CAA 2009 Williamsburg

Linking Archaeological Data

Leif Isaksen[1], Kirk Martinez[1], Nicholas Gibbins[1], Graeme Earl[2] and Simon Keay[2]

[1] School of Electronics and Computer Science, University of Southampton. UK.
[2] Department of Archaeology, University of Southampton. UK.

Abstract

The concept of Linked Data (http://linkeddata.org/)—information structured using a variety of public schemas and data sources—is beginning to take the Semantic Web out of the laboratory and into real-world applications. However, successful integration of legacy data sets requires the separation of the instances, terminologies and (frequently implicit) ontologies that constitute them so that each can be dealt with appropriately. This paper will discuss recent doctoral research seeking to provide practical solutions to this process and give some early examples of its potential benefit to archaeology.
The case study presented deals with a number of different databases pertaining to amphora and marble distribution that are being collated as part of the University of Southampton/British School at Rome "Roman Ports in the Western Mediterranean" Project. This data will be used to help understand the flow of ancient trade networks. In order to do this, a guided process, sufficiently intuitive for a wide range of archaeologists, is required to perform the mappings. Steps for mapping both amphora classification and excavation location have already been developed and temporal information will be introduced in the next phase of development.

Key words: *Semantic Web, Linked Data, RDF*

1 THE PORT NETWORKS PROJECT

The Roman Ports in the Western Mediterranean Project[1] (hereafter referred to as the Port Networks Project), directed by Prof. Simon Keay and Dr. Graeme Earl (British School at Rome/University of Southampton) is an investigation into the relationship of Portus—the main port of Rome in the Imperial era—to ports in the Western Mediterranean basin. The principal methodology involves looking at the co-presence of ceramics and marble at a range of key sites as a means of gauging fluctuating trans-Mediterranean connections during the Roman period. Source data comprise large quantities of published and unpublished harbor and shipwreck excavation databases from a variety of academic and research institutions in different countries.

Whilst the datasets all pertain to the same domain, they frequently employ mixed taxonomies and are heterogeneously structured. Normalization is rare, uncertainty frequent and variant spellings common. Different recording methodologies have also given rise to alternative quantification and dating strategies. In other words, it is a typical real-world mixed-context situation. As an international endeavor, requiring the synthesis of large quantities of data with heterogeneous format but restricted scope, it has proved an ideal opportunity to work through the issues specific to the archaeological community in deploying Semantic Web technologies.

The technological aspect of the project has been to find means by which to allow domain experts to translate their holdings into a common structure. In order to do so we are developing both a procedure and the associated technology to enable archaeological data providers to:

i) develop a common conceptual structure (domain ontology) capable of reflecting a level of inquiry relevant at an inter-site scale.

ii) cope with overlapping categorization systems

[1] http://www.bsr.ac.uk/BSR/sub_arch/BSR_Arch_05Roman.htm

iii) map local relational database schemas to the concepts represented in the domain ontology

iii) map locally used terminology with canonical (i.e. universal) identifiers

iv) export data to a centralized repository for use as a communal knowledgebase

v) export data in a format suitable for local hosting in order to promote distributed data connectivity.

2 APPROACH

The wide range of work undertaken in archaeology during the initial period of development in Semantic Web technologies has led to considerable diversity in their approaches. This makes general architectural decisions for the Port Networks Project difficult as there are still no well-established and well-documented methodologies for the full life-cycle of a Semantic Web project. Nonetheless, we can discern several key trends emerging, each with its own exemplars.

The first distinction is between processes which centralize data (MuseumFinland[2], Contexta/SR[3] and UBI-ERAT-LUPA[4], for example) and those which keep it distributed (MultiMediaN[5], eCHASE[6]). Whilst the former approach has a number of advantages in terms of simplicity, and was used frequently in early projects, there are a number of difficulties associated with it. Generally speaking, any methodology which seeks to integrate data from separate institutions which regularly update their information will have to implement an architecture that leaves them in full control of it. De-centralizing the data, however, requires a means by which to ensure that the same canonical URIs are used for mutual concepts, as well as guidance on how to make data easily discoverable by others.

