Following a STAR? Shedding More Light on Semantic Technologies for Archaeological Resources

Keith May,[1] Ceri Binding,[2] and Doug Tudhope[2]

[1] English Heritage, Strategy Dept, Fort Cumberland, Portsmouth. UK.

[2] Glamorgan University, Faculty of Advanced Technology, South Wales.UK.

Abstract

The Semantic Technologies for Archaeological Resources (STAR) project has been investigating the use of a number of emerging semantic web technologies for developing interoperability between existing data from archaeological projects in legacy systems and new project data entered in new systems, along with other data sets from previously unrelated archaeological recording systems. Initial work began at English Heritage based on the CIDOC CRM ontology for cultural heritage and the creation of archaeological domain specific extensions to the CIDOC CRM ontology for the modeling of more specific archaeological information recorded during the evaluation, excavation and post-excavation processes. The CRM modeling work has now been mapped to a number of different data sets from various derivations beginning with some from English Heritage projects, but in addition including data from other organizations in a number of different data structures and distinct formats. Work has also been carried out to incorporate domain thesauri into the project's ontological framework and the development of tools. The conceptual modeling and mappings have then been used to generate RDF triple statements using a semi-automated process, and a purpose-built data-extraction tool with the resulting RDF statements held in a triple store. The archaeological extensions (referred to as CRM-EH) have been made available in RDF format from the STAR web site (http://hypermedia.research.glam.ac.uk/kos/CRM/). This paper will set out some of the most recent findings from the STAR project, including presentation of the latest web service interfaces. It will also look at some of the main pros and cons encountered in the project work to date and try to assess the degree of interoperability provided between the different data sets and some of the cost-benefits associated with mapping the various datasets using the Conceptual Reference Models.

Keywords*: semantic interoperability, ontology, conceptual reference modeling, CIDOC CRM, CRM-EH, SKOS*

1 Background to Ontological Modeling & STAR

Further details of the development work on the ontological modeling and mapping for the STAR project have been described in CAA2008.[1] Using the 125 extension entities resulting from the archaeological ontological modeling[2] based on the CIDOC CRM, a number of further tools, prototype web based interfaces, and web services have been developed—and following feedback from user requirements workshops are still being developed and refined—to access and query data held in the RDF triple store. The STAR project has developed an initial set of semantic web services, incorporating the emerging W3C SKOS standard for thesauri representation online.[3] Several thesauri in general use within EH have been converted to the SKOS format for use in query expansion searching of controlled vocabulary fields and further work is being carried out to create SKOS versions of glossary fields and other terminologies.[4] This paper concentrates on the outcomes of the ontology based data extraction, presentation, and querying aspects of the project.

2 STAR Architecture and Technologies

The project has taken a number of archaeological data sets that were recorded using different, but related, archaeological recording systems and, using an over-arching ontological model based upon the CIDOC-

[1]K. May, C. Binding, and D. Tudhope, "A STAR is Born: Some Emerging Semantic Technologies for Archaeological Resources," *Proceedings: Computer Applications and Quantitative Methods in Archaeology*, Budapest 2008 (forthcoming).

[2]Paul Cripps et al., "Ontological Modeling of the Work of the Centre for Archaeology," *CIDOC CRM Technical Paper*, 2004. http://cidoc.ics.forth.gr/technical_papers.html.

[3]A. Miles, B. Matthews, and M. Wilson, "SKOS Core: Simple Knowledge Organisation for the Web," *Cataloging and Classification Quarterly* 43 (2007): 69–84.

[4]D. Tudhope, C. Binding, and K. May, "Semantic Interoperability Issues from a Case Study in Archaeology," in *Semantic Interoperability in the European Digital Library*, ed. Stefanos Kollias and Jill Cousins (Proceedings of the First International Workshop, SIEDL, 2008) 88–89.

CRM, has mapped these datasets to common conceptual entities within the ontology. By exporting CRM-EH based entities and relationships as triple statements in a common RDF format, the resulting RDF triple store can be searched and interrogated using query languages such as SPARQL, and a number of web services have been developed for serving the data and enabling its querying and searching using prototype application interfaces (see fig. 1).

