LOOK OUT! CAN YOU SEE SOMETHING COMING? SEMANTIC BROWSING – WHAT MIGHT IT LOOK LIKE?

K. May [a], C. Binding [b], D. Tudhope [b]

[a] English Heritage, Strategy Department, Fort Cumberland, Portsmouth PO4 9DE UK.
- Keith.May@english-heritage.org.uk
[b] Faculty of Advanced Technology, Glamorgan University, Trefforest, Wales.
cbinding@glam.ac.uk - dstudhope@glam.ac.uk

KEYWORDS: Ontology, CIDOC CRM, CRM-EH, RDF, SKOS, Semantic Web, STAR

ABSTRACT:

The Semantic Technologies for Archaeological Resources (STAR) project has been exploring the development of new tools and technologies for "semantic web" based research. The project builds upon the ontological modelling approach taken by English Heritage staff, in recent years, to modelling their information and data using the CIDOC-CRM standard (ISO 21127:2006). The ontological modelling has enabled the EH archaeological teams to more explicitly identify where 'information gaps' exist within their existing information records and flow lines, which can then be bridged using an ontological information model. The data involved can be derived from legacy datasets, current databases, and hopefully will enable incorporation of a mapping to data yet to be recorded in a newly implemented archaeological recording system. The aim is to provide a model for how new systems and technologies can be developed that enable greater interoperability and better integration of data in the rather disparate archipelago of archaeological project information.

This paper will also discuss the emerging use of the Simple Knowledge Organization System (SKOS) data model as a W3 standard for sharing and linking knowledge organization systems via the Semantic Web. The STAR SKOS web services currently provide term look up across the thesauri held in the system (including the EH Monuments Thesaurus and the former MDA Objects Thesauri), along with browsing and semantic concept expansion within a chosen thesaurus. Users may browse a concept space to explore and become familiar with specialist terminology or as part of a broader application. 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'.

The paper will also introduce a new prototype CRM Browser web service developed by the STAR project and explore ideas for how this prototype browser might be developed further in the future to enable linked searching across and between free text reports and structured data in databases, using emerging forms of semantic query languages and interfaces and what such "Semantic Browsing" might look like for end users.

1. INTRODUCTION

1.1 Overview and structure of this paper

This paper will first recount some of the background to why the ontological modelling work was carried out by English Heritage archaeological staff and how this was carried forward by the Semantic Technologies for Archaeological Resources (STAR) project. It will then give a short summary of the conceptual modelling carried out and discuss in more detail the use of extensions of the CIDOC CRM to give greater definition of the main archaeological concepts involved. The paper will present an overview of the methods used by the current STAR project for mapping from the CRM-EH model to specific archaeological data sets, including the use of the emerging W3C standard SKOS (Simple Knowledge Organisation System) for implementing controlled terminologies in ontological modelling systems. The paper will then consider some of the current issues involved in formulating, gathering, representing and reflecting user requirements for a semantic browsing system, as yet to be produced. Finally there will be some consideration of the future directions for possible semantic web browsing and searching, followed by overall conclusions.

1.2 Background and aims of the Ontological Modelling

The English Heritage Revelation project (May, S., 2004) identified early in its assessment stage that EH was not lacking in archaeological information systems. Rather the picture was of an archipelago of self-contained and isolated islands of information that had been designed over the last twenty-five years or more. Most of these EH archaeological systems were designed to fulfil individual project requirements, but without the overall planning and structure to enable the shared use and interoperability of the data being collected or created.

The data flow diagrams and entity relationship models of the existing archaeological systems helped to give a clearer picture of the baseline state of affairs. But the resulting systems documentation was still a series of rather fragmented data models for each system without a clear method for how best to integrate the data held within each of them. It was decided to attempt to model both the existing information holdings but also include further new information requirements that would be wanted in a newly designed system. This was intended to express more explicitly where the gaps were, both in and between the existing data models, and most significantly showing where those gaps might be filled or "bridged" by

modelling new relationships between those bits and pieces of information. At that point attention was drawn to Semantic Web developments and in particular the CIDOC CRM (Crofts et al. 2003) and solutions that might be provided by an ontological approach to data modelling.

