



University of the
West of England

BRISTOL

Binding, C., May, K., Souza, R., Tudhope, D. and Vlachidis, A. (2013) Semantic technologies for archaeology resources: Results from the star project. In: Contreras, F., Farjas, M. and Melero, F., eds. (2013) *Fusion of Cultures: Proceedings of the 38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Granada, Spain, April 2010*. BAR Publishing, pp. 555-561. ISBN 9781407311081 Available from: <http://eprints.uwe.ac.uk/34842>

We recommend you cite the published version.

The publisher's URL is:

<http://www.barpublishing.com/caa-2010.html>

Refereed: Yes

This material has been published in CAA 2010: Fusion of Cultures. Proceedings of the 38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Granada, Spain, April 2010, edited by M. Contreras, F. Farjas and F.J. Melero, published by BAR Publishing. This version is free to view and download for personal use only. It cannot be reproduced in any form without permission of the publisher. The complete book is available at <http://www.barpublishing.com>.

Disclaimer

UWE has obtained warranties from all depositors as to their title in the material deposited and as to their right to deposit such material.

UWE makes no representation or warranties of commercial utility, title, or fitness for a particular purpose or any other warranty, express or implied in respect of any material deposited.

UWE makes no representation that the use of the materials will not infringe any patent, copyright, trademark or other property or proprietary rights.

UWE accepts no liability for any infringement of intellectual property rights in any material deposited but will remove such material from public view pending investigation in the event of an allegation of any such infringement.

PLEASE SCROLL DOWN FOR TEXT.

Semantic Technologies for Archaeology Resources: Results from the STAR project

C. Binding¹, K. May², R. Souza¹, D. Tudhope¹, A. Vlachidis¹

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 rela-

tionships. 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 Lon-

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.

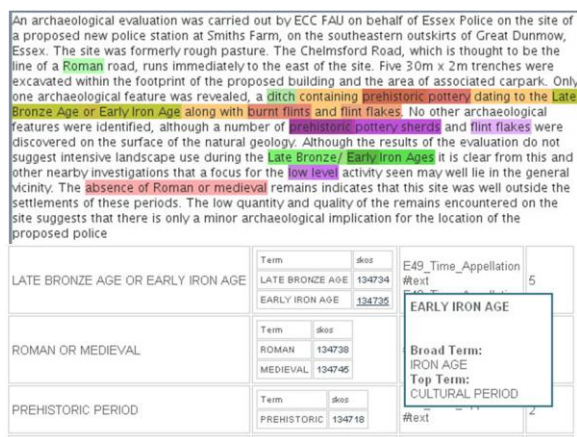


Figure 5: Example of annotation of time period entities

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.

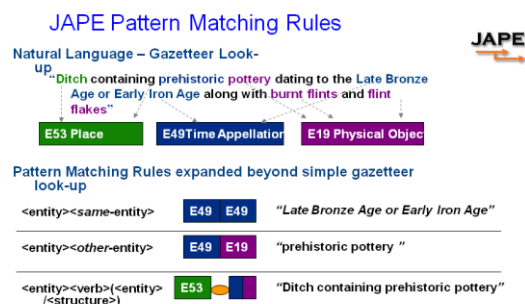


Figure 6: Example of CRM based Pattern Matching

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.

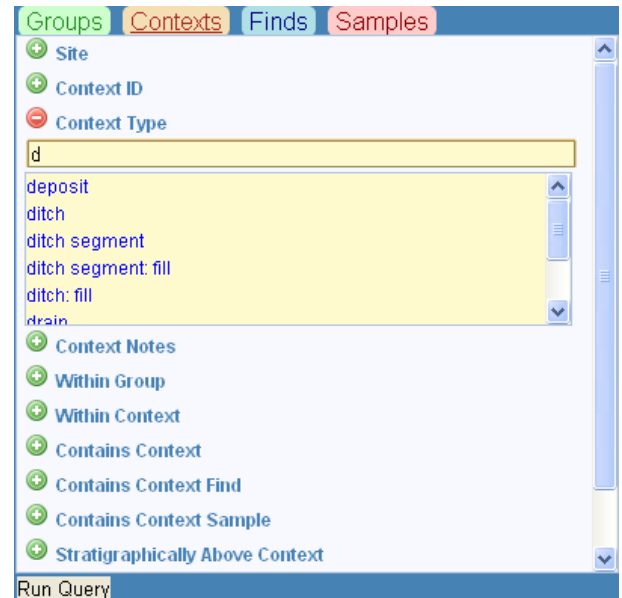


Figure 7: Screen shot from Demonstrator showing Query Builder for CRM-EH entities, with use of controlled types

