Experiences with Knowledge Organization System Services from the STAR Project

Douglas Tudhope & Ceri Binding

Cultural heritage organizations are looking to open digital collections and databases, previously confined to specialists, to a wider audience. This paper reports on initial experiences from the project, which started January 2007, in particular the conversion of vocabularies to a representation suitable for digital semantic techniques, prototype terminology services based on these representations and the inter-relation of different kinds of Knowledge Organization Systems (KOS).

There is a need for 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.

STAR (Semantic Technologies for Archeological Resources) is a 3 year project, in collaboration with English Heritage (EH), funded by the Arts and Humanities Research Council (AHRC). Its aim is to investigate the potential of semantic terminology tools for widening and improving access to digital archaeology resources, including disparate data sets and associated grey literature. This involves developing new methods for linking digital archive databases, vocabularies and the associated grey literature, exploiting the potential of a high level, core ontology and natural language processing techniques. It builds upon earlier work by the authors on semantic concept-based expansion techniques for faceted queries (Tudhope *et al.* 2006b).

Types of Knowledge Organization Systems

There are various kinds of KOS, serving different purposes (Tudhope et al 2006a).

Information retrieval KOS (such as classifications and thesauri) is intended primarily to assist retrieval of resources, originally from bibliographic databases and library catalogues and now from Digital Libraries and the Web. The design rationale is perceived assistance in future retrieval operations. These include classification and indexing, search (including browsing, query and various forms of "intelligent" searching), mapping between KOS (mono and multi- lingual), providing a framework for learning a subject domain or exploring it in order to refine a (re)search question (defining concepts and setting them in context).

KOS can be used to support manual cataloguing and also automatic cataloguing activities. KOS range from domain specific KOS to general classification systems, from two hierarchical levels to systems with great depth and breadth of coverage.

The distinction between classification and indexing is important but often misunderstood, especially in new application areas (Tudhope and Binding 2008). Classification seeks to group similar items together, whereas indexing seeks to bring out the differences between items, in order to help distinguish them during search. A KOS might be used both for classification/indexing and searching, or just searching.

Although the term is sometimes employed to refer to any form of KOS, ontologies are formal representations, derived from work in Artificial Intelligence (AI), modeling a knowledge domain with precise definitions and relationships. They are designed to be used by first order logic reasoning systems and are a knowledge representation mechanism for communication between (automatic) intelligent agents.

They are suited to applications with well-defined objects and operations. Basically they suit to situations where it is possible to reach agreement as to the precise definition of concepts (and terms) and where it is useful to define logical rules for processing relationships and possibly inferring new knowledge.

These applications tend to have a different focus than retrieval per se, for example automatic generation of new data. Examples might include many scientific applications, where the ontology is a model of currently accepted scientific knowledge and smaller subject domains, such as some business applications.

There is overhead in creating (and sustaining) formal representations and in some situations it may not be not feasible to come to commonly agreed, precise definitions on abstract or contested concepts (e.g. some descriptions of human activity). For example, in search applications, where a fuzzy notion of 'aboutness' is the basis for indexing or classifying a document, as opposed to an assertion of fact, a less formal approach may be suited.

Background to STAR project

While Web search engines have made advances in recent years, the problems of keyword searching are well known. Link popularity algorithms can yield good results for specific documents (or persons) at major sites but they are not suitable for the conceptual or subject-based searches common in academic research or serious public inquiry. Significant differences in results stem from trivial variations in search statements and from related but differing conceptualisations of a research inquiry.

The current situation within English Heritage and the archaeology domain generally is one of fragmented datasets and applications, with different terminology systems. The interpretation of a find (or free text report of an excavation) may not employ the same terms as the underlying dataset. Similarly searchers from different scientific perspectives may not use the same terminology.

The cultural heritage sector often employs KOS, such as thesauri, for indexing. However, such vocabulary tools are often not fully integrated into search tools and online practice has tended to mimic traditional print environments. The full potential of these knowledge resources in online environments has not been tapped.