More of the background to the ontological modelling of the English Heritage (EH) archaeological information domain has been presented at the Computer Applications in Archaeology conference in Prato 2004 (Cripps & May 2004, forthcoming) and further publications and outputs are available from the CIDOC CRM website including an online version of the model and accompanying documentation (Cripps et al 2004, May, 2006). Since the beginning of 2008, and in this article, the English Heritage Conceptual Reference Model (CRM) has been referred to as the "CRM-EH", to distinguish it from the CIDOC CRM ontology from which it derives, and which it is still directly related to.

1.3 Background and aims of the STAR project

STAR is a 3 year Arts and Humanities Research Council (AHRC) funded project, in collaboration with English Heritage and the Royal School of Library and Information Science Denmark, applying semantic and knowledge-based technologies to the digital information of the archaeological domain. The project aims to develop new methods for linking archived and 'live' digital databases; associated vocabularies; and, where relevant, related grey literature, exploiting the potential of a high level core ontology (CRM-EH) and natural language processing techniques.

Increasingly within archaeology, the Web is used for dissemination of reports and the associated datasets that result from fieldwork or scientific analysis of material from historic environment investigations. This contributes to the growing amount of information on the 'deep web', which a recent Bright Planet study estimated to be 500 times larger than the 'surface web'. (Bergman, 2001). 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. A higher and higher proportion of recent archaeological results and reports are appearing as grey literature, increasingly online, before or instead of traditional publication. Typically these are not indexed or made available for searching other than as ordinary web documents. It is difficult using conventional search engines to link these to datasets or indeed to search them using terminology other than that employed by the authors. Cultural heritage and memory institutions generally are seeking to expose databases and repositories of digitised items - previously confined to the realm of specialists - to a wider academic and general audience. The mapping from lay (or related subject area) terminology to technical vocabularies in a particular domain is a critical problem. There is a need for new tools to help formulate and refine searches and navigate through the information space of concepts used to describe a collection. Different people use different words for the same concept or may employ slightly different concepts and this 'vocabulary problem' is a barrier to widening scholarly access.

The historic environment sector has a rich tradition of employing Knowledge Organisation Systems (KOS – such as thesauri). However, such vocabulary tools are often not fully integrated into searching and indexing systems and online practice has tended to mimic traditional print environments. The full potential of these knowledge resources in online environments has not been tapped. The paper will explain how the emerging W3C standard - Simple Knowledge Organisation System (SKOS) - can provide the necessary semantic cross-referencing for term look up across the thesauri held in the system, and enable browsing and semantic concept expansion within a chosen thesaurus. This will allow a search to be augmented by SKOS-based vocabulary and semantic resources (assuming the services are used in conjunction with a search system). Users may also simply browse a concept space to explore and become familiar with specialist terminology or as part of a broader application. In the next section, this paper will consider how the use of an ontology such as the CIDOC CRM, especially when further enhanced by domain specific extensions (CRM-EH), can provide semantic interoperability between previously isolated datasets built using different database platforms and designed with differing data structures.

2. CONCEPTUAL MODELLING USING CIDOC CRM

The CIDOC CRM standard (Crofts et al., 2008) does not require (nor particularly recommend) any particular methodology for using it. After consultation, the approach that was adopted by the CFA was derived from general ontology building methods (Denny, 2002). and can be summarized in five main stages (Cripps, 2004):

- Acquire domain knowledge
- Organize the ontological model
- Flesh out the ontological model
- Check the work
- Commit the ontological model

In the first instance this resulted in a model that related archaeological conceptual classes directly to the CIDOC CRM entities. However on further consideration it was found that the scope notes within the CIDOC CRM that were the definitions of the semantic meanings behind the concepts, only represented the archaeological concepts in the CRM-EH model at quite a high level of conceptualisation. It was for this reason that further specifically archaeological extensions, with more specific archaeological scope notes, were made (Cripps et al 2004).

