A STELLAR Role for Knowledge Organization Systems in Digital Archaeology

by Douglas Tudhope, Ceri Binding, Stuart Jeffrey, Keith May and Andreas Vlachidis

EDITOR'S SUMMARY

Research data in archaeology is being made more accessible through the semantic efforts of the STAR and STELLAR projects of two United Kingdom universities. The goal of STAR (Semantic Technologies for Archaeological Resources) is to facilitate semantic interoperability, enabling a structured semantic search of five databases and grey literature reports using an ontology of cultural heritage in combination with other knowledge organization systems. STAR employs natural language processing to identify key concepts and generate semantic metadata to support the unified search. STELLAR (Semantic Technologies Enhancing Links and Linked data for Archaeological Resources) takes the process a step further by simplifying the job of expressing excavation data in CRM ontology terms and then generating representations as linked data. The two projects demonstrate the effectiveness of semantic interoperability methods, coordinating data and vocabularies in a shared framework, and supporting the reuse of archaeological excavation data.

KEYWORDS

archaeology
knowledge organization systems
interoperability
semantic analysis
metadata

Knowledge organization systems (KOS), such as classification systems, gazetteers, ontologies, taxonomies and thesauri, are crucial to semantic interoperability [1, 2]. Controlled vocabularies reduce ambiguity by defining the scope of terms and possibly providing synonyms. More complex KOS organize concepts via different types of semantic relationship. This semantic organization also affords a mechanism (for both human and machine) to connect different choices of terminology and different representations and structures of information.

We give an outline of two recent projects (STAR and STELLAR) that combine different kinds of KOS for semantic interoperability purposes in the archaeological domain, where there is a desire to make excavation datasets available beyond the immediate project team. This availability may be for purposes of facilitating reuse of data, linking interpretation with the underlying evidence in primary data or for general cultural resource management purposes. However, different project databases are organized differently. There may, for example, be different ways of handling archaeological Small Finds and Samples and different methods of handling

Douglas Tudhope is a professor at the University of Glamorgan where he leads the Hypermedia Research Unit and is principal investigator on the STAR and STELLAR projects. He can be reached by email at dstudhope<at>glam.ac.uk.

Ceri Binding is a research fellow at the University of Glamorgan in the Department of Computer and Mathematical Sciences and was responsible for software development on the STAR and STELLAR projects. She can be reached by email at cbinding<at>glam.ac.uk.

Stuart Jeffrey is user services manager for the Archaeology Data Service at the University of York and co-investigator for the STELLAR project. He can be reached by email at sj523<at>york.ac.uk.

Keith May is an archaeologist with responsibilities for digital research strategies at English Heritage and collaborator on the STAR and STELLAR projects. He can be reached by email at keith.may<at>english-heritage.org.uk.

Andreas Vlachidis is a research assistant at the University of Glamorgan in the Department of Computer and Mathematical Sciences and was responsible for natural language processing work. He can be reached by email at avlachid<at>glam.ac.uk.

the phasing analysis of a site, which groups excavated elements into more complex archaeological features. Furthermore, different databases will typically employ different terminology; one organization may tend to refer to *post holes*, while another organization's usual vocabulary may be *post-holes*. These variations are the perennial issues of vocabulary control in information science but are combined here with the data integration of diverse datasets. Another challenge is posed by the existence of excavation reports in the form of grey literature, unpublished reports resulting from projects funded for their excavation and basic reporting phase, but not for full publication. In fact these cases represent the majority of all archaeological projects.

The STAR Project – Cross Searching Archaeological Datasets and Grey Literature

The STAR (Semantic Technologies for Archaeological Resources) project was a collaboration between the Hypermedia Research Unit at the University of Glamorgan and English Heritage (EH) [3]. STAR began in 2007 and finished summer 2010 [4]. The project aimed to provide a degree of semantic interoperability between diverse archaeological datasets (from different projects and organizations) and archaeological reports, extracted from the OASIS (Online AccesS to the Index of archaeological investigationS) grey literature library, provided by the Archaeological Data Service [5, 6].

To this end, the CIDOC Conceptual Reference Model (CRM), an ISO standard in the realm of cultural heritage, was adopted as a unifying upper-level conceptual framework [7]. Since the CRM operates at a high level of generality, the datasets were mapped to the CRM-EH extension of the CRM, developed by English Heritage [8]. The CRM-EH specializes in the CRM classes for Physical Object and Place to archaeological subclasses such as Find and Context [9]. A mapping/extraction tool assisted the intellectual mapping of data elements to the ontology.

The online STAR Demonstrator cross searches excavation datasets from five different databases, together with an extract of excavation reports from the OASIS grey literature library. Previously cross search was not possible; each dataset remained in its own silo, and no link was made to grey literature. The Demonstrator seeks to hide the complexity of the underlying ontology, while offering structured semantic search. An interactive query builder offers search (and browsing) for key archaeological concepts, such as Samples, Finds, Contexts or interpretive Groups with their properties and relationships. As the user selects from the interface, an underlying semantic query is automatically constructed in terms of the corresponding ontological entities.

