete it provided enough information to allow many key data items to be extracted. It also allowed the Glamorgan development team to extrapolate the mappings to the other databases once the principal entities and properties of archaeological databases were more clearly understood. Subsequent Glamorgan mapping work was verified by EH in an iterative collaborative process.

4 Data Extraction

Mapping and data extraction are time-consuming and non-trivial exercises with great potential for error. A bespoke utility application was therefore created to assist with the process of data mapping, cleansing and extraction (further discussed in Section 4.4). The application allows mapping of RDF entities to database columns, construction of structured SQL queries (incorporating data cleansing functionality), and output to RDF data files. RDF data entities require unique identifiers, so key to this process was the adoption of a consistent convention for unique naming of entities.

4.1 Creation of Unique Identifiers

From the results of the mapping exercise it was found that some data would have to be an amalgamation of values from separate tables. It was therefore necessary to devise a scheme beyond just using the row ID from an individual table. In addition the data for multiple CRM-EH entity types were sometimes derived from a single table and so exhibited a 1:1 relationship - but required distinct unique identifier values. Finally, the data obviously originated from multiple databases so 'unique' identifiers were still potentially ambiguous. The identifier format adopted to deal with each of these issues was a prefixed, dot delimited URI notation, allowing the reuse of the existing database record ID values without introducing ambiguities:

prefix#entity.database.table.column.rowID
e.g. "http://tempuri/star/base#EHE0007.rrad.context.contextno.110575"

4 Not all data was deemed relevant for the STAR Research Demonstrator, which is a Demonstrator of cross search across digitally published archaeological data for scientific purposes, rather than administrative issues or immediate excavation analysis.

A temporary URI prefix (*http://tempuri/star/base#*) was added to all identifier values. Later in the project this will be globally replaced with a more persistent domain prefix.

In some instances no suitable numeric row ID was available on a table. In this case the unique identity field on a row would be comprised of textual data that could result in an invalid URI, so this necessitated XML encoding of any data used as part of an identifier.

4.2 Modelling of Events

Both CRM-EH and CRM are event based models. Events defined in the models and used to interconnect objects and places etc. were often only implicit within the original relational database structures and in the mappings created. E.g. in the CRM-EH model, *finds* would be measured via a *measurement event* resulting in *measurements*. In the translation from relational databases to an RDF graph structure it was necessary to create this event information by the formation of intermediate 'virtual' entities - data that did not necessarily explicitly exist in the underlying datasets but was required to correctly model the interconnection of entities in the resultant RDF graph.

4.3 Modelling of Data Instance Values

Being a higher level conceptual model the CRM has little intrinsic provision for the representation of actual data instance values. The approach adopted for the STAR data extraction process was to create rdf:value relationships as an additional property to model instance data for entities wherever appropriate.
(E.g. *crmeh:EHE0022.rrad.context.contextno.110575 rdf:value "98000E 56879N*").

As was experienced with the unique identifiers, some of the descriptive text fields contained problematic characters; in fact some contained HTML mark-up, so it was again necessary to encode this data to avoid producing potentially invalid data files.

4.4 Data Mapping and Extraction Utility

The data mapping information described in Section 3 was used to guide query formulation using a bespoke mapping/extraction utility to extract archaeological data conforming to the mapping specified (see **Fig. 2**). The utility consists of a form allowing the user to build up a SQL query incorporating selectable consistent URIs representing specific RDF entity and property types (including CRM, CRM-EH, SKOS, Dublin Core and others). The query is then executed against the selected database and the resultant data is displayed in tabular form (to check that the results

are as expected). This tabular data is then written directly to an RDF format file (see **Fig. 3**), and the query parameters are saved in XML format for subsequent reuse.

Although the mapping/extraction utility is a bespoke tool written specifically for the STAR project it would require minimal rework to extract data from most relational databases, using a configurable ODBC connection string.