3. ARCHAEOLOGICAL EXTENSIONS OF CIDOC CRM

In a recent paper D'Andrea succinctly summarizes one of the key problems with any attempts to integrate the digital information recorded by archaeologists, namely the nature of archaeologists themselves: "Undertaking a standardization process involving archaeologists and archaeological data may perhaps be considered as a symptom of naivety. Few scientific communities are more individualistic than this, the result being an extreme fragmentation of systems and data models." (D'Andrea 2008). Nevertheless, the work undertaken by EH staff to develop better forms of integration for their own archaeological project work in recent years, does in some way attest to an equally strong professional ethic of expecting, and feeling professionally obliged, to share the information outputs from their historic environment recording work, both with other archaeologists, but perhaps even more so with the wider

public. It was largely because of the recognition of the fundamental problems represented by the diversity of different archaeological recording systems developed by the many professional archaeological organisations in England over the last twenty five years (May, S., 2004), that English Heritage staff decide to look for an approach that would allow archaeologists to "map" their existing systems to some common and over-arching information framework, rather than trying to re-invent yet another "recording system to end all recording systems".
As a result of investigating the possibilities for a common information framework the CIDOC CRM was considered as a possible solution for bringing such a plethora of archaeological information datasets together, and building the semantic links by making the semantic relationships explicit – between the various islands in the archaeological information archipelago. However, although the CIDOC CRM was derived from the wider cultural heritage domain it was noted during the early stages of modelling the archaeological activities of EH staff, that not all the entities in the 'vanilla' CIDOC CRM were quite explicit enough to cover some of the more complex relationship. In consultation with the CIDOC CRM Special Interest Group it was decided to *extend* the entities that were represented in the EH ontological model (CRM-EH) and in order to distinguish them and their semantic meanings an additional numbering system was produced. Most importantly additional 'definitions' amounting to scope notes pertaining to the more specific archaeological entities represented by the CRM-EH entities have been written and these are now contained in the RDF 'Description' field as part of the RDF structure that forms part of the CRM-EH model. Thus a CRM-EH entity ContextFind (EHE0009) is also a CIDOC CRM entity Physical Object (E19). The CRM-EH extensions and scope notes are principally derived from EH archaeological experience, practice, and in particular the relevant information already documented in the EH archaeological recording manual, upon which the majority of any digital recording system is necessarily based. By effectively "mapping" the CRM-EH to the fields in the EH recording system the CRM-EH has to some degree been 'future-proofed' to make sure that it relates as closely as possible to any computer system that is implemented in the future to contain the data that the paper based recording manual is also meant to record.

This is why the CRM-EH is referred to as an *extension* of the CIDOC CRM – although the provision for such extensions is an integral part of the CIDOC CRM's implementation as stated on the opening page of the CIDOC CRM: "By its very structure and formalism, the CRM is extensible and users are encouraged to create extensions for the needs of more specialized communities and applications". (Crofts, N., et al 2008 p i). This paper will show how the use of the CRM-EH extensions has further aided the process (if not the simplicity) of mapping between different archaeological datasets, although it is noted that it is not an absolute requirement to make extensions to the CIDOC CRM in order to map data and indeed in many cases within the STAR implementation a choice has been made to simply map directly to 'vanilla' CIDOC CRM without any requirement to use a CRM-EH extension. At present it remains to be seen exactly what further issues of interoperability may emerge as more and more archaeological data sets are mapped, and more and more complex searches and queries are attempted. As D'Andrea notes: "Considering some of these mapping procedures, it may be noticed that there are alternative ways of representing the same conceptual archiving practice. While the English Heritage mapping chose to base on the creation of new sub-classes of "IsA" type specializing the original CIDOC-CRM and making it richer, the Italian ICCD mapping preferred to maintain a full compatibility with CIDOC-CRM, fitting the starting source with the destination ontology only through the semantic equivalence between corresponding classes".
It should be clarified here that the CRM-EH is fully compatible with the CIDOC CRM – indeed every CRM-EH entity also bears the related CIDOC CRM entity and "E" number. What the extensions have allowed is greater 'specialisation' of specific issues pertaining to how EH archaeologists (and probably many others) actually record the archaeology that they investigate. More details of the RDF extensions and associated scope notes can be found by downloading the RDF files from:
http://hypermedia.research.glam.ac.uk/media/files/documents/2008-04-01/CIDOC_v4.2_extensions_eh_.rdf

