Semantic Technologies for Archaeology Resources: Results from the STAR project

C. Binding[1], K. May[2], R. Souza[1], D. Tudhope[1], A. Vlachidis[1]

[1] Hypermedia Research Unit, University of Glamorgan, UK
[2] English Heritage, UK
{cbinding, rsouza, dstudhope, avlachid}@glam.ac.uk; Keith.May@english-heritage.org.uk

Abstract

Outcomes from the STAR Project are presented. The underlying rationale is the need to widen access to archaeological datasets, which will allow third parties to cross search different datasets and investigate the basis for interpretations in the underlying data. The semantic technologies employed are based on standard representations of domain vocabularies and the underlying core ontology, an archaeological extension of the CIDOC CRM. Methods for extracting semantic RDF representations from the datasets are described, together with Natural Language Processing techniques for information extraction from a selection of OASIS grey literature. STAR web services and semantic search implementations are presented. The need for controlled terminology is emphasised. Illustrative results from the semantic search Demonstrator are discussed.

Keywords: Cross searching, CIDOC CRM, core ontology, grey literature, information extraction, natural language processing, semantic interoperability, semi-automatic mapping tool, thesaurus

1. Introduction

Cultural heritage organisations are looking to open digital collections and databases, previously confined to specialists, to a wider audience. 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. Additionally, different datasets may employ different schema for semantically equivalent information. Entities may have different names but refer to the same underlying concept.

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.

Ontologies offer a high level domain conceptualisation with formal definition of roles and semantic relationships. Within cultural heritage, 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.

2. STAR Project aims

STAR (Semantic Technologies for Archaeological Resources) is funded by the UK Arts and Humanities Research Council (AHRC). It is a collaboration between the Hypermedia Research Unit at the University of Glamorgan and English Heritage (EH).

STAR has aimed to address these semantic interoperability concerns by utilising semantic terminology tools to link digital archive databases, vocabularies and the associated grey literature, exploiting the potential of the CRM-EH.

The STAR system cross searches over excavation datasets from different database schemas, including Raunds Roman, Raunds Prehistoric, Museum of London, Silchester Roman (LEAP) and Stanwick sampling. The system also cross searches across an extract of excavation reports from the OASIS index of grey literature, operated by the Archaeological Data Service (ADS).

Figure 1 shows the general architecture of the STAR system and the various data and conceptual resources.

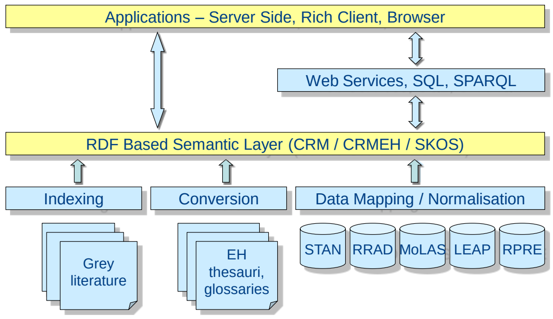

Applications – Server Side, Rich Client, Browser
Web Services, SQL, SPARQL
RDF Based Semantic Layer (CRM / CRMEH / SKOS)
Indexing
Conversion
Data Mapping / Normalisation
Grey literature
EH thesauri, glossaries
STAN
RRAD
MoLAS
LEAP
RPRE

Figure 1: *STAR architecture*

STAR employs a web service architecture for accessing the data and ontology expressed as RDF statements, together with terminology resources, such as thesauri (Figure 2).

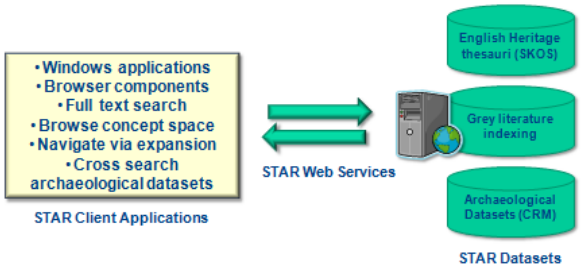


Figure 2: *STAR Web Services and Client Applications*

2.1. EH extension of CRM

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 in English Heritage (CRIPPS et al. 2004, MAY et al. 2008, 2009). 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 (an example is given in Figure 3). Working with May, an implementation of the CRM-EH has been produced as a modular RDF (Resource Description Framework) extension, currently referencing the published (v4.2) RDFS implementation of the CRM.