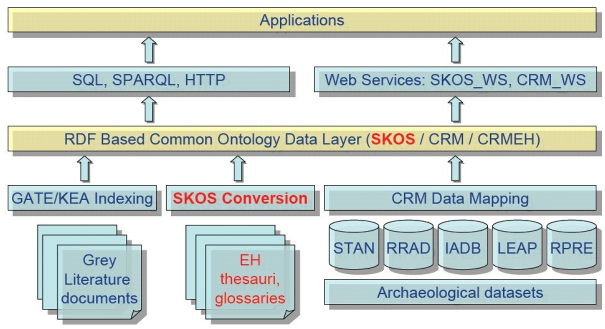

Figure 1. *STAR project general architecture*.

The project is investigating how the use of the W3C SKOS standard for controlled vocabularies can improve search retrieval mechanisms and how Information Extraction techniques incorporating the CRM-EH ontology might be used to create semantic "bridges" between grey literature reports and related data. For the purposes of this paper we will focus upon outputs and interfaces developed for the results of the CRM data mapping and extraction elements of STAR.

Data mapping between the CRM-EH ontological model and a number of archaeological data sets was initially produced in simple spreadsheet format. The entities and relationships were further transposed into RDF format, which could be navigated, managed and edited as necessary using Protégé ontology editing software. In order to produce the actual instances of RDF statements based upon the "Subject—Predicate—Object" triple statements represented by the CRM-EH modeling, a bespoke mapping/extraction utility has been developed to extract archaeological data conforming to the mapping specified in a semi-automated manner (see fig. 2).

The utility consists of a form with entry boxes corresponding in turn to the Entity-Relationship-Entity elements of the CRM-EH statement. As the parts of the statement are created, the user can view the resulting SQL query building up in the "Generated SQL" display panel. This form-based interface enables the user to build up an 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.

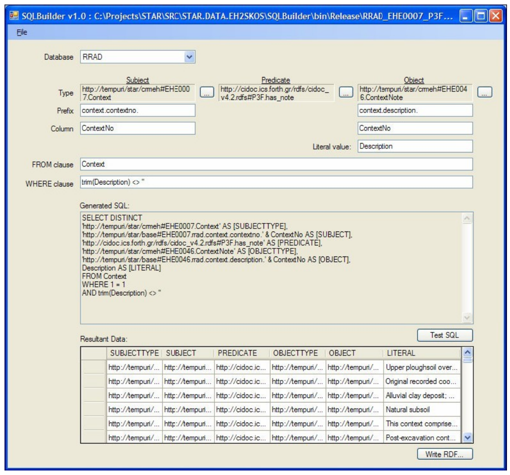

Figure 2. *The data mapping and extraction utility. A query has been built and tabular data has been extracted from the selected database and displayed.*

```
<?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.con
textno.1">
<crm:P3F.has_note>
<crmeh:EHE0046.ContextNote
rdf:about="http://tempuri/star/base#EHE0046.rrad.context.des
cription.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.
```

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

Although the mapping/extraction utility is a bespoke tool written specifically for the STAR project, it would not require a great deal of reworking to extract data from most relational databases, using a configurable ODBC connection string.

For ease of identification and cross-checking, the files containing extracted data were named according to the relationships they contained. For example, file *EHE0007_P3F_EHE0046.rdf* would contain all the extracted data for the RDF triple relationship *EHE0007.Context→P3F.has_note→EHE0046.Context*

Note. Using this extraction process on the four main datasets processed to date, in the region of 305 RDF files were created representing all the main "subject-predicate-object" triples expressed in the CRM-EH model.

As a guide to the scale of the resulting triple store and by way of comparison with existing archaeological datasets, when the data extraction process was applied to the four main datasets used so far by STAR, it resulted in just under three million triple statements being produced (table 1).

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**

Table 1. *Statistics for extracted data.*

It is roughly estimated that the four databases between them contained approximately 25,000 context records with associated finds and environmental data.[1]

3 GRANULARITY OF ONTOLOGICAL MODEL AND RESULTING DATA MAPPING

While the project has produced RDF statements for all the relationships that are represented in the CRM-EH model, that is not the same thing as having data from all the datasets that map to each of those RDF statements. The mapping process tends to look for commonality (interoperability) between different datasets, and thereby tends to focus on "core" data concepts that are common to all the systems involved. This has led to the identification of a loosely termed CRM-EH "core" of concepts, which seem to be most readily identifiable across different archaeological recording systems and which hold relationships that are most central to interoperability within and between different datasets, at least within the current scope of the STAR project (see also discussion of cost-benefits in section 6 below).