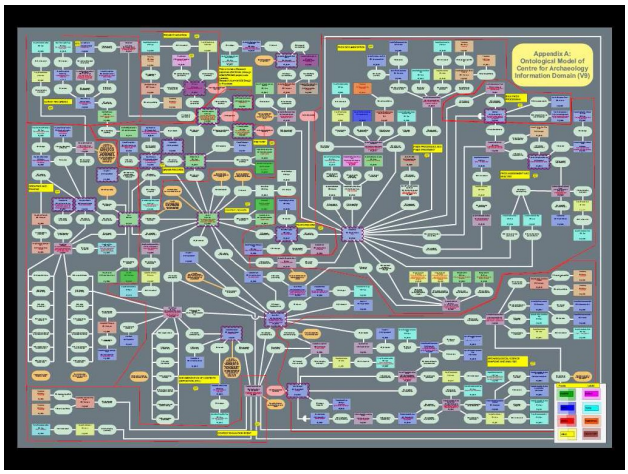

Figure 1: Ontological Model of the English Heritage Archaeological Information Domain
See also online version at
http://cidoc.ics.forth.gr/docs/AppendixA_DiagramV9.pdf

4.STAR INITIAL METHODOLOGY FOR MAPPING FROM THE CRM-EH MODEL TO DATA SETS

This section will set out some of the main issues addressed in establishing a methodology for mapping from the CIDOC CRM ontology and CRM-EH extensions to actual specific data fields in various selected test-bed datasets.

4.1 Dataset selection

To test the possible implementation, and prove the interoperability of the modelling, a number of data sets were needed. The test-bed data were selected by various criteria: data from EH legacy systems such as Delilah (itself over 25 years old); current or more recent databases being used by teams in the EH archaeological centre at Fort Cumberland, including some specialist archaeological science data from the Environmental Archaeology Branch at EH (mostly a variety of different project databases using various versions of MS Access or specialist datasets in MS Excel); external data in a database designed by a different archaeological organisation not using EH software or systems; and some data that reflected

recent developments in online publication of integrated data and reports. The LEAP Silchester data, structured using IADB (http://www.iadb.org.uk/), and published in Internet Archaeology as an example of "Integrated Publication" (Miles, 2004, Richards 2004) satisfied both the latter criteria. In addition some attempt was made to choose datasets that broadly covered a range of archaeological periods, but ones that would allow some meaningful archaeological cross-project searches. The initial datasets chosen were Raunds Roman Archaeological Database (RRAD) along with Raunds Prehistoric data, Raunds environmental sampling data and the Silchester LEAP data.
As well as being from quite differing database origins these data sets were also from different stages in the historic environment 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 approaching 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, A et al. 2007). These characteristics of the initially selected four datasets are summarized in Table 1 below.

	Database Type	**Main Archaeology Periods**	**MoRPHE Project stages**
Raunds Prehistoric	Delilah - CSV	Neolithic & Bronze Age	Execution - Excavation
Raunds Environmental	MS Excel - DBF	Neolithic to late Roman	Execution - Assessment
Raunds Roman (RRAD)	MS Access - MDB	Roman & Iron Age	Execution – Assessment to Analysis
Silchester LEAP data	MySQL - MYD	Roman & Late Iron Age	Execution - Publication

Table 1: Summary of initial test-bed data sets for STAR prototype CRM browser

4.2 Data mapping and generation of RDF triples

In order to map the datasets to the CRM-EH (and thereby to the CIDOC CRM) the approach taken was to identify a "core" of key archaeological concepts from the larger CRM-EH data model and then relate these "core" entities to the key fields identified in the RRAD, RPRE and IADB databases. From this starting point further data fields could be mapped to the CRM(s), as and when they needed to be included in the resulting merged test-bed RDF dataset. This intellectual mapping required 'domain' archaeological knowledge of the data and the CIDOC CRM and CRM-EH ontologies. Initial mappings were performed by May and communicated via spreadsheets to the team in Glamorgan. The process of 'domain mapping' is time consuming, and requires considerable focus on the complex semantic and conceptual issues being addressed. It is not therefore something that can be easily slotted in to a few minutes here and there within a general work schedule. Fortunately it seems that once an initial mapping is produced for the archaeological domain, the process of mapping a further system dataset is considerably aided by being able to recognize similarities (or exact matches) in work-patterns or conceptual activities and then simply using the same relevant parts of the CRMEH model. From the initial "core" mappings, it did generally become easier for some subsequent mappings to be performed by others in the project team (i.e. non archaeological domain experts) using the initial spread sheet mappings as a guide, with domain expert validation by May as and when required.