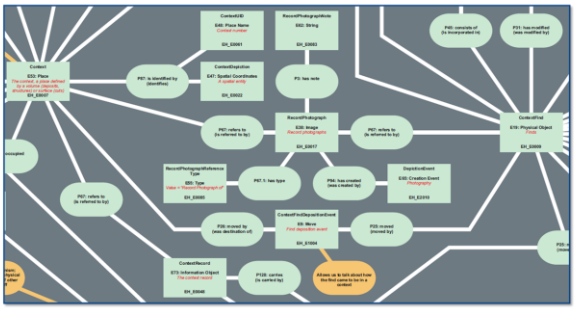

ContextUID
E48: Place Name
Context number
EH_E0061
RecordPhotographNote
E62: String
EH_E0063
P46: consists of (is incorporated in)
P31: has modified (was modified by)
Context
E53: Place
The context, a place defined by a volume (deposits, structures) or surface (cuts)
EH_E0007
P87: is identified by (identifies)
ContextDepiction
E47: Spatial Coordinates
A spatial entity
EH_E0022
P3: has note
occupied
P67: refers to (is referred to by)
RecordPhotograph
E38: Image
Record photographs
EH_E0017
P67: refers to (is referred to by)
ContextFind
E19: Physical Object
Finds
EH_E0009
RecordPhotographReference Type
E55: Type
Value = "Record Photograph of"
EH_E0085
P67.1: has type
P94: has created (was created by)
DepictionEvent
E65: Creation Event
Photography
EH_E2010
P67: refers to (is referred to by)
P26: moved by (was destination of)
ContextFindDepositionEvent
E9: Move
Find deposition event
EH_E1004
P25: moved (moved by)
ContextRecord
E73: Information Object
The context record
EH_E0048
P128: carries (is carried by)
Allows us to talk about how the find came to be in a context

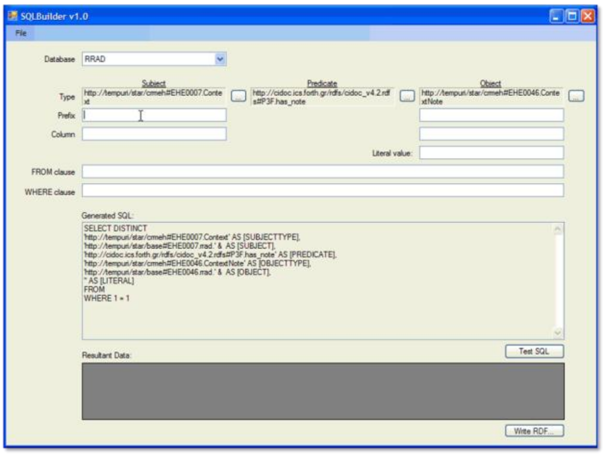
Figure 3: *Example of the CRM-EH showing Context and ContextFind*

3. Methods

The work has required methods to be developed in mapping datasets to the core ontology, extracting semantic web representations in RDF, applying Natural Language Processing (NLP) techniques to free text grey literature documents in order to index them with CRM-EH entities and developing semantic search techniques that operate over RDF generated from both datasets and grey literature.

3.1. Mapping datasets to CRM-EH

Domain expert, May, generated spreadsheets showing the key mappings from the various datasets to the CRM-EH. These selections from the different databases were extracted via SQL queries, and stored as separate RDF files (simplifying the process). This intellectual work was significantly assisted by a mapping and data extraction tool (BINDING et al. 2008), which automatically generates the RDF statements after the interactive mappings are set up. Figure 4 shows an example, where the user is extracting RDF statements for a context note from the selected database fields.



Database: RRAD

	Subject	Predicate	Object
Type	http://tempuri/star/crmeh#EHE0007.Context	http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P3F.has_note	http://tempuri/star/crmeh#EHE0046.ContextNote
Prefix			
Column			
Literal value:			

FROM clause:

WHERE clause:

Generated SQL:

```
SELECT DISTINCT
'http://tempuri/star/crmeh#EHE0007.Context' AS [SUBJECTTYPE],
'http://tempuri/star/base#EHE0007.rrad.' & AS [SUBJECT],
'http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#P3F.has_note' AS [PREDICATE],
'http://tempuri/star/crmeh#EHE0046.ContextNote' AS [OBJECTTYPE],
'http://tempuri/star/base#EHE0046.rrad.' & AS [OBJECT],
'' AS [LITERAL]
FROM
WHERE 1 = 1
```

Resultant Data:

Figure 4: *STAR mapping and data extraction tool*

3.2. SKOS Terminology Services

STAR employs the Simple Knowledge Organization System (SKOS) standard as the representation format for domain thesauri and related Knowledge Organization Systems (KOS). SKOS is a W3C Recommendation

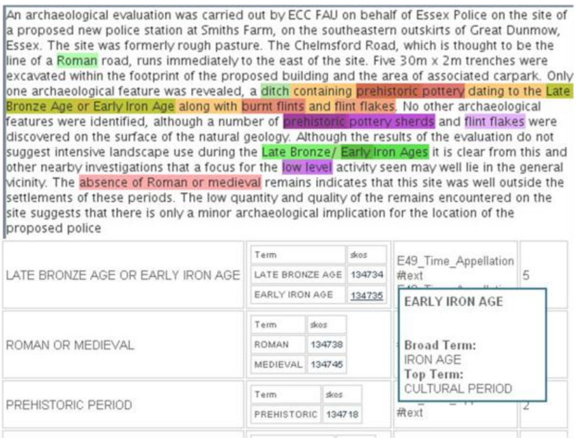
RDF representation for KOS and is based on a formal data model. It is intended as a formal RDF/XML representation standard for KOS designed for information retrieval purposes. This offers a cost effective approach for dealing with thesauri and dataset glossaries. In STAR SKOS representations are connected to the CRM ontology via the CRM Type entity.