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.

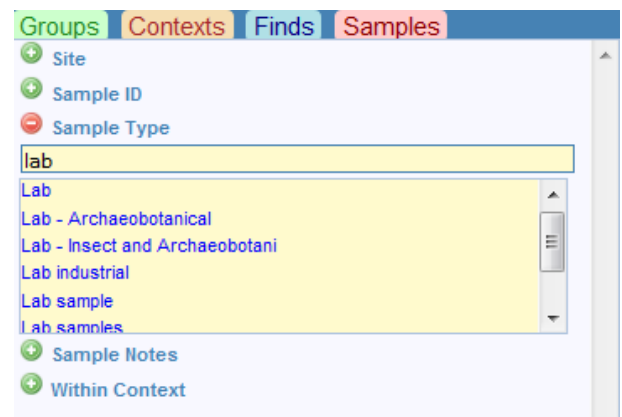


Figure 8: Screen shot from Demonstrator showing Query Builder and problem of uncontrolled types

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.

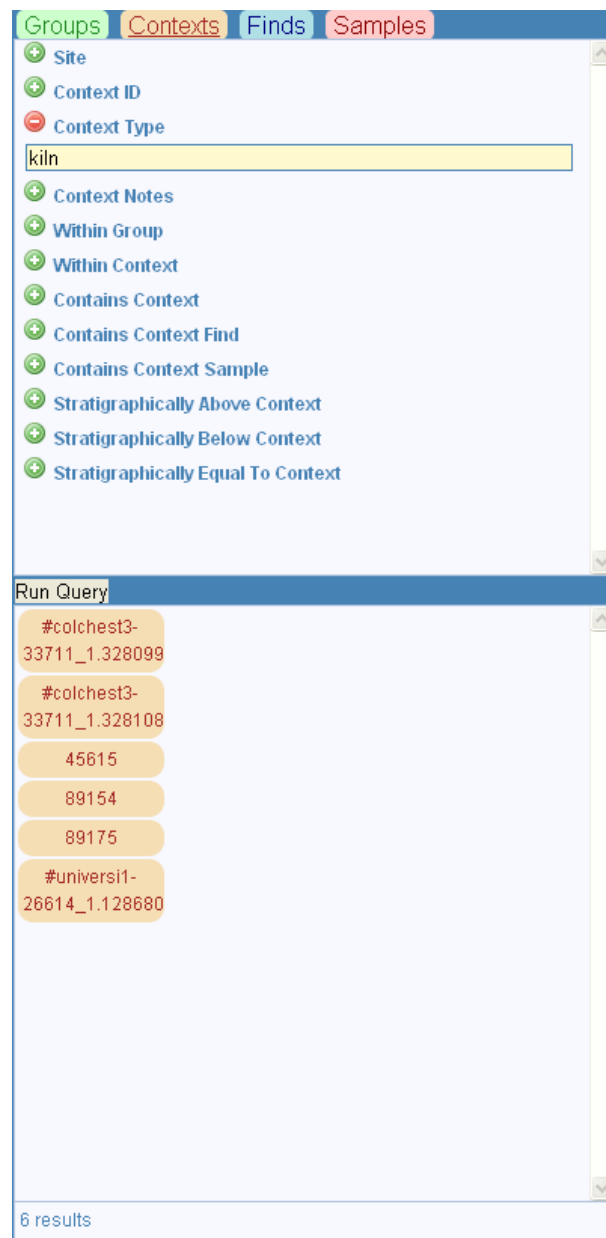


Figure 9: Screen shot from Demonstrator showing cross search of datasets and grey literature

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*.