STAR employs a web-service architecture for programmatic access to the CRM and various glossaries and thesauri. The latter are represented in the W3C standard SKOS format, a formal RDF/XML representation [10]. The EH Thesauri [11] are available for programmatic access via the API, which is based on a subset of the SWAD [12] Europe SKOS API, with extensions for semantic concept expansion [13, 14]. The SKOS (Simple Knowledge Organization System) services are accompanied by a variety of user interface widgets (browser neutral) that can be integrated into browser-based user interfaces, where browsing of concept structures or concept-based search is required.

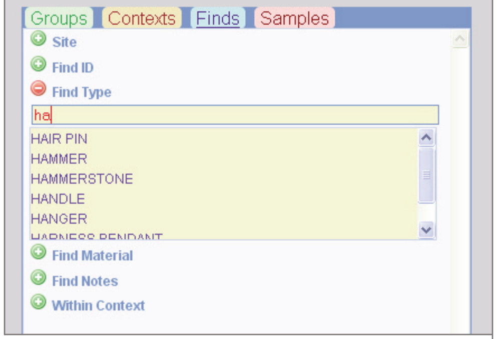
Natural language processing (NLP) information extraction techniques were applied to key concepts in the grey literature, producing semantic metadata in the same CRM-EH based representation as the extracted data. This NLP-produced metadata allows unified searching of the different datasets and the grey literature in terms of the semantic structure of the CRM-EH ontology.

The CRM and CRM-EH do not supply a vocabulary of concepts beyond the class names in the ontology. Therefore, various KOS were used in conjunction with the ontology for various purposes. An (extended) set of EH glossaries were closely identified with associated fields in the datasets. In some cases, fields were effectively controlled, but in other cases an intellectual alignment operation was required for semi-controlled fields – an important aspect of the work. These SKOS glossaries afforded semantic search in the Demonstrator based on URI identifiers, with controlled terms being suggested by the query builder in the Demonstrator. No glossary was available for the Sample fields and the resulting difference can be observed

CONTENTS TOP OF ARTICLE < PREVIOUS PAGE NEXT PAGE > NEXT ARTICLE >

in the lack of terminology control. Figure 1 gives an example of controlled term suggestion for archaeological Finds from the STAR Demonstrator.

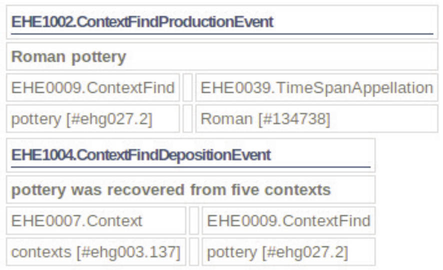
FIGURE 1. Controlled term suggestion in the STAR Demonstrator.



In addition to the glossaries, several thesauri were available. A materials thesaurus supplied controlled vocabulary for the Demonstrator. Other thesauri were useful terminology resources but somewhat loosely related to the datasets, being designed for broader purposes. An early pilot client application employed the (SKOS) terminology web service for query expansion, effectively acting as a search thesaurus over the free text note fields in the datasets.

An informal mapping between glossaries and thesaurus hierarchies with CRM (and CRM-EH) classes supported the NLP. In basic operation, the glossaries provide the vocabulary for named entity recognition, one element of the rule-based information extraction system. Both SKOS concepts and CRM (and CRM-EH) classes can be associated with phrases in the grey literature reports. In recall enhancing mode, where a match between a glossary term and a thesaurus concept has been established, broader and narrower expansion possibilities are available. Of course, the usual precision/recall trade-offs such as false positives apply to this expansion and to the NLP techniques more generally. It is therefore important to retain the provenance of triples generated by NLP methods, for appropriate visualization in downstream applications consuming this data. Figure 2 gives an example of automatic SKOS and CRM-EH annotations using NLP techniques.

FIGURE 2. Information extraction example: "Roman pottery was recovered from five contexts."


EHE1002.ContextFindProductionEvent

Roman pottery

EHE0009.ContextFind	EHE0039.TimeSpanAppellation
pottery [#ehg027.2]	Roman [#134738]

EHE1004.ContextFindDepositionEvent

pottery was recovered from five contexts

EHE0007.Context	EHE0009.ContextFind
contexts [#ehg003.137]	pottery [#ehg027.2]

The STELLAR Project – Tools for Generating CRM/SKOS-Based RDF and Linked Data

The STAR project served as the launching point for the current STELLAR project. STELLAR (Semantic Technologies Enhancing Links and Linked data for Archaeological Resources) aims to generalize and extend the data extraction tools produced by STAR to facilitate their adoption by third-party data providers [15]. This project is a collaboration between Glamorgan and the Archaeological Data Service at the University of York, with English Heritage as project partners. The extracted data is represented in standard RDF formats that allow the datasets to be cross searched and linked by a variety of Semantic Web tools, following a linked-data approach.