As discussed above, ontologies typically provide a higher level domain conceptualisation with more formal definition of roles and semantic relationships. Within archaeology, the CIDOC Conceptual Reference Model (CRM) is emerging as a standard core ontology (Doerr 2003).

The CRM is the result of 10 years effort by the CIDOC Documentation Standards Working Group and is an ISO Standard (ISO 21127:2006). It encompasses cultural heritage generally and the intention is that it can mediate between different sources and types of information. In order to supply an umbrella framework to integrate different datasets and thesauri, EH have designed a core ontology based on the CIDOC CRM standard (the CRM-EH), extending the CRM with key archaeological concepts and relationships.

Star Project initial work

The CIDOC CRM deals with concepts at a high level of generality. For mapping to datasets at a detailed level, we worked with the CRM-EH extension of the CRM, developed by our collaborators (May) in English Heritage (Cripps et al. 2004, May 2006).

The CRM-EH models the archaeological excavation and analysis workflow. Thus it introduces concepts such as find and context, specialising the original CRM concepts for object and place. Working with May, an implementation of the CRM-EH has been produced as a modular RDF extension referencing the published (v4.2) RDFS implementation of the CRM.

Initial mappings were made from the CRM-EH to three different database formats, where the data has been extracted to RDF and the mapping expressed as an RDF relationship. The data extraction process involved selected data from three archaeological datasets, based on three different database formats.

Selections from the different databases were extracted via SQL queries, and stored as separate RDF files, simplifying the process. These selections can be recombined as required. The extracted data corresponded to a subset of the CRM-EH model. For the initial phase we limited the scope of the data extraction work to data concerning contexts and their associated finds. A mapping and data extraction tool, developed for the project, facilitated the (significant) manual work involved (Binding et al. 2008).

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. The SemWeb library was employed to aggregate the extracted data files into a single SQLITE database. The resultant database of aggregated data was 193MB overall and consisted of 268,947 RDF entities, 168,886 RDF literals and 796,227 RDF statements (triples). The SemWeb library supports SPARQL querying against the database, but the SQLITE database itself also supports direct SQL queries.

Conversion of KOS

STAR employs SKOS Core as the representation format for domain thesauri and related KOS. SKOS Core is a W3C Working Draft RDF/XML representation for KOS, based on a formal data model. SKOS is intended as a formal RDF/XML representation standard for the family of KOS, with a lightweight semantics designed for information retrieval purposes. This offers a cost effective approach for dealing with thesauri.

Thesaurus data was received from English Heritage National Monuments Record Centre, as CSV format files. Initially we converted the CSV files to XML, and wrote an XSL transformation to export the data to SKOS RDF format. This worked for the smaller thesauri.

However XSL transformation of the data files was a resource intensive operation for the larger thesauri, with the PC running out of memory on occasion. Thus we moved to another approach, which imported the CSV files into a Microsoft Access database, with a custom C# application then exporting the data into SKOS RDF format (Tudhope et al. 2008).

Separate RDF files were produced for each thesaurus and validated using the W3C RDF validation service. All files passed this basic RDF syntax validation test without problems. The files were then checked using the W3C SKOS validation service, which is a series of SKOS compatibility and thesaurus integrity checks. A few minor anomalies were detected by these tests, including legacy features such as orphan concepts.

This information was passed back to the developer of the thesauri to feed into routine maintenance. Any future updates to thesaurus data can be processed easily in a similar fashion. The SKOS files will be used in the STAR project by query expansion and domain navigation tools (see for example, figure 4 below).