Terminology web services (BINDING et al. 2010) have been developed based upon SKOS thesaurus representations. The service is based on a subset of the SWAD Europe SKOS API, with extensions for semantic concept expansion (BINDING et al. 2004).

These terminology services allow access to the SKOS thesauri and glossaries in a variety of (browser neutral) user interface widgets. They can be employed in a wide variety of applications for both data entry and display purposes, where access to controlled terminology, browsing of concept structures or query expansion is required.

3.3. Natural Language Processing of Grey Literature

Cross searching of the grey literature is based on annotations produced to the same CRM-EH based RDF representation as the database extractions. NLP techniques are employed via the General Architecture for Text Engineering (GATE) toolkit. This is driven by the domain glossaries and thesauri. NLP rules produce CRM (and CRM-EH) conformant annotations, which are subsequently extracted as RDF triples, in the same form as those produced by the dataset extraction. Figure 5 shows an example of the rule-based annotation of a summary of a grey literature report.



An archaeological evaluation was carried out by ECC FAU on behalf of Essex Police on the site of a proposed new police station at Smiths Farm, on the southeastern outskirts of Great Dunmow, Essex. The site was formerly rough pasture. The Chelmsford Road, which is thought to be the line of a Roman road, runs immediately to the east of the site. Five 30m x 2m trenches were excavated within the footprint of the proposed building and the area of associated carpark. Only one archaeological feature was revealed, a ditch containing prehistoric pottery dating to the Late Bronze Age or Early Iron Age along with burnt flints and flint flakes. No other archaeological features were identified, although a number of prehistoric pottery sherds and flint flakes were discovered on the surface of the natural geology. Although the results of the evaluation do not suggest intensive landscape use during the Late Bronze/ Early Iron Ages it is clear from this and other nearby investigations that a focus for the low level activity seen may well lie in the general vicinity. The absence of Roman or medieval remains indicates that this site was well outside the settlements of these periods. The low quantity and quality of the remains encountered on the site suggests that there is only a minor archaeological implication for the location of the proposed police

LATE BRONZE AGE OR EARLY IRON AGE	Term skos LATE BRONZE AGE 134734 EARLY IRON AGE 134735	E49_Time_Appellation #text	5
ROMAN OR MEDIEVAL	Term skos ROMAN 134738 MEDIEVAL 134745		
PREHISTORIC PERIOD	Term skos PREHISTORIC 134718	#text	2

EARLY IRON AGE

Broad Term:
IRON AGE
Top Term:
CULTURAL PERIOD

Figure 5: *Example of annotation of time period entities*

Gazetteer lists have been derived from EH and other database glossaries. Mapping between gazetteer terms and ontological classes is facilitated in GATE, using the Ontogazetteer utility. The ontological classes E49.Time Appellation, E19.Physical Object, and E53.Place have been mapped against individual gazetteer lists, containing terms originating from EH Thesauri and relevant glossaries. Figure 6 gives an example of the effect of pattern matching rules driven by these CRM ontological classes.

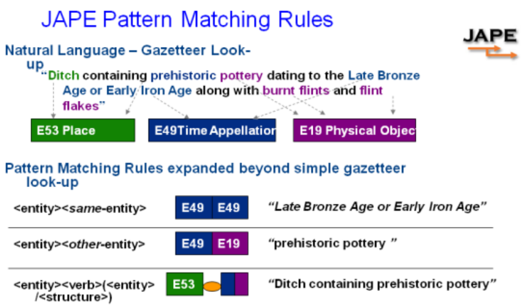

JAPE Pattern Matching Rules
JAPE
Natural Language – Gazetteer Look-up
"Ditch containing prehistoric pottery dating to the Late Bronze Age or Early Iron Age along with burnt flints and flint flakes"
E53 Place
E49Time Appellation
E19 Physical Object
Pattern Matching Rules expanded beyond simple gazetteer look-up
<entity><same-entity>
E49 E49
"Late Bronze Age or Early Iron Age"
<entity><other-entity>
E49 E19
"prehistoric pottery "
<entity><verb>(<entity>/<structure>)
E53
"Ditch containing prehistoric pottery"

Figure 6: *Example of CRM based Pattern Marching*

The CRM-EH model has also been used in the definition of more specialised pattern matching rules to exploit hierarchical relationships. Matching rules incorporated a simple ontological parent-child relationship between the parent entity E53.Place and child CRM-EH entities EH_E0007 Context and EH_E0005 Group. Within the archaeology domain, the term context is specified as the context of an archaeological excavation and is modelled as a Place.

Pattern matching techniques include linguistic evidence of combinations between entities and verb phrases in the form of <entity><verb><entity>. This technique investigates the assumption that text phrases carry information which describes relationships between CRM-EH entities and that linguistic evidence in the form of pattern matching rules can be employed to extract such textual instances. Extraction of the entity Context_Find which refers to physical objects found during excavation can benefit from the above technique. Phrases denoting association of a physical object with a Period and a Context strengthen the validity of the rule matching mechanism by assuming a physical object as a find, as for example in the phrase *' ...pits containing Iron Age flint...'*.