In addition some of the exchange of the mappings between entities and core data fields and, particularly, keeping the evolving CRM-EH modelling up to date, was aided by the use of Protégé ontology editing software (http://protege.stanford.edu/) or Altova SemanticWorks (http://www.altova.com/products/semanticworks/semantic_web_rdf_owl_editor.html) although in many cases the complexity of the modelling diagram was beyond the graphical visualization capabilities of these primarily text based ontology editing programmes. One more recent development of the modelling is an attempt to add more dimensions and "granularity" to the modelling diagram with the use of multiple colours and colour shading, but this also has drawbacks for displaying greater complexity on any computer screen, not least it's inaccessibility for colour-blind readers. The further process of actually extracting the data from the datasets in accordance with the "core" mappings, was also time consuming and would not be human scaleable over a large number of datasets. Therefore an automated data mapping and extraction utility - using SQL queries with query parameters saved in XML format for subsequent reuse - was developed by Binding to assist the processing of the end data, and the resulting output is an RDF format file. The automated mapping 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 extracted data was imported into a MySQL RDF triple store database on the Glamorgan server, using the SemWeb RDF library.

4.2 ID format adopted

The RDF entities in the RDF triple store, require unique identifiers. Some of the data being extracted was an amalgamation of records from separate tables – e.g. *EHE0009.ContextFind* actually contained records from RRAD.Object & RRAD.Ceramics tables. It was therefore necessary to devise a unique ID for all RDF entities beyond just using the record ID from an individual table. The format adopted to deal with all these issues was a simple dot delimited notation as follows:

[URI prefix]entity.database.table.column.ID
e.g. "EHE0008.rrad.context.contextno.100999"

This format (although verbose) allowed the use of existing

DB record ID values without introducing ambiguities. In RRAD database, Ceramics and Objects were both instances of *EHE0009.ContextFind*. This therefore involved the combination of data from two tables:

• *EHE0009.rrad.object.objectno.105432 [an EHE0009.ContextFind record from theRRAD object table]*
• *EHE0009.rrad.ceramics.ceramicsno.105432 [an EHE0009.ContextFind record from the RRAD Ceramics table, with the same ID value]*

The format also allowed the same base record ID to be used for both *EHE0009.ContextFind* and E*HE1004.ContextFindDepositionEvent* (these records actually originated from the same table and had a 1:1 relationship), using a different entity prefix to disambiguate the records:

• *EHE0009.rrad.object.objectno.105432 [The ContextFind record ID]*
• *EHE1004.rrad.object.objectno.105432 [The ContextFindDepositionEvent recordID]*

Finally an arbitrary URI prefix (http://tempuri/) was added to all ID values. According to need, this can be replaced with a more persistent prefix.

4.3 Date/Time and coordinate formats adopted

Although there is nothing dictated in the CIDOC CRM ISO or CRM-EH about date/time representation formats, it was important to maintain a consistent date format throughout the merged data. For the purposes of the data extraction to keep all data consistent we used a "big endian" format (i.e. from most to least significant) compatible with both W3C standards and ISO8601 ("Data elements and interchange formats – Information interchange – Representation of dates and times"). The format is as follows:
CCYY-MM-DDThh:mm:ss e.g. "2007-05-03T16:19:23"
This format does not introduce any restrictions on how dates & times are eventually displayed or used within applications; it merely provides a common string representation mechanism for interoperability of data.

Spatial co-ordinates with various formats are used in a number of different fields in each of the tes-bed datasets. RRAD coordinates were 6 digit numeric values in separate "Easting" and "Northing" columns. RPRE coordinates were slash separated string values, sometimes with an extra 4 digit value appended (i.e. either *nnnnnn/nnnnnn/nnnn* or *nnnnnn/nnnnnn*). IADB coordinates were numeric values in separate "Easting" and "Northing" columns (and appeared to be relative to a site local reference datum). CRM/CRM-EH requires a single string to represent a spatial co-ordinate value. The consistent format chosen for output was 6 digit space delimited Easting and Northing values, with an optional Height value (Above Ordnance Datum). These values were all assumed to be in metres:
nnnnnnE nnnnnnN [nn.nnnAOD] e.g. "105858E 237435N 125.282AOD"