Mapping between SKOS and other representations

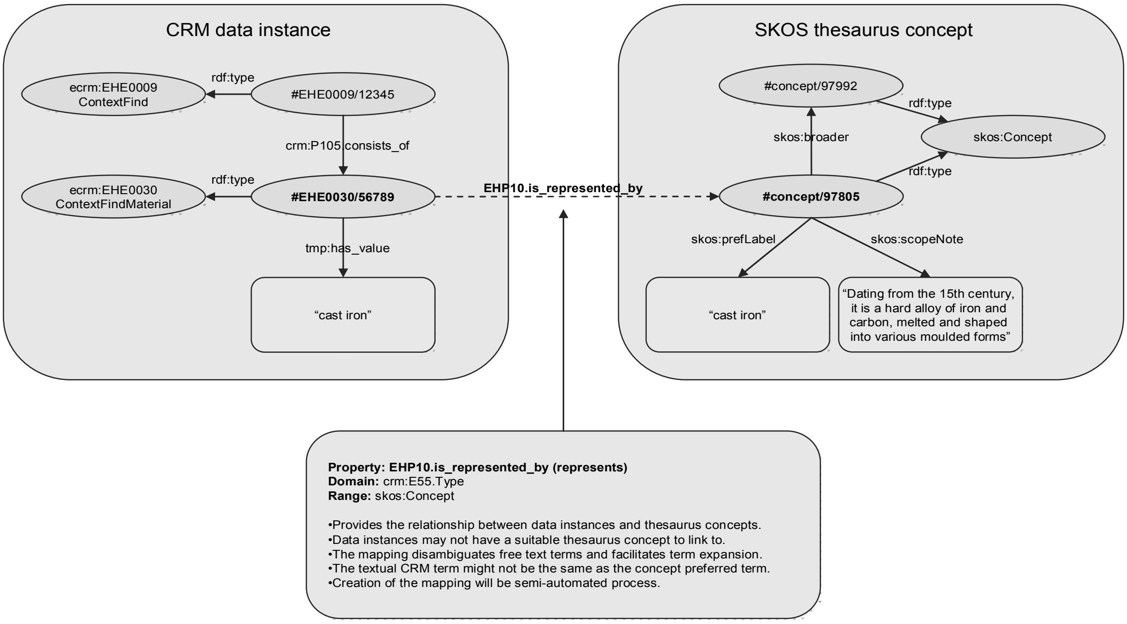
The next phase of the research will investigate the appropriate connections between the thesauri (expressed in SKOS) both to information (data) items and to an upper (core) ontology, in this case the CRM-EH. Illustrating both issues, Figure 1 shows the model we have adopted for integrating SKOS thesauri with the CRM. This illustrates two points with SKOS RDF data: (a) the connection between a SKOS concept and the data item it represents and (b) the connection between the CRM and SKOS.

(a) Connecting SKOS concepts and data

STAR employs a project specific is represented by relationship to model the connection between a SKOS concept and an information item (Figure 1). Using a project relationship for this allows the possibility of modifying it to take account of subsequent standards development regarding these issues.

The standard *DC: Subject of* would be a possibility if appropriate. However, this does not quite capture the application of a SKOS concept to information items for STAR purposes. It might be argued that this concept-referent relationship should be modeled in SKOS.

However, we believe it would be more appropriately expressed in a separate indexing or vocabulary use standard. There tend to be differences in the usual use cases informing the application of Library Science KOS (intended for information retrieval purposes) and most (AI) formal ontology applications, which often model a mini-world, with a form of *Instance* relationship.



Property: EHP10.is_represented_by (represents)
Domain: crm:E55.Type
Range: skos:Concept

- Provides the relationship between data instances and thesaurus concepts.
- Data instances may not have a suitable thesaurus concept to link to.
- The mapping disambiguates free text terms and facilitates term expansion.
- The textual CRM term might not be the same as the concept preferred term.
- Creation of the mapping will be semi-automated process.

Figure 1: Model for combining SKOS and CIDOC CRM

(b) Connecting SKOS concepts and an upper ontology

The method of modeling the link between an upper ontology and domain thesauri (or related KOS) depends upon the intended purposes. A successful connection depends on a good alignment of the ontology and domain KOS, the number of different KOS intended to be modeled and the use cases to be supported. This is similar to the considerations and likely success factors for mapping between thesauri or KOS generally.

When the purpose is to support automatic inference, formalising the domain KOS and completely integrating them within a formal (OWL) ontology may be a good solution. Automatic inferencing can then be applied to the more specific concepts of the domain KOS. On the other hand, this is likely to be a resource intensive exercise.