An exercise with an initial rule set produced a set of annotations types relating to the entities Time Appellation, Physical Object, and Place, together with the assignment of unique ontological identifiers and terminological references to the individual annotations.

Pilot evaluation results with the initial rule set were encouraging. For Time Appellation concepts, the extraction mechanism achieved recall and precision rates over 70%. The evaluation regarding the extraction of Physical Objects and Places revealed the capability of the mechanism to target such concepts. However, the pilot evaluation demonstrated that the limited amount of glossary terms covering Places influenced the performance of the mechanism.

Subsequent work extended the coverage of Place (in the sense of archaeological excavation Contexts), drawing on the English Heritage glossary of Simple Names for Deposits and Cuts and other resources, including the English Heritage Thesaurus of Monument Types. This was also extended by terms contributed by a noun phrase frequency analysis conducted over OASIS grey literature documents. Rules have also been refined, allowing the extraction of CIDOC CRM entities, such as Physical Object, Material, (time) Period and more specialised CRM-EH entities, such as Context, Context-Find. The information extraction process can be run in various configurations, such as recall enhancing or (the stricter) precision enhancing. Work is ongoing to further refine the rules, to generalise the method to related cultural heritage domains and to explore the possibilities of web service realisations.

4. STAR Demonstrator

5. The STAR Demonstrator supports cross search and browsing of the instance data and the conceptual model. A common RDF data store holds the CRM-EH ontology, associated thesauri and the extracted instance data, both from the datasets and the grey literature.

Previous versions of the STAR demonstrator supported conventional free text search over database text fields (with interactive query expansion via the SKOS terminology services), along with browsing of the full native ontology structure (May et al. 2009). The current demonstrator focuses upon semantic search (via SPARQL, the Semantic Web RDF query language) afforded by a user interface which seeks to hide the complexity of the underlying ontology. The interface is based upon scenarios informed by a series of workshops with collaborators and archaeological domain experts. Queries can be constructed (by selecting the relevant tab) to search for Samples, Finds, Contexts or interpretive Groups (of contexts). Each entity has various properties, for example a Context can be within a Group, can contain another Context, can contain Finds, Samples. As the user interactively defines a query using the interface, an underlying SPARQL query is automatically constructed in terms of the corresponding CRM-EH ontological entities. Figure 7 shows the user interface for Context queries, which can also be defined in terms of the stratigraphic relationships in the ontological model.

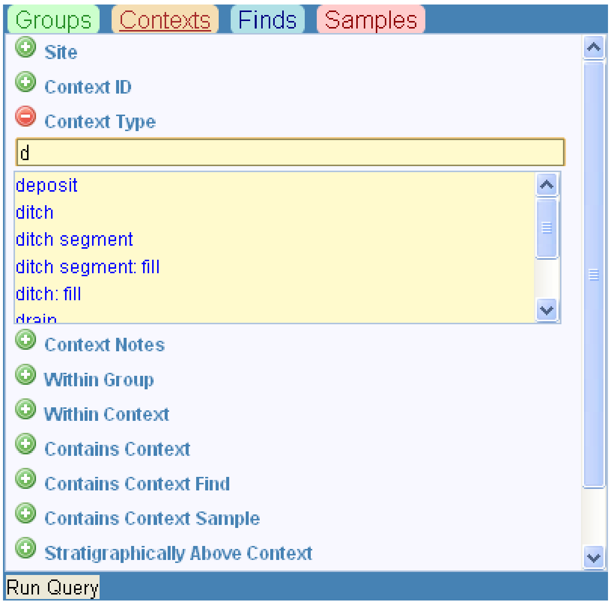


Figure 7: *Screen shot from Demonstrator showing Query Builder for CRM-EH entities, with use of controlled types*

Figure 7 also illustrates the use of controlled types from a relevant Context type SKOS vocabulary. As the user types into the field, an auto-completion widget offers a list of controlled types that is narrowed down as the user continues to type. In the current implementation, only concepts that have been used to index the data are returned.

In contrast, Figure 8 illustrates the problem when no terminology control has been applied. Currently, the Demonstrator has no controlled vocabulary for Sample types. Consequently all the variants, both within a given dataset (perhaps entered by different archaeologists) and between the different datasets are replicated within the RDF representation. Here we see *Lab*, *Lab sample*, *Lab samples* and, in fact, scrolling down *Laboratory* is also to be encountered.