4.4 Modelling notes/annotations

The EH recording manuals and the current datasets contain several kinds of note fields. For the purposes of disambiguating
all the different types of notes that show up in the RDF triples, a core set of EH archaeological note types have been identified. These are:
• Comments (the most general category of 'catch-all' notes)
• Method of excavation
• Interpretation (likely to be further refined to specific cases)
• Siting description (reasons relating to location of a sample)
• Site treatment (relating to samples)
While it might potentially be restrictive to model notes as strings (notes have other implicit attributes such as language, author/source etc.), this is the current position within the CRM (*E1.CRM Entity _ P3.has_note _ E62.String)*. However, taking the RDFS encoding of CIDOC CRM recommendation, we intend to create sub properties of *P3.has_note* e.g. *EHPxx1.has_InitialInterpretation*, *EHPxx2.has_RevisedInterpretation*, as part of future work.
The CIDOC CRM has a modelling construct in the form of "properties of properties". For example, property *P3.has_note* has a further property *P3.1.has_type* – intended to model the distinction between different types of note. Unfortunately, this construct does not translate well to RDF. As evidence of this, property *P3.1.has_type* is not actually part of the current RDFS encoding of CRM on the CIDOC website (in the comment header there is a suggestion to create specific sub properties of *P3.has_note* instead). The more recent OWL encoding of CRM also avoids including the construct.

4.5 Modelling of Events

Many of the events defined in the CRM-EH modelling and used to interconnect objects and places in the CRM-EH model were largely only *implicit* within the earlier relational database structures. The ability to model these events more *explicitly* for further data recording and analysis was one of the key reasons the CIDOC CRM was chosen for the event based ontological modelling, allowing a new approach to systems modelling and design. As a result therefore, in certain cases during the mapping from the data to the CRM model (and its further translation from relational data structures to an RDF graph structure) it was necessary to create this *explicit event* information by the formation of intermediate '*virtual entities*', but with no current data to actually fill such entities. The aim is that for any newly implemented digital recording system used by EH, these explicit events will be recorded as part of those new recording systems. 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.

4.6 Initial mapping of data fields to extended CRM

When imported into the SemWeb MySQL triple store database the combined data files produced the following results:

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 2. Statistics for extracted data

The number of statements (triples) contained in the resultant RDF files is **2,988,005.** Some triples (e.g. *rdf:type* statements) were duplicated due to entities occurring within multiple files, but any duplication was removed during the aggregation process. A number of separate RDF files were combined in the aggregation process including the CRM itself, the CRM-EH extension, alternative language labels for the CRM, and various EH domain thesauri. This scale of triple store has proved to provide perfectly adequate browsing capabilities for the project team even when using a relatively remote 3G wireless connection to the data on the Glamorgan server from a fieldwork site and during a mainline train journey. The data files produced were each validated against the W3C RDF validation service. Whilst this did not prove the validity of the data relationships or even conformance to CRM-EH, it did at least give confidence in the validity of the basic RDF syntax.

4.7 Prototype Search / Browse Application

An initial prototype client application has been produced (see Fig 2), capable of cross searching and exploring the amalgamated data in the RDF triples store. The application utilizes 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, and also working upwards from data items to concepts as desired.