Since information retrieval KOS and AI ontologies tend to be designed for different purposes, this conversion may change the underlying structure of the KOS. It may involve facet analysis to distinguish orthogonal facets in the domain KOS, which should be separated to form distinct hierarchical facets.

It may also involve modeling to much more specific granularity of concepts if the upper ontology is intended to encompass many distinct domain KOS. For example, the need for disambiguation may well not be present in the KOS considered separately but greater detail is required when several thesauri are integrated together.

Such highly specific modeling should be considered in terms of costs and benefits. It is important to consider the use cases driving full formalisation, since information retrieval KOS, by design, tend to express a level of generality appropriate for search and indexing purposes and driving down to greater specificity may yield little cost benefit for retrieval or annotation use cases.

We would argue that the SKOS representation offers a cost effective approach for many retrieval oriented applications that don't use first order logic in indexing, search and browsing. The W3C Semantic Web Deployment Working Group is developing a set of guidelines for combining SKOS and OWL generally.

A variant of the above approach, which allows the more tractable option of SKOS representation, is to consider the domain KOS as leaf nodes of an upper ontology, expressing this, with some form of subclass or type relationship. This corresponds to Leaf Node Linking in Zeng & Chan's (2004) review of mapping. One approach recommended for the CIDIC CRM is to assert an Instance relationship between a Type property of a CRM class and the top of a thesaurus hierarchy.

In some cases, the domain thesauri may not fit neatly under the upper ontology, the thesauri being designed separately for different purposes. In the STAR project, from the initial discussions with EH collaborators, the appropriate connection may sometimes be a looser SKOS mapping (broader) relationship between groups of concepts rather than complete hierarchies.

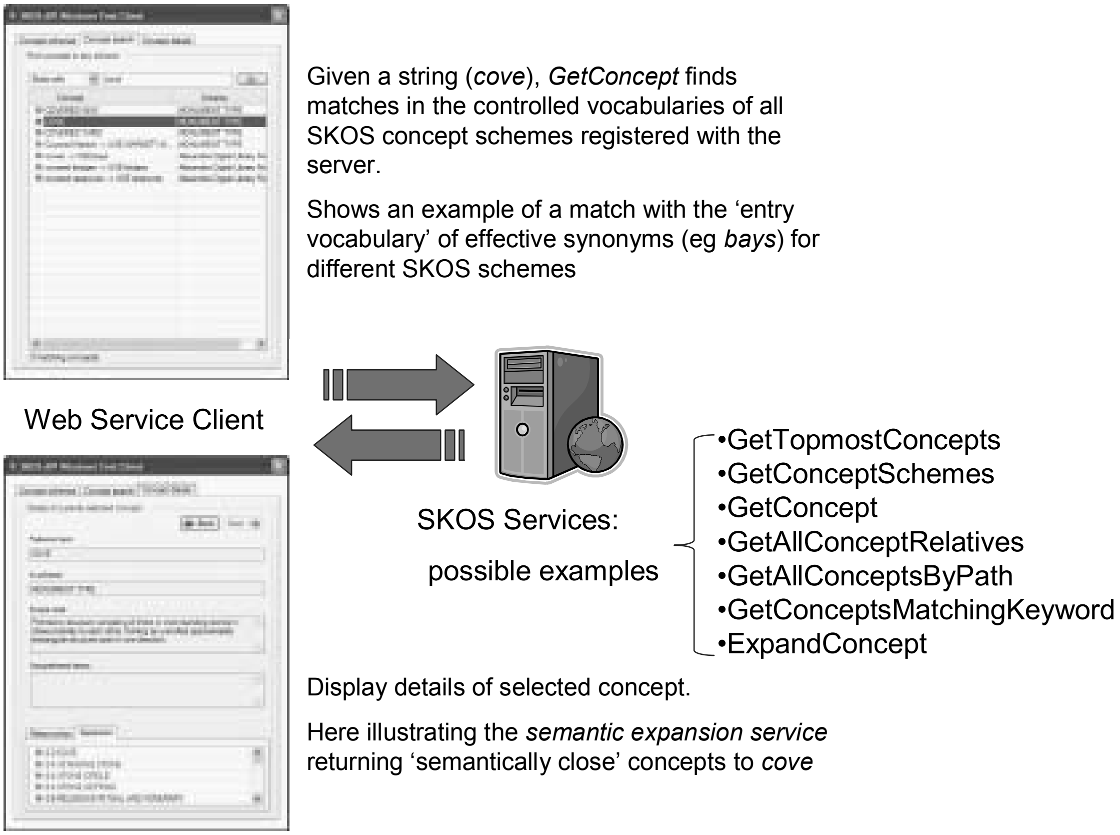
Yet another possibility, which avoids the issue, is illustrated in Figure 1, which shows a data instance mapped to a CRM entity and where the data items are also indexed with thesaurus concepts. In this case, there is a mapping between data and the integrating upper ontology and a separate mapping between database fields and domain thesauri.