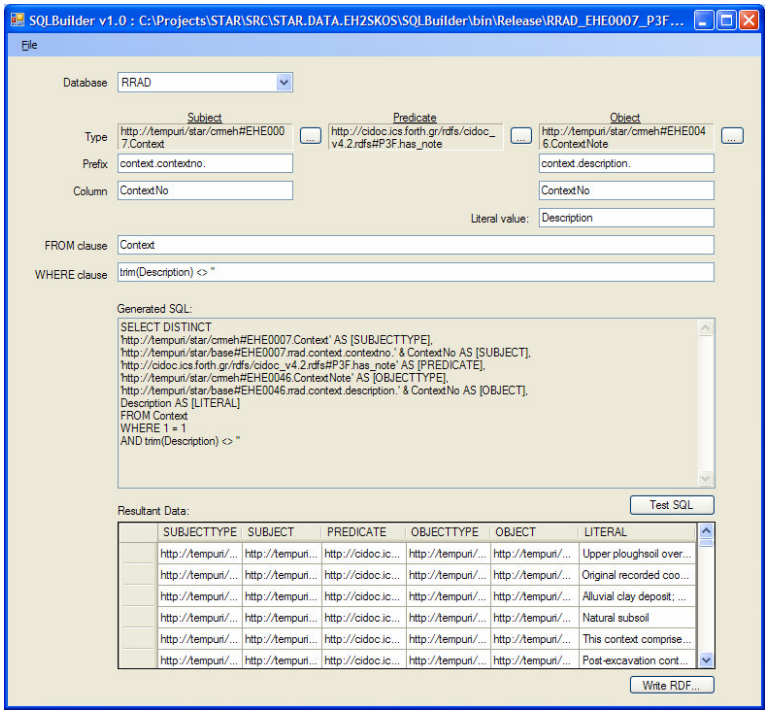

Fig. 2. The data mapping and extraction utility. A query has been built and tabular data has been extracted from the selected database and displayed.

4.5 Modular Approach Adopted

The mapping/extraction tool facilitated building and saving queries for the extraction of RDF data from the five databases. Each query resulted in the extraction of data instances conforming to discrete modular parts of the CRM-EH ontological model. This allowed the data extracts to be later selectively combined as required, and for any query to be revised and re-run if necessary. This assisted in improving overall co-ordination and consistency, preventing the process from becoming unnecessarily complex and unwieldy.

Files containing extracted data were named according to the relationships they contained. E.g. file *RRAD_EHE0007_P3F_EHE0046.rdf* would contain all extracted

data for the relationship *EHE0007.Context→P3F.has_note→EHE0046.ContextNote*, taken from the RRAD database. A total of 305 RDF files were created in this way for the initial extraction exercise.

```
<?xml version="1.0"?>
<rdf:RDF xml:base="http://tempuri/star/base#"
         xmlns:crm="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#"
         xmlns:crmeh="http://tempuri/star/crmeh#"
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<crmeh:EHE0007.Context
rdf:about="http://tempuri/star/base#EHE0007.rrad.context.contextno.1">
<crm:P3F.has_note>
<crmeh:EHE0046.ContextNote
rdf:about="http://tempuri/star/base#EHE0046.rrad.context.description.1">
<rdf:value>Upper ploughsoil over whole site no Sub-division for the convenience of
finds processing '1' contains finds contexts '3759', '3760' and '3763'.</rdf:value>
</crmeh:EHE0046.ContextNote>
</crm:P3F.has_note>
</crmeh:EHE0007.Context>
   Etc.
```

Fig. 3. RDF data is automatically generated by the extraction utility and written to a file.

5 Utilising the Extracted Data

Recalling that the original aim of the exercise was to demonstrate the potential benefits in cross searching data conforming to a common overarching conceptual structure, the extracted data was next imported into a MySQL RDF triple store database, using the SemWeb RDF library [16]. At this point any entity/statement duplication was resolved, and any gross errors with RDF/XML formatting would be readily highlighted (no errors of this kind were actually encountered - another benefit of using a consistent data extraction tool). When imported into the SemWeb MySQL triple store database the combined data files produced the following results:

Table 1. Statistics for extracted data

Database	Entities	Literals	Statements
RRAD (inc. STAN)	919,017	126,691	2,383,216
RPRE	114,105	20,482	317,085
IADB	85,694	21,592	209,582
LEAP	30,066	7,954	78,122
Totals:	**1,148,882**	**176,719**	**2,988,005**

5.1 Prototype Search / Browse Application

An initial prototype client application was produced (see **Fig. 4**), capable of cross searching and exploring the amalgamated data extracted from the previously separate databases. The application utilises a bespoke CRM based web service for all server interaction (the underlying SemWeb library does also support SPARQL querying). Boolean full-text search operators facilitate a measure of query refinement and result ranking. Retrieved query results are displayed as a series of entry points to the structured data; it is then possible to browse to other interrelated data items, by following chains of relationships within the CRM-EH, beaming up from data items to concepts as desired.