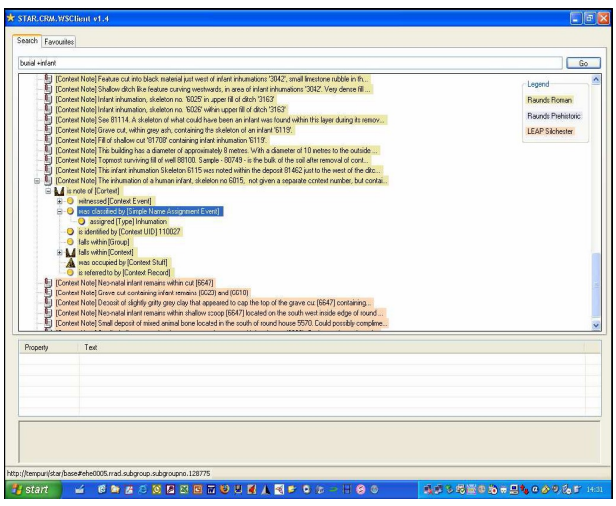

Figure 2: Prototype CRM Browser Web Service API

The results to date have successfully demonstrated the potential for searching across at least four very separate datasets with quite different origins and a variety of different data structures and content. At present the types of searching that can be carried out are inevitably limited by the constraints of the actual datasets selected for test purposes – although even with these constraints some very interesting archaeological queries have already been formulated that could not easily be explored without the use of the STAR browser. Further testing is planned to introduce additional datasets for query generation and expansion testing purposes, but also to test the practicalities of how much data the triple store can function with. At present How this CRM semantic browser could be developed in order to best enable archaeologists to search the various historic environment datasets will be discussed further in section 6.

5. THE USE OF SKOS BY STAR

The STAR project has adopted the emerging W3C standard SKOS Core as the representation format for historic environment domain thesauri, such as the English Heritage National Monuments Thesauri (http://thesaurus.english-heritage.org.uk/), and related Knowledge Organization Systems (KOS). "In general, thesauri conforming to the SI/NISO/ISO standards should map in a fairly straight forward manner to SKOS. However, there may need to be judgments on how to deal with non-standard features (Tudhope 2008). The STAR investigations to date have encountered some potential problems associated with the use of Guide Terms or facet indicators in some thesauri. Other issues may arise from the need to create URIs for concept identifiers as part of the conversion and the potential for validation.

6. ATTEMPTS TO GATHER USER REQUIREMENTS FOR A SEMANTIC BROWSER.

With the CRM Browser operational the next task has been to try and gather ideas for how archaeologists (as the CRM-EH domain experts) and other more general users may wish to use the capabilities to search across a range of different data sets. One immediate problem in judging this is that users have very little, if any, experience of what "semantic browsing" might look or 'feel' like. In the first instance a number of workshops are being held to gather 'user requirements' in a fairly orthodox manner. These have already led to a discussion of possible interfaces for searching in systematic ways between finds and contexts and navigating around groups of contexts to understand better all the different types of data that are associated semantically with a single unit of archaeological Stratigraphy and then working up these queries across associated archaeological groups to search across an associated group of archaeological features (in different datasets) and further still to answer the sort of complexities posed in the 'scenario' "Excavated examples of 'Roman' corn drying ovens where archaeobotanical analysis has taken place and where rye (seacale cereale L) is found in deposits from the 'Iron Age' in Europe" or "What is the first record for Centaurea Nigra (Black Knapweed) in England"? Another area of interest for archaeologists will be the potential for geospatial browsing using GIS, (something not particularly within the scope of the current STAR project although some exploration in this area will be attempted).

The current CRM Browser is just a prototype of further tools that the technologies will enable to be developed. At the present time, more use of the cross-terminology searching enabled by SKOS will also need to be developed and one use of the SKOS web service for archaeologists is likely to be a web service that allows archaeologists to better specify what particular archaeological terminology list they are referring to when they deposit their digital archives or write their archaeological reports. One further style of user interface that has been postulated, is the development of more interactive methods for users to select from a suite of 'tools' for semantic searching, perhaps allowing a spatial, temporal and text based series of search mechanisms to be available in the "semantic query toolkit".

6.1 Future directions for linking reports to data

One major area that STAR is still investigating is the ability to use the CIDOC CRM and CRM-EH ontologies to aid the semantic searching of reports linked to data and visa versa. In particular there seems some potential to use Natural Language Processing (NLP) techniques such as those employed in the GATE architecture (http://gate.ac.uk/) to annotate reports with commonly identified archaeological concepts. This could thereby allow both authors and researchers to develop new approaches to formulating conceptual 'threads' through the information written about an archaeological investigation, and enable researchers and other users to 'drill down' into underlying data once they have identified a good 'semantic match' using a suitable semantic query interface, be that a basic text-based query ("more like *this* please?" whatever ***this*** represents); or GIS based (spatial - show me *where* these things come from?); or some sort of time-line (temporal – *when* did these sorts of things first start occurring?); a photo or even part of a photo (image based "give me more that look like this bit of the image"?)… or in the future most probably some combination of all these.