We aim to take a twin-track approach. We start by providing a centralized vocabulary of canonical concept URIs, such as amphora types or ontology terms, hosted at http://archvocab.net. More extensive, and therefore more contentious, information about these concepts will be held in a publically available triplestore at http://archaeology.rkbexplorer.com. The reason for keeping these separate is to make it transparent to users that the canonical URI for a concept is separate from any statements about it – it simply provides a means for us to agree that we are talking about the same thing.

Once these stable and centralized resources for universal concepts have been set up, instance data, that is to say RDF produced about specific excavations, can then be dealt with more flexibly. For the purposes of the Port Networks Project we will establish a centralized triplestore in which project partners can store their own data, making analysis easier to coordinate. We also provide project partners with an XML/RDF version of the data which they will be able to host on their own websites. Should they choose to do so, it makes it openly available to by the wider research community and thus greatly improves the sustainability of the project.

The next consideration is whether instance data should be exported to an RDF store prior to querying and integration, or whether it should be mapped dynamically in real time. There are currently few, if any, cultural heritage applications that utilize the second approach but it is beginning to become more common elsewhere with DBpedia, a semantic service derived from Wikipedia, being a notable example. For systems that chiefly consist of large, centralized repositories, dynamic systems have the advantage of providing a 'live-update', so that information entered into a relational database does not need to be regularly exported, but they are dependent on a mapping server such as D2R

[2] Hyvönen et al., "MuseumFinland—Finnish museums on the semantic web."

[3] Astudillo, Inostroza, and Moncada, "Contexta/SR."

[4] Doerr, Schaller, and Theodoridou, "Integration of complementary archaeological sources."

[5] van Ossenbruggen et al., "Searching and Annotating Virtual Heritage Collections with Semantic-Web Techniques."

[6] Addis et al., "The eCHASE System for Cross-border Use of European Multimedia Cultural Heritage Content in Education and Publishing."

Server[7] which provide a SPARQL interface for querying. Dumping the data, on the other hand, requires either that users or an automated process export their data to the RDF store regularly. As this process can be resource intensive, it may also cause unwelcome performance issues at export time (although it is likely to improve performance all round at other times).

We have opted to go for the 'export' option for three reasons. Firstly, the source data is widely distributed and predominantly held in small, isolated, desktop systems. A dynamic approach would constantly be victim to downtime at any or all of these sources, leading to perpetually differing results. Most instance data is also fairly stable, with updates occurring over the course of an excavation season. Thus, there is not likely to be any need for a 'real-time' view of it. Finally, if a database is altered in such a way as to no longer be compatible with its RDF mapping, this can be identified at export time and the old data used until the issue has been resolved. With a live system, such problems are likely to interfere with the integrity of the output dataset as a whole.

Having established these general architectural principles, the next step is to set out the stages needed to implement them. They have been broken down into two phases of development, each of which was specifically intended to facilitate the conversion of diverse data holdings to a common structure:

1. Specification of a common ontology for both classificatory and excavation instance data;

2. Implementation of a workflow process that allows data holders to export their data as ontology-compliant RDF.

3 SPECIFYING A COMMON ONTOLOGY & VOCABULARY

Ontology specification

The first step is the design of an ontology capable of describing archaeological excavation data that pertains to marble and amphorae finds. This has been done in conjunction with a wide range of domain experts in order to ensure that key data necessary for a comprehensive inter-site summary can be described adequately by it, and that strategically useful research questions can be answered. Interestingly we have found that, due to the inherently incomplete nature of archaeological data, many of the minutiae and methodological differences between sites were agreed to be of minimal relevance for broad-scale analyses. Fig. 1 gives a (provisional) rendition of the ontology.

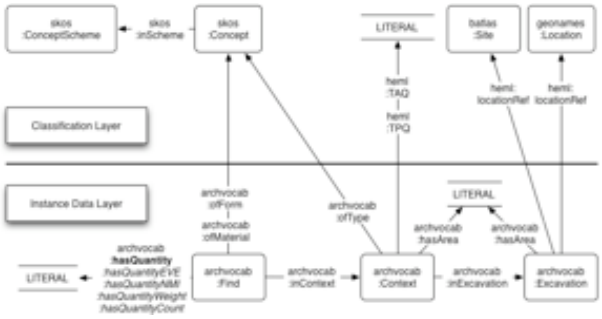

Figure 1. Excavation ontology diagram

The ontology is separated into 'Classification' and 'Instance data' layers so that independent datasets are only linked by canonical classificatory and singleton concepts. These canonical URIs provide a vocabulary of concepts that may be common to any instance data set: typology, location, period, form or material. It also makes deliberate reuse of vocabularies used elsewhere, including SKOS[8], and HEML[9]. The overall design is simple and stable enough for domain experts to easily interpret its relation to their own datasets. An RDFS description of the ontology is at http://archvocab.net/excavation/ontology.rdf.