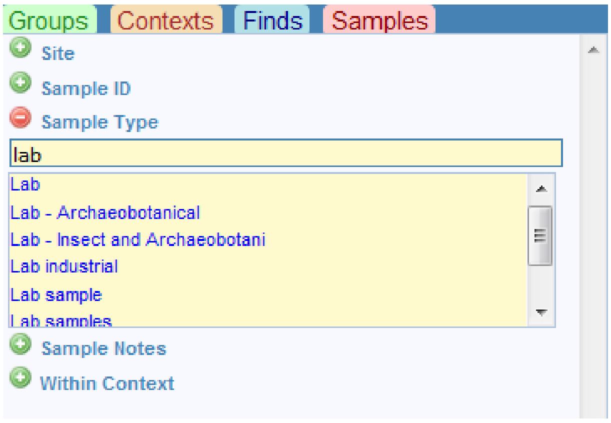


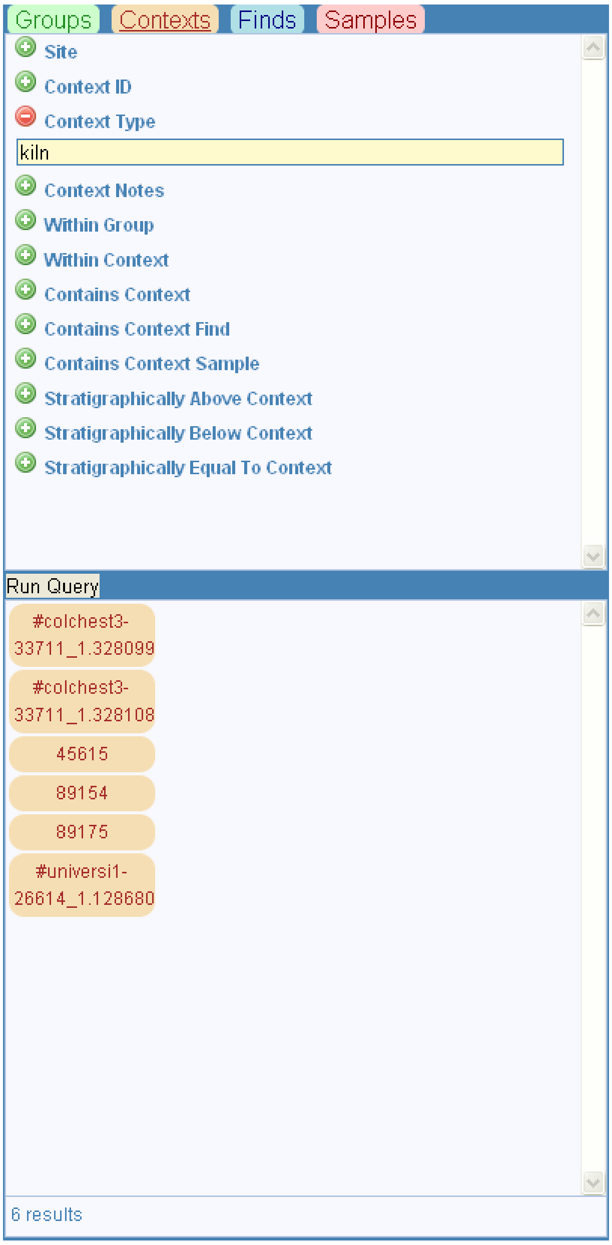
Figure 8: *Screen shot from Demonstrator showing Query Builder and problem of uncontrolled types*

For a given result, it is possible to view properties extracted into RDF for the CRM-EH entities, corresponding to Groups, Contexts, Finds, Samples (which are colour coded in the interface). The single column Figure 11 at the end of the paper shows the results from a query for Contexts of type *Post-hole* containing a Sample with a Note mentioning the term, *grain*. Six results are returned (three from one dataset and three from another). We see that Context result 903 contains both Finds and a Sample, while forming part of a Group of type *Drain*. Corresponding Find and Sample views have been explored for this result in the Figure. The Sample view, however, illustrates a consequence of uncontrolled terminology when searching free text fields, in that the term, *grain*, is ambiguous. Three of the results refer to the presence of a cereal grain in the sample, while in the other three results, *grain* refers to a unit of measurement (as in Figure 11).

This illustrates the importance of the process of *semantic enrichment* of key fields in the datasets, such that types of Contexts and Finds etc are represented by unambiguous SKOS identifiers from standard vocabularies. Where the data is not controlled however, this carries a resource overhead in transforming the data for semantic search. Depending on the situation, the overhead of enforcing controlled terminology on all free text notes within the datasets may not be cost effective and some ambiguity may be left to the user to resolve. The need for disambiguation of natural language is also critical for the NLP techniques employed on the grey literature (section 3.3) and several rules have been developed for this purpose. In future work, we intend to extend this to significant free text fields within datasets.

Due to the standard Semantic Web RDF representation and the common RDF store for all data entities, it is possible to cross search, not just over the five different datasets but also the extract of the (OASIS) grey literature, which is represented in a similar manner, as discussed in Section 3.3. Figure 9 shows a search for Contexts of type kiln.

In addition to three dataset results, we see three results from the grey literature information extraction, expressed as RDF.