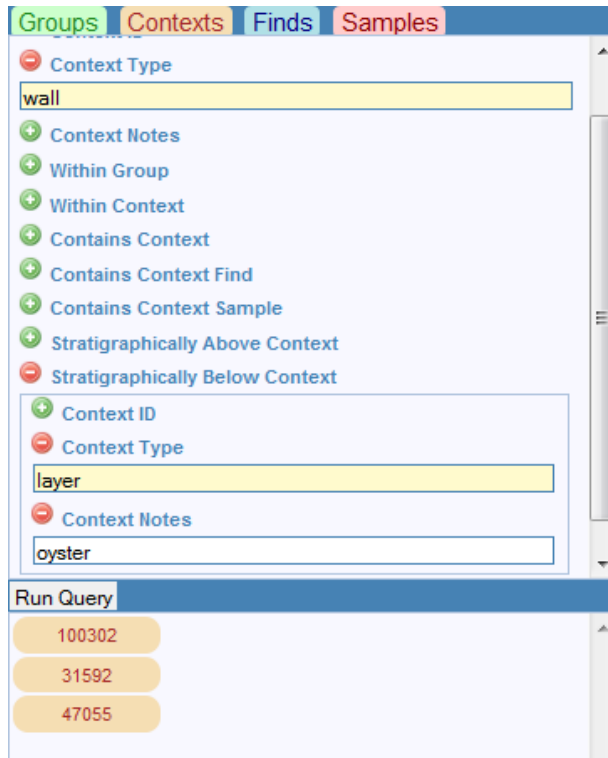


Figure 10: Screen shot from Demonstrator showing semantic search via stratigraphic relationships.

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*, 24(3), 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)

SKOS. Simple Knowledge Organization System, <http://www.w3.org/2004/02/skos> (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.

STAR PROJECT. Semantic Technologies for Archaeological Resources. <http://hypermedia.research.glam.ac.uk/kos/star> (accessed 25 January 2010)

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.

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.

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

The screenshot displays the STAR Demonstrator interface. On the left, a sidebar contains filters for 'Site', 'Context ID', 'Context Type' (with 'post-hole' selected), 'Context Notes', 'Within Group', 'Within Context', 'Contains Context', 'Contains Context Find', 'Contains Context Sample', 'Sample ID', 'Sample Type', 'Sample Notes' (with 'grain' selected), and 'Stratigraphically Above Context'. Below this is a 'Run Query' section showing a list of results: 665, 903, 970, 65966, 67349, and 68122. The main area is divided into several panels: 'Group Details' shows a hierarchy with group ID 130 and samples 535 537 and 903 904 970 971; 'Context Details' shows a hierarchy with context ID 903 and samples 21 220 221 758 7 and 161; 'Context Sample Details' shows sample ID 903 and find ID 161; 'Context Find Details' shows find ID 903 and sample ID 21. A table of 'Sample Measurements' is also visible, with columns for Type, Value, and Unit.

Type	Value	Unit
FLOT_VOL	300	-
PROC_VOL	10	-

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

The screenshot displays a software interface for semantic search and browsing of stratigraphic relationships. The interface is organized into several panels:

- Contexts Panel (Left):** Shows a search for "wall" and "oyster". It includes a "Run Query" section with results 100302, 31592, and 47055. Below the search results, it lists "3 results".
- Group Details Panel (Top Middle):** Displays a hierarchy of contexts. A green box is connected to two orange boxes below it.
- Context Details Panel (Top Right):** Shows a stratigraphic relationship between contexts. A hierarchy of orange boxes is shown, with context 31564 connected to 47013. Below the hierarchy, there is a "Notes" section with text: "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." and a URL: "http://tempuri/star/base#ehe0007.rrad.context.contextno.31564".
- Context Sample Details Panel (Bottom Middle):** Shows a sample of type "oyster".
- Context Find Details Panel (Bottom Right):** Shows a find of type "oyster".

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