SKOS based Terminology Services

Terminology web services based upon SKOS thesaurus representations are illustrated in Figure 2. Details of the API and a pilot demonstrator are available from the STAR project website. An earlier version of the current service was integrated with the DelosDLMS prototype next-generation Digital Library management system (Binding et al. 2007).

The service is based on a subset of the SWAD Europe SKOS API, with extensions for concept expansion. The services currently provide term look up across the thesauri held in the system, along with browsing and semantic concept expansion within a chosen thesaurus. This allows search to be augmented by SKOS-based vocabulary and semantic resources (assuming the services are used in conjunction with a search system).

Queries may be expanded by synonyms or by semantically related concepts. For example, a query is often expressed at a different level of generalisation from document content or metadata, or a query may employ semantically related concepts. Such expansion is based on a measure of 'semantic closeness'. Semantic expansion of concepts for purposes of query expansion yields a ranked list of semantically close concepts (Binding and Tudhope 2004; Tudhope et al. 2006).



Given a string (*cove*), *GetConcept* finds matches in the controlled vocabularies of all SKOS concept schemes registered with the server.

Shows an example of a match with the 'entry vocabulary' of effective synonyms (eg *bays*) for different SKOS schemes

Web Service Client

SKOS Services:

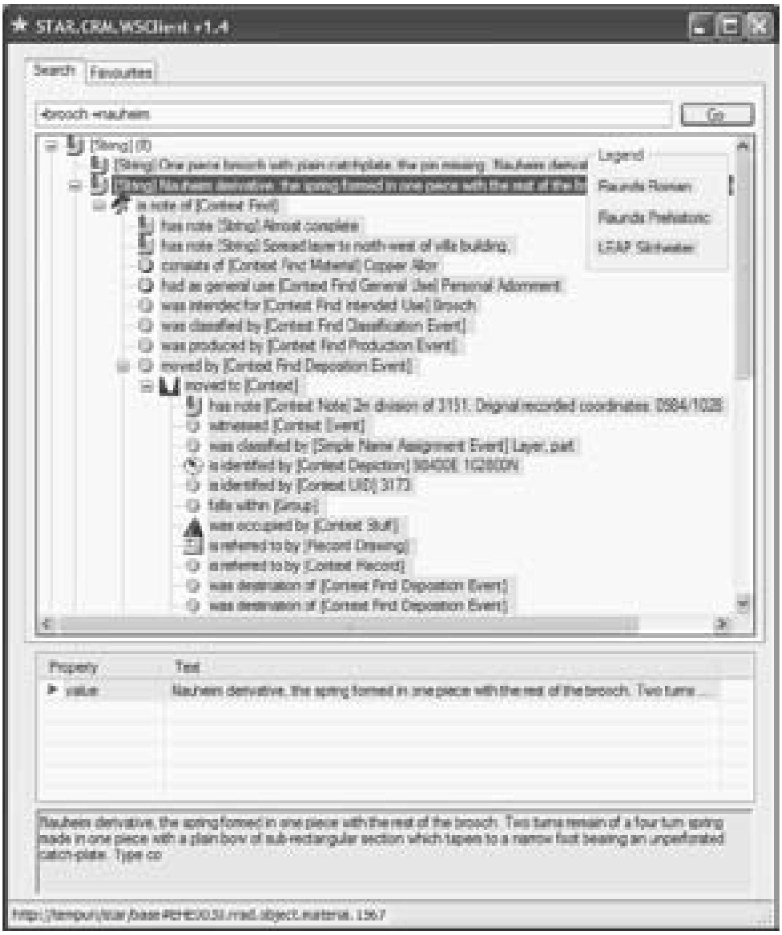
possible examples

- •GetTopmostConcepts
- •GetConceptSchemes
- •GetConcept
- •GetAllConceptRelatives
- •GetAllConceptsByPath
- •GetConceptsMatchingKeyword
- •ExpandConcept

Display details of selected concept.

Here illustrating the *semantic expansion service* returning 'semantically close' concepts to *cove*

Figure 2: SKOS web services



+brooch +nauheim

```
[String] (8)
  [String] One piece brooch with plain catchplate, the pin missing. Nauheim derivat
  [String] Nauheim derivative, the spring formed in one piece with the rest of the b
    is note of [Context Find]
      has note [String] Almost complete
      has note [String] Spread layer to north west of villa building.
      consists of [Context Find Material] Copper Alloy
      had as general use [Context Find General Use] Personal Adornment
      was intended for [Context Find Intended Use] Brooch
      was classified by [Context Find Classification Event]
      was produced by [Context Find Production Event]
      moved by [Context Find Deposition Event]
        moved to [Context]
          has note [Context Note] 2nd division of 3151. Original recorded coordinates: 0984/1028
          witnessed [Context Event]
          was classified by [Simple Name Assignment Event] Layer, part
          is identified by [Context Depiction] 904008 1028000N
          is identified by [Context UID] 3173
          falls within [Group]
          was occupied by [Context Stuff]
          is referred to by [Record Drawing]
          is referred to by [Context Record]
          was destination of [Context Find Deposition Event]
          was destination of [Context Find Deposition Event]
```

Property	Text
▶ value	Nauheim derivative, the spring formed in one piece with the rest of the brooch. Two turns ...

Nauheim derivative, the spring formed in one piece with the rest of the brooch. Two turns remain of a four turn spring made in one piece with a plain bow of sub-rectangular section which tapers to a narrow foot bearing an unperforated catch-plate. Type co