[7] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/

[8] http://www.w3.org/2004/02/skos/core

[9] http://www.heml.org/rdf/2003-09-17/heml

Universal classification concepts

- ***Classes****: skos:Concept, skos:ConceptScheme, geonames:Location, batlas:Site*
- ***Properties:*** *skos:inScheme*

Instance data concepts

- ***Classes****: archvocab:Excavation, archvocab:Context, archvocab:Find*
- ***Properties:*** *archvocab:inContext, archvocab:inExcavation, heml:locationRef, archvocab:ofForm, archvocab:ofMaterial, archvocab:ofType, heml:TerminusAnteQuem, heml:TerminusPostQuem, archvocab:hasQuantity[EVE|NMI|Weight|Count]*

A notable observation during this design process was that, whilst ceramics experts will emphasize the shape of an amphorae over its fabric when classifying, marble specialists generally focus on the material first, with each using the term 'type' differently. As a result, we have created properties *archvocab:ofForm* and *archvocab:ofMaterial* so that both can be described in the same manner without ambiguity.

Classification service

The second step is creating the service that provides these canonical URIs for classification categories. Fortunately, because classification types form a reasonably small and stable body of information, it is feasible to define the URIs with a mixture of semi-automated processing and human intervention in a way that is not possible with instance data. Standard amphorae and marble typologies have been taken from a variety of digital and non-digital sources including the Archaeological Data Service Amphorae Database, and the Institut Català d'Arqueologia Clàssica Marble Catalogue. As mentioned previously, these URIs are provided at http://archvocab.net/amphora.

Archaeological typology data can be hard to compare as it frequently uses a mixture of different, overlapping typology series, using different terms for the same type. As not all of these overlapping types are agreed upon, it is not possible to compile them all into a single 'supertaxonomy'. In order to handle these separate schemata we are using the SKOS vocabulary. This not only allows us to describe separate concepts but also to map them across classification schemes. Each Form is related to its Type Series using the *skos:inScheme* predicate. Being uncontentious, this information is provided along with the URIs themselves. Thereafter, we can use the *skos:exactMatch*, *skos:narrowMatch* and *skos:broadMatch* predicates to identify types which are identical or similar to types in other schemata. Aggregation can then be done efficiently at query time and without needing to reclassify the instance data. The RDF to describe these relationships is also being created through a combination of structured querying and hand-correction, and, as it is open to archaeological debate, will be hosted in a separate triplestore at http://archaeology.rkbexplorer.com.

4 MAPPING

With a stable URI base for linking to, the second objective is to provide tools and a workflow by which data curators can map and export their holdings as RDF with minimal support. To do this, another two-step process has been developed. The Mapping Stage is a one-off activity in which a data curator generates an XML concordance between their local terminology and schema and the canonical property URIs described above. The Export Stage is then fully automatic and can be repeated as often as desired. Both processes are being prototyped as standalone applications written in the Java programming language. A website that provides the same functionality is likely to replace them in a future development phase.

The first tool takes data curators through a guided process by which they can map the local terms and database schema to the ontology and classification schemes described above. Using basic Natural Language Processing, it predicts probable mappings which the user can correct or extend using a Graphical User Interface. The results are stored as an XML configuration file specific to the dataset.

The process starts when the application, called a 'Data Inspector Wizard', is pointed at a digital resource such as a database or spreadsheet. A number of parameters are provided by the user, including logon details, the nature of the repository

(whether it contains amphora or marble finds), the relevant database Table or View and the desired namespace of the excavation data (Fig. 2). Ideally this should be a registered domain name owned by the data curator so that XML/RDF output can be hosted locally.