The hierarchy of the granularity is not formally represented in the modeling at present and it remains an area of possible further investigation as to whether, and if so how, to express the different granularity of conceptualization and details that are inherent in different conceptual entities. This may be an area for further development in the user interfaces, as it may reflect how archaeologists could conceive concepts to be structured for navigation.

4 USER REQUIREMENTS AND UNDERSTANDING THE INFORMATION LANDSCAPE

The STAR project has held a number of user focused surveys, workshops and trials in order to elicit feedback from users on the different ways they might want to navigate or search through the sort of semantically enabled information that STAR is generating. One immediate issue has been to find ways of presenting the complexity of the semantic inter-relationships between the data in a way that users can begin to get enough of a feel for the sorts of query that might be possible. To this end an initial prototype CRM browser was developed that enabled initial query entry of free-text search terms followed by the option to navigate the results of returned queries using a clickable, expandable, icon-enhanced interface (see fig. 4.)

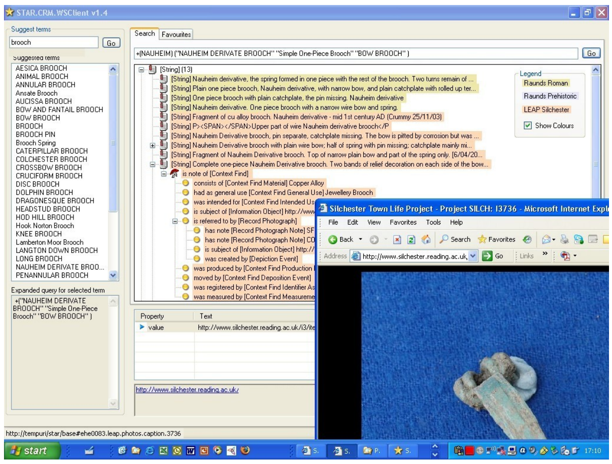

Figure 4. *Prototype CRM Browser showing expanded query results including hyperlinks to related data.*

This prototype browser interface has enabled the team to test and demonstrate the fundamental principle of interoperability between the various previously un-integrated datasets. It has also better enabled the project team to gather feedback and test examples of potential queries with archaeological end users. It has demonstrated the ability of the STAR CRM browser to generate query results used to link, via returned URL hyperlinks, to Silchester data running live on the server at Reading University. The prototype has also incorporated the SKOS based thesauri browsing interface to enable some disambiguation of terms for searching, although it still remains for this element to be incorporated automatically into the search and retrieval process.

Nevertheless, the first prototype interface still exhibited a number of the key presentational problems when trying to depict all the relationships represented in the ontology-based data in the triple store. First, it was

[1] C. Binding, D. Tudhope, and K. May, "Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM," *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries* (Aarhus: 2008) 280–290; *Lecture Notes in Computer Science* 5173 (Berlin: Springer, 2008).

necessary to distinguish between results that were being returned from different original datasets. In the prototype this was done by simply color coding the three principal data sets, so results for Raunds Roman project data were highlighted in yellow, Raunds Prehistoric data in grey, and Silchester LEAP data in red. In the example illustrated (see fig. 4) of a search for Nauheim brooch, the browser has returned all records that contain the term "Nauheim". The user can then search any one of these records and "drill" deeper to see that the result contains a note from a Context Find that has various associated records, including a reference to a photograph, which itself is the subject of an information object that is the URL link to an actual digital photograph on the Silchester Project[1] online database at Reading University. While it is important to be aware of these interrelationships for demon-strating the principle of interoperability, it is probably unrealistic to expect most users to directly browse the full ontological set of relationships (although that might be an option for an advanced interface) and the team is currently working on a simpler user search and retrieval interface.

The results of the user testing suggest that users will want to search in at least two significantly different ways (and probably many more if possible) through archaeological data. The two distinct types of query can be categorized as Inter-site and Intra-site queries. For Inter-site queries researchers want to do things like comparing types of site at local, regional, or national scale; or asking which projects across a country/region/area have evidence for excavated prehistoric field systems; or which regions have finds of certain types of pottery or certain types of samples from given types of deposits. On the other hand, Intra-site queries will search on information derived from just one particular site but may then query on complex interrelationships between different data held about that particular site. Examples of Intra-site queries might include: interrogating and understanding relationships between Phases and Groups of features within a specific site; finding and comparing examples of types of objects from a particular site; or comparing the contents of different samples taken from different areas within a site.