http://tempuri/star/base#EHE0031.rrad.object.material.1367

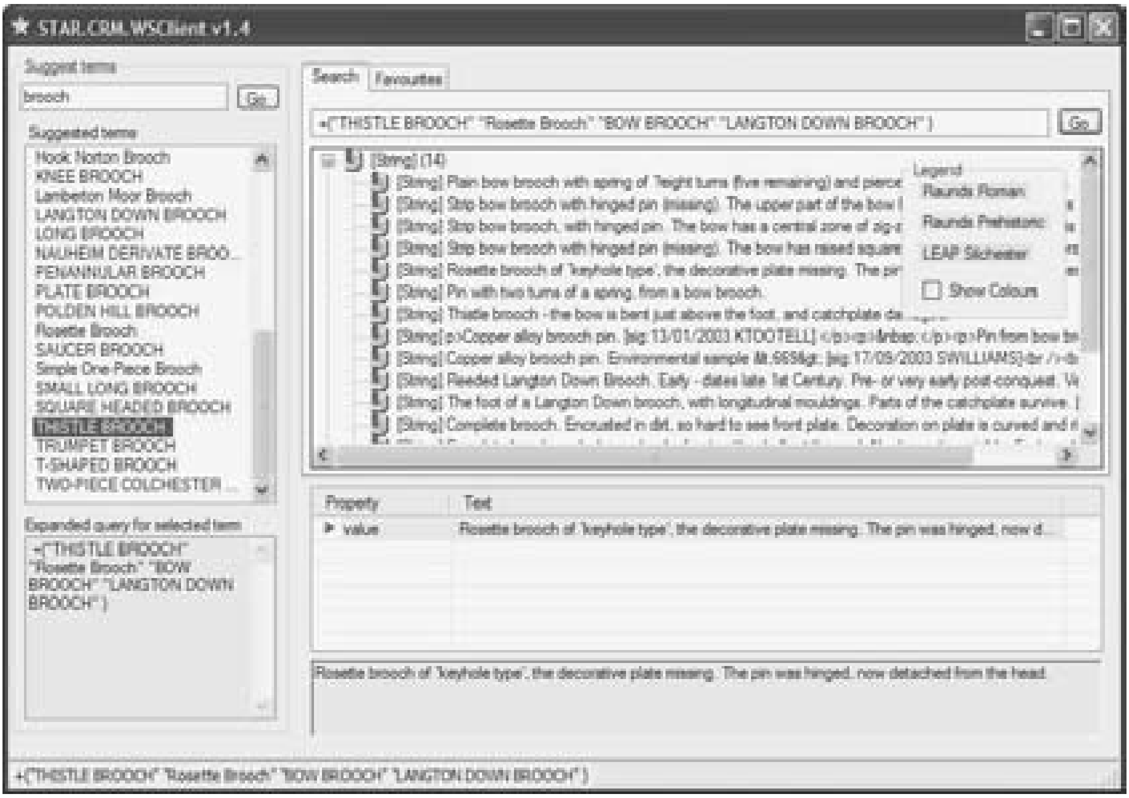
Figure 3: Initial prototype search and browse application

Prototype CRM-EH service

An initial prototype STAR client application (Figure 3), supports cross searching and exploring the amalgamated data extracted from the previously separate databases, which include free text descriptions. This is based on a (STAR Project) CRM based web service for all server interaction. Boolean full-text search operators are available. Result items offer entry points to the structured data; allowing a user to browse to related data items, by following chains of relationships within the CRM-EH, beaming up from data items to concepts as desired.

Figure 3 illustrates a search for a particular kind of brooch using Boolean full-text search operators. Double-clicking a result reveals various properties and relationships to other entities and events, which may be double clicked to continue browsing. Figure 4 shows another version of the STAR client prototype, incorporating the SKOS terminology services (described above) as an initial stage of the interactive search process.

Here the SKOS service suggests various controlled terminology corresponding to the entry term, *brooch*. Some of these specialist terms have been selected for a highly specific query, yielding a particular subset of the many instances of brooch in the combined datasets.



brooch

Hook Norton Brooch
KNEE BROOCH
Lamberton Moor Brooch