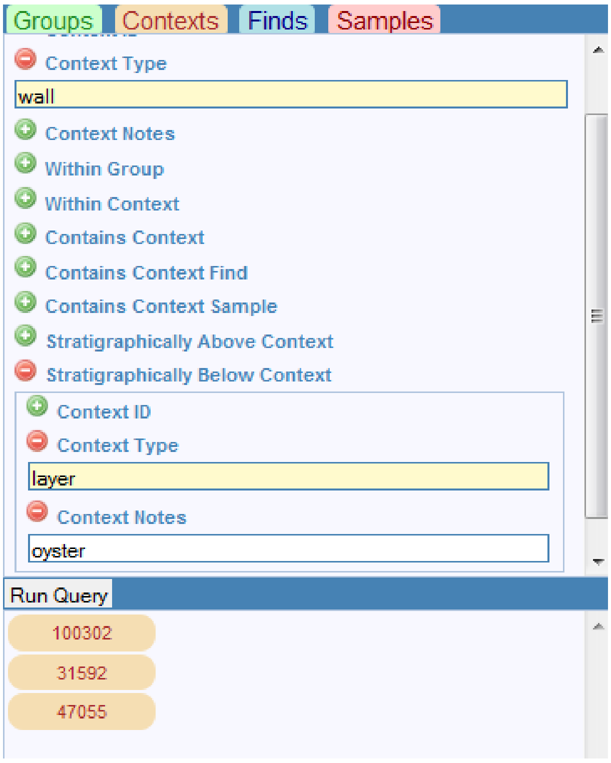
Context Type: kiln

#colchest3-33711_1.328099
#colchest3-33711_1.328108
45615
89154
89175
#universi1-26614_1.128680

6 results

Figure 9: *Screen shot from Demonstrator showing cross search of datasets and grey literature*

The demonstrator also affords semantic search via stratigraphic relationships. For example, Figure 10 shows a query for Contexts of type *Wall*, stratigraphically below Contexts of type *Layer*, which have an associated Note that mentions the term, *oyster*.



Context Type

wall

Context Notes

Within Group

Within Context

Contains Context

Contains Context Find

Contains Context Sample

Stratigraphically Above Context

Stratigraphically Below Context

Context ID

Context Type

layer

Context Notes

oyster

100302

31592

47055

Figure 10: *Screen shot from Demonstrator showing semantic search via stratigraphic relationships.*

Three results are shown for the query. It is possible to browse the Context results, either hierarchically (to Groups or containing Contexts), or stratigraphically. The single column Figure 12 at the end of the paper illustrates this for result 31592.

6. Conclusions

This paper presents the current results from the STAR project, particularly the semantic search interface across disparate datasets, combined with CRM-based information extracted from archaeological grey literature.

The Demonstrator affords cross searching and exploring the amalgamated data extracted from the previously separate databases, which also include free text descriptions. This is based on a (STAR Project) CRM based web service for all server interaction. Result items offer entry points to the structured data; allowing a user to browse to related data items, where the user interface hides the details of the underlying chains of relationships within the CRM-EH ontology.

The information extraction techniques allow RDF triples to be extracted in the same CRM-EH format as the extracted data for cross search purposes.

We believe that the combination of semantic technologies and standards with NLP techniques holds significant promise for the archaeology domain. In future work, we intend to develop these techniques. Immediate steps are to generalise the mapping/extraction tool and the NLP techniques for wider use beyond the immediate project. We also intend to explore the possibilities of semantic terminology services and associated widgets, along with 'linked data' representations.

References

ADS. Archaeological Data Service.

http://ads.ahds.ac.uk/ (accessed 25 January 2010)

BINDING C., TUDHOPE D., 2004. KOS at your Service: Programmatic Access to Knowledge Organisation Systems. Journal of Digital Information, 4(4), http://journals.tdl.org/jodi/article/view/110/109 (accessed 25 January 2010)

BINDING C., TUDHOPE D., MAY K., 2008. 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, 5173, pp: 280–290. Berlin: Springer.

Binding C., Tudhope D. 2010 forthcoming. Terminology services. Knowledge Organization 37(4).

CIDOC Conceptual Reference Model (CRM). http://cidoc.ics.forth.gr (accessed 25 January 2010)

CRM-EH Extension to CIDOC CRM ontology.

http://hypermedia.research.glam.ac.uk/kos/CRM/ (accessed 25 January 2010)

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 (accessed 25 January 2010)

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

ENGLISH HERITAGE.

http://www.english-heritage.org.uk/ (accessed 25 January 2010)

ENGLISH HERITAGE THESAURI.

http://thesaurus.english-heritage.org.uk/

GATE. General Architecture for Text Engineering, http://gate.ac.uk/ (accessed 25 January 2010)

MAY K., BINDING C., TUDHOPE D., 2008. A STAR is born: some emerging Semantic Technologies for Archaeological Resources. Proceedings Computer Applications and Quantitative Methods in Archaeology (CAA2008), Budapest.

May K., Binding C., Tudhope D. 2009. Following a STAR? Shedding more light on Semantic Technologies for Archaeological Resources. Proceedings Computer Applications and Quantitative Methods in Archaeology (CAA2009), Williamsburg.