7.CONCLUSIONS

The results to date from the combined work of the ontological modelling and the outputs from STAR have demonstrated that a degree of semantic interoperability has already been attained. While this is encouraging and is a positive driver for further research, it remains to be adequately assessed whether a full implementation of the CRM-EH modelling is achievable in a cost-effective manner. This is in part because a full implementation of the CRM-EH model is only going to be really achievable once a new system for digital recording has been introduced at EH, in the next year, that will enable staff to begin to collect data for all the new entities in the model (and include data in those fields that accordingly map to the 'virtual entities') and it will then take some further time for enough examples of projects with this new type of related data to be recorded and analysed, to test the effectiveness of the new interoperability and how well it integrates the new data with the legacy data. Nevertheless the degree to which at least partial interoperability between a range of different datasets has been achieved, across at least three different types of database software and archaeological recording system structures, shows that the methodology is sound and only by the further testing of more and more archaeological datasets – mapped to the CRM-EH, and thereby to the CIDOC CRM – will reveal increasing tests and measures of the scalability of the current system and hopefully will enable the most appropriate ways to implement scaleable solutions for both hardware and software infrastructures. Work on the early stages of using NLP software for annotating key relationships within reports has also looked promising and if the same level of 'mapping' across different organisational report structures can begin to be attempted then it may be possible to incorporate many of the 'core' elements of the CRM-EH ontology into some form of developmental report and data search and browser mechanism.

7.1 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 wider work of the STAR project team and domain experts at English Heritage.

7.2References

Bergman, M., 2001. WHITE PAPER: The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing*, vol. **7**, no. 1, August, 2001 (accessed 16 June 2008)

Clarke, Amanda., Fulford, Mike G, Rains, Mike and Tootell, Klare., 2007. Silchester Roman Town Insula IX: Development of a roman property C. AD 40-50 - C. AD 250. *Internet Archaeology* 21. *http://intarch.ac.uk/journal/issue21/silchester_index.html (accessed 28 May. 2008)*

Cripps P., Greenhalgh A., Fellows D., May K., Robinson D., 2004. Ontological Modelling of the work of the Centre for Archaeology, CIDOC CRM Technical Paper. http://cidoc.ics.forth.gr/technical_papers.html

Cripps P., May K., (forthcoming). To OO or not to OO? Revelations from ontological modelling of an archaeological information system. *Proceedings of Computer Applications in Archaeology* Prato 2004.

Crofts, N. Doerr, M. Gill, T. Stiff, M. and Stead, S. (Eds.) Nov 2008. *Definition of the CIDOC Conceptual Reference Model and Crossreference manual. Version 4.2.4.* Official release of the CIDOC CRM. (Accessed 9 June 2008)| http://cidoc.ics.forth.gr/official_release_cidoc.html

D'Andrea. A., Niccolucci, N., 2008. "Mapping, Embedding and Extending: Pathways to Semantic Interoperability. The Case of Numismatic Collections". First Workshop on Semantic Interoperability in the European Digital Library **(**SIEDL 2008).

Denny, M., 2002. *Ontology Building: A Survey of Editing Tools.* http://www.xml.com/pub/a/2002/11/06/ontologies.html

Doerr, M., 2003. *The CIDOC CRM – an Ontological Approach to Semantic Interoperability of Metadata.* AI Magazine, Volume 24, Number 3 (2003).

English Heritage., 2007 *Management of Research Projects in the Historic Environment (MoRPHE).* http://www.englishheritage.org.uk/MoRPHE (16 June 2008)

May, K., 2006 “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

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

Miles, D., 2004 Digital dissemination and archiving. *Internet Archaeology* **15**. http://intarch.ac.uk/journal/issue15/preface (accessed 16 June 2008)

Richards, J., 2004 Online Archives. *Internet Archaeology* **15**. http://intarch.ac.uk/journal/issue15/richards_index.html (accessed 16 June 2008)

SKOS: Simple Knowledge Organization Systems, http://www.w3.org/2004/02/skos

STAR Project: Semantic Technologies for Archaeological Resources, http://hypermedia.research.glam.ac.uk/kos/star

Tudhope, D., Binding, C., May, K., 2008 Semantic interoperability issues from a case study in archaeology. First Workshop on Semantic Interoperability in the European Digital Library **(**SIEDL 2008).