The aim is to make it easier for data owners who are not ontology specialists to express their excavation data in terms of the CRM (and CRM-EH) and to generate semantic/linked-data representations. The STELLAR tools convert archaeological data to RDF in a consistent manner, without requiring detailed knowledge of the underlying ontology.

These tools work from a set of templates (implemented as XSL stylesheets) that express commonly occurring patterns encountered in the STAR project. The current set of templates corresponds to the general aim of cross-searching excavation datasets for inter-site analysis and comparison. Different templates that drew on other areas of the ontology could be designed for purposes such as project management or detailed intra-site analysis. Each template is a combination of various optional elements with a mandatory ID. The ID is prefixed with a namespace (a tool parameter) to generate URIs. Thus the RDF output is produced in a form that facilitates subsequent expression as linked data.

In addition to CRM-based templates, there is a template allowing a glossary/thesaurus connected with the dataset to be expressed in SKOS. The CRM templates have elements giving the (preferred) option of expressing controlled data items as SKOS URIs (either to local vocabularies generated by the SKOS template, or to Linked Data publications of a major SKOS vocabulary). The templates are available from the STELLAR project website, along with the tools that operate over the templates. To generate RDF, the user chooses a template for a particular data pattern and supplies the

CONTENTS
TOP OF ARTICLE
< PREVIOUS PAGE
NEXT PAGE >
NEXT ARTICLE >

corresponding input from their database. Documentation and a tutorial are available on the website. People interested in applying the tools or making use of the templates are encouraged to get in touch. The final stage of the project will publish a collection of linked data from a selection of the excavation datasets archived by the Archaeological Data Service, which have been extracted by the STELLAR tools.

Conclusion

In general, the information science community has innovated in many different aspects of knowledge organization systems, providing exemplars that other communities can follow. This article provides an overview of two specific projects conducted by partnerships among the University of Glamorgan, the Archaeology Data Service and English Heritage. The STAR project has demonstrated the viability of the methods for semantic interoperability between diverse archaeology datasets and grey literature. The current STELLAR project, which will finish later this year, aims to make it easier to express archaeological data and vocabularies in a common semantic framework. This framework serves the twin goals of opening data for reuse and opening up the much broader range of research questions that might be answered when we connect currently isolated excavation datasets.

Acknowledgements

The STAR and STELLAR projects were funded by the U.K. Arts and Humanities Research Council (AHRC). Thanks are due to Phil Carlisle (English Heritage) for assistance with EH thesauri and the Archaeological Data Service for assistance with the OASIS grey literature digital library. ■

Resources Mentioned in the Article

[1] Tudhope, D., Koch, T., & Heery, R. (September 15, 2006). *Terminology services and technology: JISC state of the art review.* Bath, England: UKOLN. Retrieved March 7, 2011, from www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_technology_review_sep_06.pdf.

[2] Golub, K., & Tudhope, D. (March 7, 2009) *Terminology Registry Scoping Study (TRSS): Final report (JISC Report).* Bath, England.: UKOLN. Retrieved March 7, 2011, from www.jisc.ac.uk/media/documents/programmes/sharedservices/trss-report-final.pdf.

[3] *English Heritage:* www.english-heritage.org.uk/.

[4] *STAR Project:* http://hypermedia.research.glam.ac.uk/kos/star.

[5] *Archaeological Data Service:* http://archaeologydataservice.ac.uk/.

[6] *OASIS (Online AccesS to the Index of archaeological investigationS).* Available from the Archaeological Data Service at www.oasis.ac.uk/.

[7] *CIDOC CRM (Conceptual Reference Model):* http://cidoc.ics.forth.gr.

[8] *CRM-EH: English Heritage Extension to CRM for the archaeology domain:* http://hypermedia.research.glam.ac.uk/kos/CRM/.

[9] 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). Paris: ICOM. Retrieved March 7, 2011, from http://cidoc.ics.forth.gr/technical_papers.html.

[10] *SKOS: Simple Knowledge Organization Systems - W3C Semantic Web Deployment Working Group:* www.w3.org/2004/02/skos.

[11] *English Heritage Thesauri:* http://thesaurus.english-heritage.org.uk/.

[12] *SWAD-Europe* (Semantic Web Advanced Development for Europe) was a project that ran from May 2002 to October 2004 to support W3C's Semantic Web initiative in Europe: www.w3.org/2001/sw/Europe.

[13] Tudhope, D., & Binding, C. (June/July, 2006). Towards terminology services: Experiences with a pilot web service thesaurus browser. *Bulletin of the American Society for Information Science and Technology, 32*(5), 6–9. Retrieved March 7, 2011, from www.asist.org/Bulletin/Jun-06/tudhope_binding.html.

[14] Binding, C., & Tudhope, D. (2010). Terminology web services. *Knowledge Organization, 37*(4), 287–298.

[15] STELLAR Project: http://hypermedia.research.glam.ac.uk/kos/stellar/.

CONTENTS
TOP OF ARTICLE
< PREVIOUS PAGE
NEXT PAGE >
NEXT ARTICLE >