How people carry out such searches or navigate around this "information landscape" bears some similarities to how they might traverse a transportation system. One analogy is a train system where "mainline" routes are established for frequently recurring direct travel between major stations, while "branch" or "local" routes are used for more "neighborhood" connections and are perhaps less frequently travelled by the majority of users. As with train journeys, some queries might go partly via a "mainline" route but then switch to a "local" route for a more localized piece of information. Of course this does not preclude that some people might chose to reach their destination by other forms of transport such as car, boat, or plane or their information retrieval equivalents!

5 Examples of short-cutting search and query interfaces

A particular approach for how to navigate through the inter-relationships of the CRM-EH modeled data was reinforced by the user testing. The idea is to present short-cuts to the user for traversing the commonly followed relationships between key entities in the CRM-EH. This also builds upon the analogy of building "mainline" routes between commonly traversed enquiries.

For this purpose of short-cutting we have focused on the concept of an archaeological context and its most closely associated key relationships:

- Context => Relationships to => Find
- Context => Relationships to => Sample
- Context => Relationships to => Context (Stratigraphic, Spatial, Temporal)
- Group => Relationships to => Context

The resulting prototype interfaces enable a user to select first from the "core" concepts of Group, Context, Find, Sample, and then develop specific queries for entities and relationships that pertain to those specific core concepts. The example below (see fig. 5) shows a similar search for a Nauheim brooch when entered into the more structured query interface. The interface for enabling more directed querying seems promising, although more work is needed to improve on the presentation of the results of such searches; moreover, as with other search mechanisms, there may be some performance issues if extremely complex combinations of relationships are searched on.

An example of a similar use of short-cutting is given in the CIDOC CRM by the relationship "P53 has former or current location." P53 is a short-cut for a longer chain of events and relationships, which describes the movement of a physical object by a "move" event using either a "moved to" or a "moved from" relationship to a new spatial location. This more detailed representation has been used in the CRM-EH to model the way a finds object is deposited in a context as the result of a deposition event. It would also be possible using the P53 short-cut to model this, as the find has the relationship P53 "former or current location" defined by

[6]A. Clarke et al., "Silchester Roman Town Insula IX: The Development of a Roman Property c. AD 40–50-c. AD 250," *Internet Archaeology* 21 (2007).

its spatial coordinates within the broader spatial place defined by a context.[1]

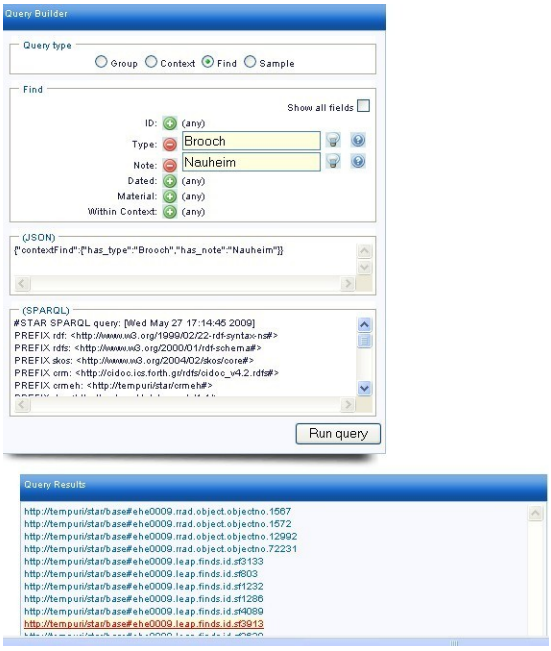

Figure 5. *Prototype Query Builder interface showing example search for "Nauheim Brooch".*

6 Investigation of Cost-Benefit Issues

At time of writing, the STAR project is only about two-thirds completed, so assessment at this stage of any outcomes remains provisional. Nevertheless, it was one of the project's aims to try to assess the cost benefits of the approaches, particularly in relation to wider application within the archaeological domain. This section sets out some of the major points for consideration and further investigation.

Beginning on the positive side, it already seems well demonstrated that the approach of using the CRM-EH extensions to the over-arching CIDOC CRM ontology has enabled the STAR project to implement complex cross-searching of otherwise un-integrated datasets. This is achieved without needing to alter the data in existing systems, although it does therefore require additional effort in exporting data from a number of systems into a commonly formatted RDF triple store. Fundamentally, the overarching aim of achieving genuine interoperability between otherwise unrelated archaeological datasets has been demonstrated.