LANGTON DOWN BROOCH
LONG BROOCH
NAUHEIM DERIVATE BROO...
PENANNULAR BROOCH
PLATE BROOCH
POLDEN HILL BROOCH
Rosette Brooch
SAUCER BROOCH
Simple One-Piece Brooch
SMALL LONG BROOCH
SQUARE HEADED BROOCH
THISTLE BROOCH
TRUMPET BROOCH
T-SHAPED BROOCH
TWO-PIECE COLCHESTER ...

+("THISTLE BROOCH" "Rosette Brooch" "BOW BROOCH" "LANGTON DOWN BROOCH")

+("THISTLE BROOCH" "Rosette Brooch" "BOW BROOCH" "LANGTON DOWN BROOCH")

[String] (14)
[String] Plain bow brooch with spring of ?eight turns (five remaining) and pierce
[String] Strip bow brooch with hinged pin (missing). The upper part of the bow I
[String] Strip bow brooch, with hinged pin. The bow has a central zone of zig-z
[String] Strip bow brooch with hinged pin (missing). The bow has raised square
[String] Rosette brooch of 'keyhole type', the decorative plate missing. The pi
[String] Pin with two turns of a spring, from a bow brooch.
[String] Thistle brooch - the bow is bent just above the foot, and catchplate da
[String] p>Copper alloy brooch pin. [sig 13/01/2003 KTOOTELL] </p><p>&nbsp;</p><p>Pin from bow br
[String] Copper alloy brooch pin. Environmental sample <#.999>. [sig 17/09/2003 SWILLIAMS]
<b
[String] Reeded Langton Down Brooch. Early - dates late 1st Century. Pre- or very early post-conquest. V
[String] The foot of a Langton Down brooch, with longitudinal mouldings. Parts of the catchplate survive. I
[String] Complete brooch. Encrusted in dirt, so hard to see front plate. Decoration on plate is curved and r