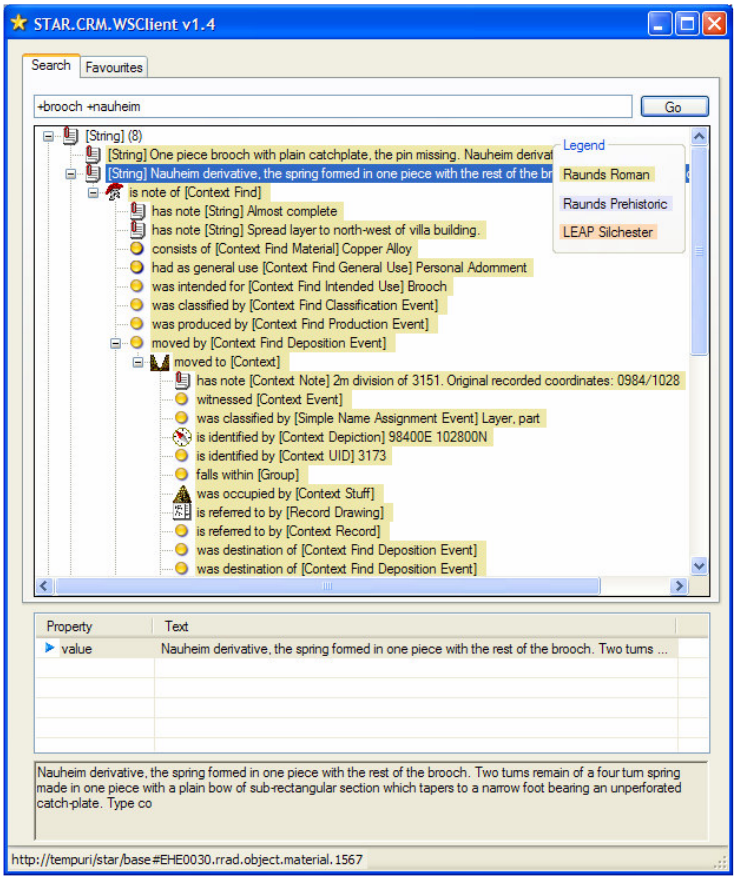

Fig. 4. Initial prototype search and browse application

Fig. 4 shows an example of a search for a particular kind of brooch using Boolean full-text search operators. One of the retrieved results has been selected and double-clicked to reveal various properties and relationships to further entities and events, any of which may then be double clicked to continue the browsing. Local browsing of the CRM-EH structured data can immediately reveal a good deal of information about the find e.g. a description, a location, the material it was made of, it's condition, how it was classified by the finds specialist, various measurements, the constituents of the surrounding soil, other finds in the immediate vicinity etc.

6 SKOS-Based Terminology Services

To complement the CRM based web service used by the search / browse application described in Section 5, the project has also developed an initial set of terminology services [17], based upon the SKOS thesaurus representation [18], [19]. The services are a further development of the SKOS API [20] and have been integrated with the DelosDLMS prototype next-generation Digital Library management system [21]. Functionality includes a facility to look up a user provided string in the controlled vocabularies of all KOS known to the server, returning all possibly matching concepts. The ability to browse concepts via the semantic relationships in a thesaurus is provided, along with semantic expansion of concepts for the purposes of query expansion [22]. The experimental pilot SKOS service is currently available on a restricted basis (see http://hypermedia.research.glam.ac.uk/kos/terminology_services) operating over EH Thesauri [23], and a demonstration client application is also available.

7 Conclusions

This paper discusses work in extracting and exposing archaeological datasets (and thesauri) in a common RDF framework assisted by a semi-automatic custom mapping tool developed for the project. The extensions to the CRM and the mapping/extraction tool have potential application beyond the immediate STAR project. The viability of the approach is demonstrated by implementations of CRM and SKOS based web services and demonstrator client applications. The initial prototype client application demonstrates useful cross searching and browsing functionality and provides evidence that the data mapping and extraction approach is viable. The next phase of the project will investigate interactive and automated traversal of the chains of semantic relationships in an integrated data/concept network, incorporating the EH thesauri to improve search capability.

Recent mapping exercises by the BRICKS and Perseus/Arachne projects from databases to the CIDOC CRM (see Section 3) have highlighted various issues in detailed mappings to data. Some findings are replicated by the STAR experience to date. Semi-automated tools improved consistency in mapping and data extraction work, although intellectual input from domain experts was still necessary in identifying and explaining the most appropriate mappings. Data cleansing and a

consistent unique identifier scheme were essential. In some cases, it was necessary to explicitly model events not surfaced in data models, in order to conform to the event-based CRM model. As with BRICKS, it proved necessary to create technical extensions to the CIDOC CRM to deal with attributes required for practical implementation concerns.

STAR experience differs from previous work regarding the abstractness of the CRM. The EH extension of the CRM (the CRM-EH) models the archaeological excavation/analysis workflow in detail and this is a distinguishing feature of the STAR project. The ambiguity of mappings from data to the CRM has not arisen to date in STAR. While this may be due to the more detailed model of the archaeological work flow, unlike BRICKS all the mappings were performed by the same collaborative team. However, a tentative conclusion to date is that a more detailed model does afford more meaningful mappings from highly specific data elements than the (non-extended) CRM standard. The object oriented CRM structure is intended to be specialised for particular domains and the representation of both the CRM-EH extension and the technical extensions of the CRM as separate RDF files offers a convenient route for integrating optional extensions to the standard model. The CRM-EH extension is the result of a significant effort, and the cost/benefit issues around the granularity of modelling for cross dataset search and more specific retrieval, along with user interface issues, will be a key concern in the next phase of STAR project work.