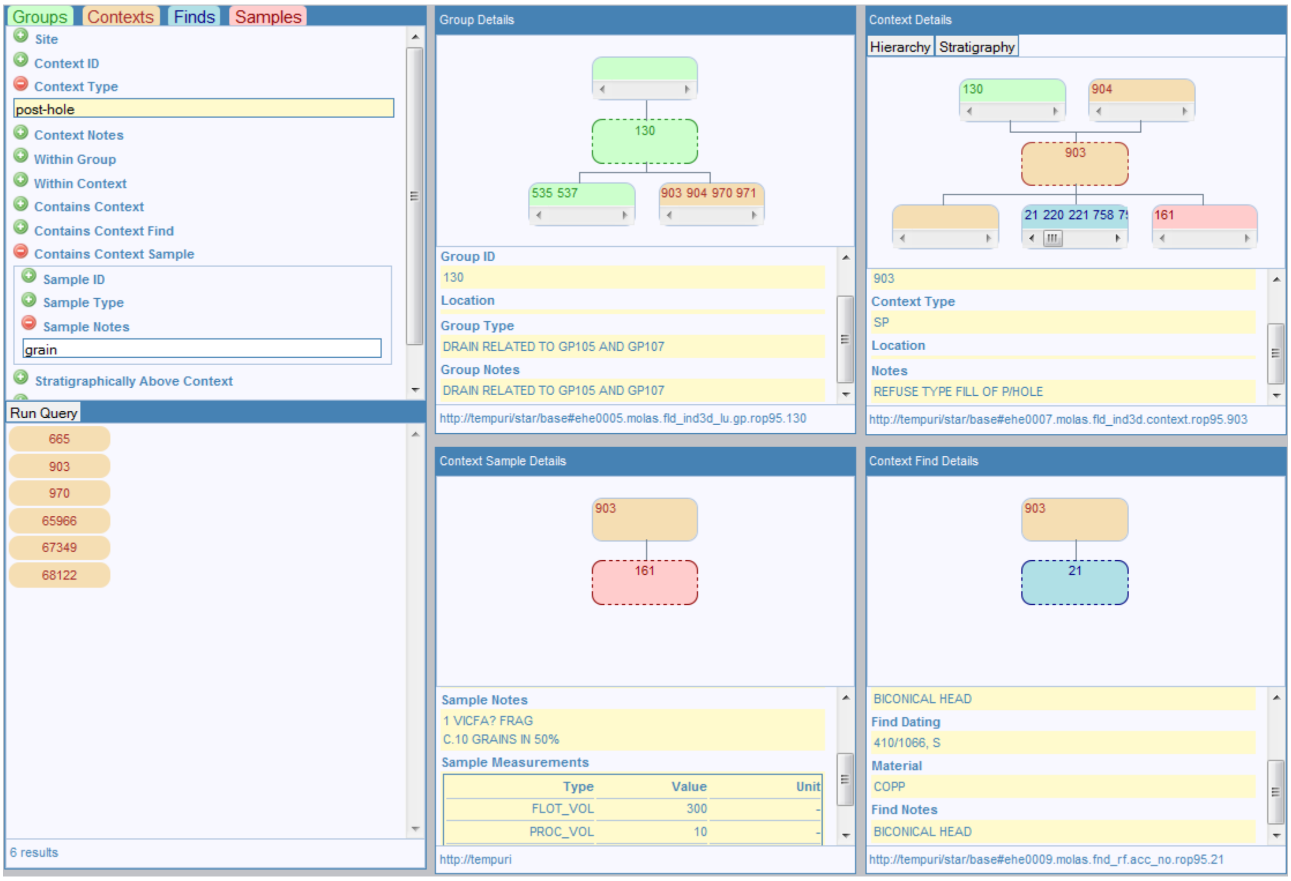
OASIS. Online AccesS to the Index of archaeological investigations. http://www.oasis.ac.uk/ (accessed 25 January 2010)

SKOS. Simple Knowledge Organization System, http://www.w3.org/2004/02/skos (accessed 25 January 2010)

STAR PROJECT. Semantic Technologies for Archaeological Resources.

http://hypermedia.research.glam.ac.uk/kos/star (accessed 25 January 2010)

Vlachidis A, Binding C, May K, Tudhope D. 2010 forthcoming. Excavating Grey Literature: a case study on the rich indexing of archaeological documents via Natural Language Processing techniques and Knowledge Based resources. ASLIB Proceedings journal, Vol. 62, No. 4&5.