Property	Text
▶ value	Rosette brooch of 'keyhole type', the decorative plate missing. The pin was hinged, now d...

Rosette brooch of 'keyhole type', the decorative plate missing. The pin was hinged, now detached from the head.

+("THISTLE BROOCH" "Rosette Brooch" "BOW BROOCH" "LANGTON DOWN BROOCH")

Figure 4: Integrated SKOS and CRM web services

Conclusions

Work to date has demonstrated the extraction and storage of relational data from multiple databases into separate RDF files based on CRM-EH structure and its subsequent integration to support search and browsing across datasets and from data instance to CRM-EH entities. Combining this data with the CRM-EH ontology opens up the possibility of automated traversal across known relationships, via the SKOS and CRM based web services. Future work includes further data extraction, more advanced search capabilities, information extraction based on the CRM-EH and evaluation with users.

Acknowledgements

The STAR project is funded by the UK Arts and Humanities Research Council (AHRC). Thanks are due to Phil Carlisle & Keith May (English Heritage). Select use has been made of material from a review on Terminology Services and Technology: (Tudhope *et al.* 2006), conducted for and funded by the Joint Information Systems Committee (JISC).

References

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

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)

Binding C., Tudhope D., May K. Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM. Proceedings (ECDL

2008) 12th European Conference on Research and Advanced Technology for Digital Libraries, Aarhus. Lecture Notes in Computer Science, Berlin: Springer. (2008 forthcoming)

CIDOC Conceptual Reference Model (CRM). http://cidoc.ics.forth.gr

CRM-EH Extension to CRM. http://hypermedia.research.glam.ac.uk/kos/CRM/

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

Doerr, M. The CIDOC Conceptual Reference Module: an Ontological Approach to Semantic Interoperability of Metadata. AI Magazine, 2493, 75--92 (2003)

English Heritage. http://www.english-heritage.org.uk/

English Heritage Thesauri. http://thesaurus.english-heritage.org.uk/

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

SEMWEB RDF Library for .NET, http://razor.occams.info/code/semweb

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

SKOS API. SWAD_EUROPE Thesaurus Project Output (2004) http://www.w3.org/2001/sw/Europe/reports/thes/skosapi.html

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

Tudhope D., Koch T., Heery R. Terminology Services and Technology: JISC state of the art review. (2006a) http://www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_technology_review_sep_06.pdf

Tudhope D., Binding C., Blocks D., Cunliffe D. Query expansion via conceptual distance in thesaurus indexed collections. Journal of Documentation, 62 (4), 509–533. Emerald (2006b)

Tudhope D., Binding C., May K. Semantic interoperability issues from a case study in archaeology. In: Stefanos Kollias & Jill Cousins (eds.), Semantic Interoperability in the European Digital Library, Proceedings of the First International Workshop SIEDL 2008, 88–99, associated with 5th European Semantic Web Conference, Tenerife. (2008)

Tudhope D., Binding C. Faceted Thesauri. Axiomathes, 18(2), 211–222, Springer. (2008)

Zeng M, Chan L. Trends and issues in establishing interoperability among knowledge organization systems. Journal of American Society for Information Science and Technology, 55(5): 377 – 395. (2004)

About the writers

Douglas Tudhope, BSc, PhD, Reader
Hypermedia Research Unit, University of Glamorgan
email. dstudhope@glam.ac.uk

Ceri Binding, Research Fellow
Hypermedia Research Unit,
University of Glamorgan
email. cbinding@glam.ac.uk