Acknowledgements
The STAR project is funded by the UK Arts and Humanities Research Council (AHRC). Thanks are due to Phil Carlisle (English Heritage) for assistance with EH thesauri.

References

1. Bergman, M.K.: The Deep Web: Surfacing Hidden Value. BrightPlanet Corp. White Paper (2001), http://www.brightplanet.com/images/stories/pdf/deepwebwhitepaper.pdf
2. Patel M., Koch T., Doerr M., Tsinaraki C.: Report on Semantic Interoperability in Digital Library Systems. DELOS Network of Excellence, WP5 Deliverable D5.3.1. (2005)
3. STAR Project: Semantic Technologies for Archaeological Resources, http://hypermedia.research.glam.ac.uk/kos/star
4. CIDOC Conceptual Reference Model (CRM), http://cidoc.ics.forth.gr
5. RDFS Encoding of the CIDOC CRM, http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs
6. Doerr M., Hunter J., Lagoze C.. Towards a Core Ontology for Information Integration. Journal of Digital Information, 4 (1), http://journals.tdl.org/jodi/article/view/jodi-109/91 (2003)
7. English Heritage http://www.english-heritage.org.uk/
8. English Heritage Ontological Model http://cidoc.ics.forth.gr/docs/AppendixA_DiagramV9.pdf
9. May, K.: Integrating Cultural and Scientific Heritage: Archaeological Ontological Modelling for the Field and the Lab. CIDOC CRM Sig Workshop, Heraklion (2006) http://cidoc.ics.forth.gr/workshops/heraklion_october_2006/may.pdf

10. May K.: Report on English Heritage Archaeological Application of CRM. CIDOC CRM Sig Workshop, Edinburgh (2007)
11. EPOCH Archive Mapper for Archaeology Project. http://www.epoch.eu/index.php?option=com_content&task=view&id=222&Itemid=338
12. Babeu, A., Bamman, D., Crane, G., Kummer, R., Weaver, G.: Named Entity Identification and Cyberinfrastructure. 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL07) Budapest, 259-270 (2007)
13. Kummer, R.: Towards Semantic Interoperability of Cultural Information Systems - Making Ontologies Work. MA Thesis. University of Koln (2007), http://old.hki.uni-koeln.de/studium/MA/MA_kummer.pdf
14. Nußbaumer, P., Haslhofer, B.: CIDOC CRM in Action – Experiences and Challenges. Poster at 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL07), Budapest (2007) http://www.cs.univie.ac.at/upload//550/papers/cidoc_crm_poster_ecdl2007.pdf
15. Nußbaumer, P., Haslhofer, B.: Putting the CIDOC CRM into Practice – Experiences and Challenges. Technical Report, University of Vienna (2007) http://www.cs.univie.ac.at/publication.php?pid=2965
16. SEMWEB RDF Library for .NET, http://razor.occams.info/code/semweb
17. Tudhope D., Koch T., Heery R.: Terminology Services and Technology: JISC State of the art review (2006) http://www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_technology_review_sep_06.pdf
18. Miles A., Matthews B., Wilson M.: SKOS Core: Simple Knowledge Organisation for the Web, Alistair Miles, Brian Matthews and Michael Wilson, Proceedings of the International Conference on Dublin Core and Metadata Applications, 5-13 (2005)
19. SKOS: Simple Knowledge Organization Systems, http://www.w3.org/2004/02/skos
20. SKOS API. SWAD_EUROPE Thesaurus Project Output (2004) http://www.w3.org/2001/sw/Europe/reports/thes/skosapi.html
21. Binding C., Brettlecker G., Catarci T., Christodoulakis S., Crecelius T., Gioldasis N., Jetter H-C., Kacimi M., Milano D., Ranaldi P., Reiterer H., Santucci G., Schek H-G., Schuldt H., Tudhope D., Weikum G.: DelosDLMS: Infrastructure and Services for Future Digital Library Systems, 2nd DELOS Conference, Pisa (2007) http://www.delos.info/index.php?option=com_content&task=view&id=602&Itemid=334
22. Binding C., Tudhope D.: KOS at your Service: Programmatic Access to Knowledge Organisation Systems. Journal of Digital Information, 4(4), (2004) http://journals.tdl.org/jodi/article/view/jodi-124/109
23. English Heritage Thesauri http://thesaurus.english-heritage.org.uk/
24. Doerr, M.: The CIDOC Conceptual Reference Module: an Ontological Approach to Semantic Interoperability of Metadata. AI Magazine, 2493, 75--92 (2003)
25. Cripps P., Greenhalgh A., Fellows D., May K., Robinson D.: Ontological Modelling of the work of the Centre for Archaeology, CIDOC CRM Technical Paper (2004) http://cidoc.ics.forth.gr/technical_papers.html