Context Type: post-hole

Sample Notes: grain

6 results

Group ID
130
Location

Group Type
DRAIN RELATED TO GP105 AND GP107
Group Notes
DRAIN RELATED TO GP105 AND GP107

http://tempuri/star/base#ehe0005.molas.fld_ind3d_lu.gp.rop95.130

903
Context Type
SP
Location

Notes
REFUSE TYPE FILL OF P/HOLE

http://tempuri/star/base#ehe0007.molas.fld_ind3d.context.rop95.903

Sample Notes
1 VICFA? FRAG
C.10 GRAINS IN 50%
Sample Measurements

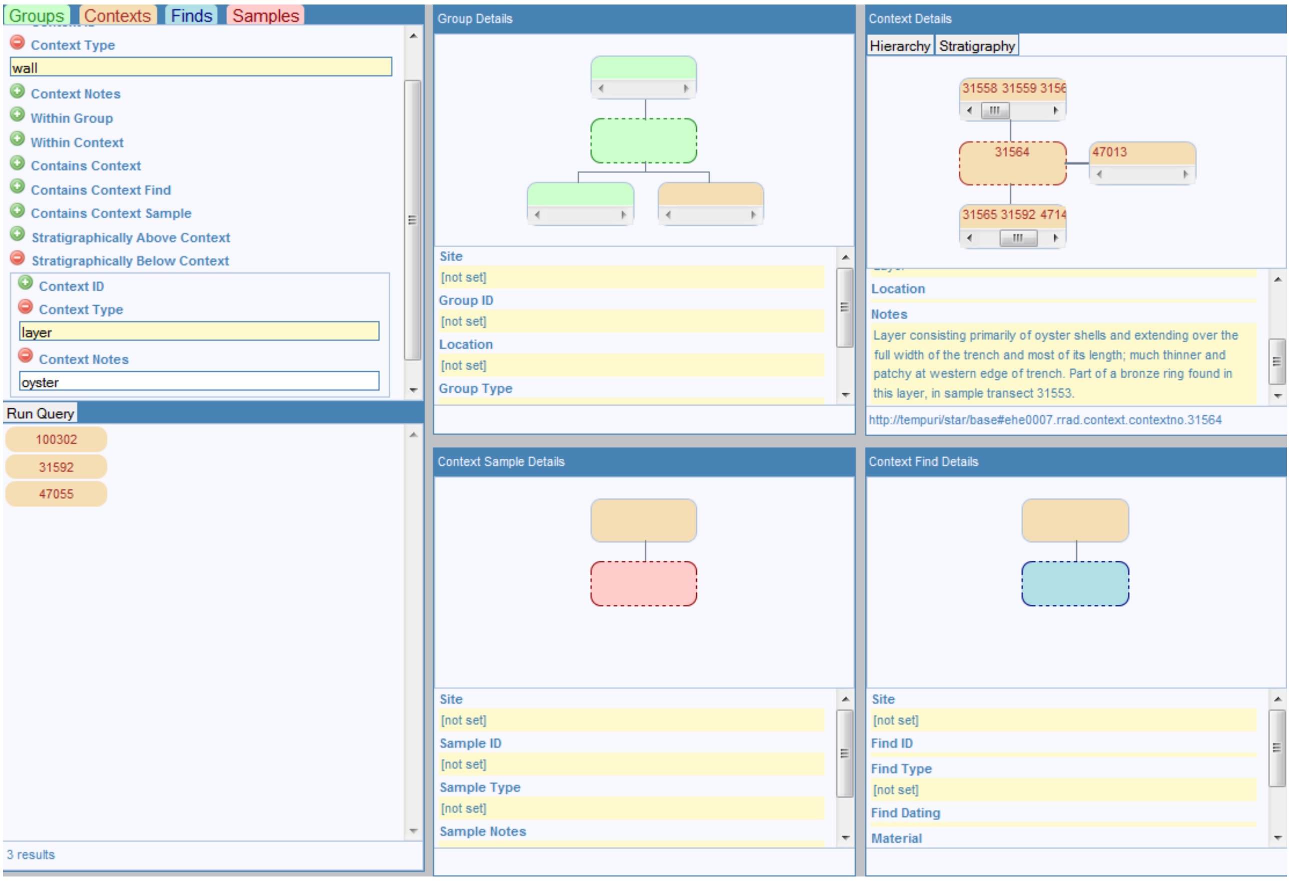
Type	Value	Unit
FLOT_VOL	300	-
PROC_VOL	10	-

http://tempuri

BICONICAL HEAD
Find Dating
410/1066, S
Material
COPP
Find Notes
BICONICAL HEAD

http://tempuri/star/base#ehe0009.molas.fnd_rf.acc_no.rop95.21

Figure 11: *Screen shot from Demonstrator showing all properties for given result*



Context Type: wall
Context Notes
Within Group
Within Context
Contains Context
Contains Context Find
Contains Context Sample
Stratigraphically Above Context
Stratigraphically Below Context
Context ID
Context Type: layer
Context Notes: oyster

100302
31592
47055

3 results

Site: [not set]
Group ID: [not set]
Location: [not set]
Group Type

Location
Notes: Layer consisting primarily of oyster shells and extending over the full width of the trench and most of its length; much thinner and patchy at western edge of trench. Part of a bronze ring found in this layer, in sample transect 31553.
http://tempuri/star/base#ehe0007.rrad.context.contextno.31564

Site: [not set]
Sample ID: [not set]
Sample Type: [not set]
Sample Notes

Site: [not set]
Find ID
Find Type: [not set]
Find Dating
Material

Figure 12: *Screen shot from Demonstrator showing semantic search and browsing of stratigraphic relationships (top right)*