Once the overall ontological model has been established and verified as applicable across the domain—a matter of not inconsiderable effort—and users are able to identify that the model does effectively represent the fundamental relationships inherent in their data, the work required to then map further archaeological datasets is more straightforward (although not without some additional intellectual endeavor). The lack of commonly used methodologies or easy to use off-the-shelf software to help with this is an issue, but as a baseline, producing a simple spreadsheet that shows data fields from archaeological databases mapped to CRM ontological entities is manageable. Indeed it can be seen as one of the major benefits of not needing to re-engineer existing databases, that users can export their data into an RDF triple store without compromising the existing structure of their data (i.e. they don't have to rebuild existing systems to benefit from using the CRM ontological modeling and RDF interoperability technologies).

The mapping of archaeological data fields to "core" concepts is relatively straightforward, once the person doing the mapping is familiar with the ontological model used and, if anything, it seemed to become easier as more datasets were addressed (although so far only five completely different archaeological data systems have been attempted). That said, it was still felt by those who were generous enough to share their data with the project that the mapping process was relatively specialized, and that they were happier for the STAR project team to map their data to the CRM-EH rather than attempt to do so themselves; there was a pragmatic issue here too about how much time others could give freely to the project. It does point to a more fundamental issue with the approach of mapping data to an ontology that it is meant to enable engagement with domain expertise. Often the required level of engagement from domain experts less familiar with the relevant ontology is a considerably steep learning curve to fully represent all aspects of their data. The development of the semi-automated data extraction utility may make the process of extracting data into RDF format a reasonably cost-effective approach to generating RDF versions of existing data sets by computer literate archaeological practitioners, but in its current form it is still relatively specialized technology.

This leads to further questions of how to quantify the relative benefits of the process. One principal improvement offered by this type of approach to cross-searching data is that it enables researchers to answer queries that would otherwise not have proved possible. The prototype query interfaces have shown an improved ability to interrogate across the different datasets, but as the nature of possible queries becomes more complex, so there are reciprocal demands to develop better interfaces to enable the user to assimilate and navigate around the complexity of returned results.

Another issue that will need to be investigated further is how best to assess the appropriate degree of interoperability that is necessary for different types of data resource. Two particular aspects of this have been highlighted by STAR. First, in order to test the viability

[1] Nick Crofts et al., ed. "Definition of the CIDOC Conceptual Reference Model," March 2009. http://cidoc.ics.forth.gr/official_release_cidoc.html.

of the archaeological ontological modeling, the project has mapped data sets at different stages in the archaeological process. This resulted in data from the excavation stage, the analysis stage, and the publication/dissemination stage of projects being included in the resulting triple store. While this enables querying across different aspects of an archaeological record for the same site within the parameters of the STAR demonstrator, it does raise the practical issue of how to present pre-published data results in a web-of-data alongside either partially analysed or fully published information. This is principally a data management issue and might be addressed either by only exporting data from similar stages of work to the same triple stores for searching, or by agreeing on some protocols for identifying the stage at which data is made available (e.g., is it interim data, un-synthesised analysis, or final publication). The second aspect of this issue to consider is how to reflect the level of detail, or granularity of the data, that is made interoperable. To date STAR has focused on "core" data fields from the datasets that it has mapped to (see section 3). This means that mainly the principal context, finds, samples and group/phase records from the projects have been mapped to. It was felt that this enabled a good level of querying about the broader nature of each site (e.g., what finds records came from round houses or trackways of a certain period) while also allowing quite detailed queries (e.g., what types of seeds were recorded in samples taken from corn-drying ovens). But there were still areas of each dataset that remained un-mapped and therefore un-integrated. To some degree this was a pragmatic choice by the project to enable further datasets to be tested in the time available, rather than attempting to extend the ontological modeling to map to every single data field in the datasets used by the project. It also reflects the general level of granularity expressed by the CRM-EH as it currently stands. It was always known that, despite the relative breadth and complexity of the existing model, that it by no means models all data fields held within EH archaeological recording systems—let alone other organizations' recording systems.

One further qualifier that may have most influence on how and when data is made available for integration might come from the fact that the majority of archaeological data (certainly in the UK) is recorded by the commercial sector. Such data is therefore less likely to be made available across the sector until formally published. One might therefore envisage two stages of use of the technologies in the future. The first stage might be used by single organizations to create an RDF repository for their own projects, in order to enable cross-project and intra-site searching and analysis of related data. The second stage might then be to disseminate "published" RDF data to more public repositories once analysis and publication has been completed.

7 CONCLUSIONS TO DATE AND FURTHER WORK

Interoperability between a range of previously un-integrated archaeological datasets has been achieved using the CIDOC CRM based archaeological extensions of the CRM-EH ontological modeling. Work continues to map other data sets to the CRM-EH and CIDOC CRM, although as noted above there may be issues to be resolved with final dissemination of currently unpublished datasets; the ideal scenario would still be for other organisations to map their own data to the CIDOC CRM and CRM-EH to enable further integration and cross-searching.

Further work needs to be done on testing the scalability of both the hardware and software solutions for greater numbers and sizes of resource. To fully demonstrate the cost effectiveness of these approaches, there is a need for further data to be made available in a format that is mapped to the same ontological model to allow more comprehensive searching across a critical mass of relevant data. The current position with semantic interoperability brings to mind Metcalfe's law for Ethernet networks, which is usually stated as "the value of a telecommunications network is proportional to the square of the number of users of the system," but the reality for networks of internet proportions is likely to be much more complex.[1]

There are also other broader issues that do not just pertain to the interoperability or semantic searching of archaeological data. One of the main issues still to be resolved on a global level for the internet is how to handle persistent URI references, so that the referencing of information items such as within the STAR demonstrator can be maintained in a sustainable manner for others to cross-reference. There are also further issues to be addressed about how scaleable the triple store infrastructure for RDF data is, how scaleable the infrastructure for this would be across the web, and how the query interfaces will perform when faced with multiple inter-relationships.

The STAR project is working towards building an online demonstrator and the project team is aiming to publish the server software as open source by the end of the STAR project.

[1]Simeon Simeonov, "Metcalfe's Law: More Misunderstood than Wrong?" *High Contrast: Innovation & Venture Capital in the Post-broadband Era* (July 26, 2006).

ACKNOWLEDGEMENTS

The STAR project is funded by the UK Arts and Humanities Research Council (AHRC). The views expressed here are principally our own, but they are based on the research work of various domain experts at English Heritage, including Gill Campbell, Vicky Crosby and Phil Carlisle. We would particularly like to acknowledge Paul Cripps for his hard work in the initial development of the archaeological ontological model and Mike Rains for providing valuable feedback and help with Silchester LEAP data.

BIBLIOGRAPHY

Binding, C., D. Tudhope, and K. May. "Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM," in *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries,* Aarhus: 2008, 280–290; *Lecture Notes in Computer Science* 5173 (Berlin: Springer, 2008).

Clarke, Amanda, Mike G. Fulford, Mike Rains, and Klare Tootell. "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).

Cripps, Paul, Anne Greenhalgh, Dave Fellows, Keith May, and David-Earle Robinson. "Ontological Modeling of the work of the Centre for Archaeology." *CIDOC CRM Technical Paper*, 2004.

Crofts, Nick, Martin Doerr, Tony Gill, Steve Stead, and Matthew Stiff, eds. "Definition of the CIDOC Conceptual Reference Model." March 2009. http://cidoc.ics.forth.gr/official_release_cidoc.html.

May, K., C. Binding, and D. Tudhope. "A STAR is Born: Some Emerging Semantic Technologies for Archaeological Resources." *Proceedings Computer Applications and Quantitative Methods in Archaeology* (CAA2008). Budapest, 2008 (forthcoming).

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

Simeonov, Simeon. "Metcalfe's Law: More Misunderstood than Wrong?" *High Contrast: Innovation & Venture Capital in the Post-broadband Era* (July 26, 2006). http://simeons.wordpress.com/2006/07/26/metcalfes-law-more-misunderstood-than-wrong/.

Tudhope D., C. Binding, and K. May. 2008. "Semantic Interoperability Issues from a Case Study in Archaeology," in *Semantic Interoperability in the European Digital Library. Proceedings of the First International Workshop SIEDL 2008*, edited by Stefanos Kollias & Jill Cousins, 88–99.

WEB REFERENCES

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?CFID=186346&CFTOKEN=77468588.

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

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