Figure 2. Data Inspector Wizard. Basic database information

The Wizard starts by matching table column names against the RDF triplestore at http://archaeology.rkbexplorer.com which contains linguistic terms associated with the key ontology concepts. It then creates a provisional mapping between them which the curator can modify if desired (Fig. 3). Column name mappings (but not data) are then returned to the triplestore so that they can be used to improve the predictive process over time.

The following stages form a modular workflow that is dependent upon the nature of the local repository and what elements of the ontology it has data for. The different elements of the workflow are currently under development but two individual stages, for mapping Amphora Form and Location concepts, are given as examples below.

Figure 3. Data Inspector Wizard. Ontology-to-column mapping.

Amphora Form Mapping

Local amphora terms are generally divided into up to four ordered elements, of which only the first is mandatory.

1. a Type Series name (e.g. 'Dressel' or an abbreviation such as 'dr.'),
2. a Type number (e.g. '20' but occasionally in roman numerals),
3. additional information (frequently the Material type or an alternative identification),
4. a marker of uncertainty (often a question mark).

The result is that the following entries could both refer to the same type and even come from the same database:

- Dr. 20?
- Dressel XX with tituli picti

The software breaks these local terms down into their component parts, assuming the first numeric value that it comes across to be the type number (if there is no number, it is assumed to be a Type Series with a single class). Because it is much easier to identify a Type once the Type Series is known, the Wizard aggregates all instances in the dataset with the same Type Series value and predicts the Type Series to which it refers. Once

again, these results are presented to the user for correction and new mappings are added to the classification repository to improve future guesses (Fig. 4).

It is interesting to note that, although mapping is reasonably low across all terms used in a dataset (generally below 50%), the proportion of records mapped correctly without user intervention is generally very high: often 90% or above. This is because deviation from an easily predictable norm is most frequently due to typological errors in un-normalized source data.

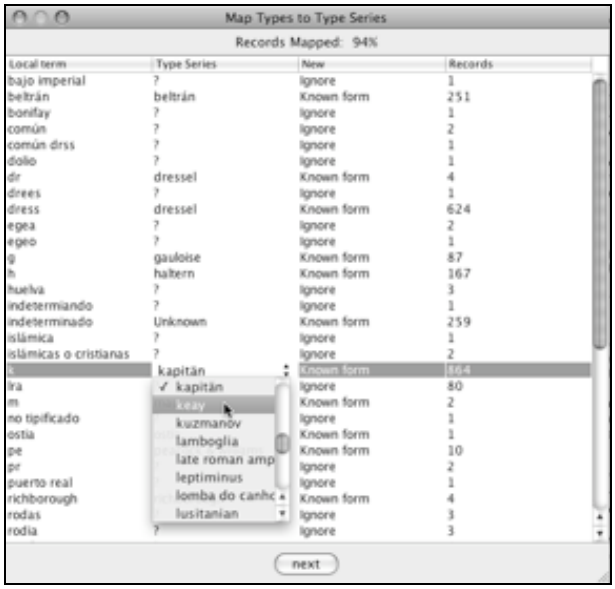

Figure 4. Data Inspector Wizard. Amphora type series mapping

With this done, the wizard uses the corrected Type Series mapping in order to predict the actual Form type. Results are usually quite accurate (>90%) as the estimation process chiefly relies on number-matching. Once again, the user is able to correct misassignments or expunge problematic instances (Fig. 5). It is worth noting that final output will frequently map multiple local terms to a single canonical term (for example, 'Dr. 20, and 'Dressel 20' may both refer to Dressel 20 amphorae in the same database), but the inverse is not true (for a given database, 'K. 2' might refer to Keay 2 amphorae or Kapitän 2 amphorae but will not refer to both).

Figure 5. Data Inspector Wizard. Amphora type mapping

Location

Location can be recorded in two fundamentally different ways: Spaces and Places. Spaces are discrete areas that we can describe using a number of formalisms, such as a National Grid Reference. The problem with them is that they are complex to process and give absolute boundaries to locations which are likely to have expanded and contracted, or even moved entirely, over time. Places are much closer to how we discuss locations in natural language. Someone can talk about a Place simply by referring to one of its toponyms, without ever having to know its precise geographical location or boundary. Although it is important not to confuse two places that have the same name, for inter-site analysis it is usually sufficient to know that a find came from, say, Seville rather than Barcelona. Knowledge of their specific geographical situation can be introduced later if necessary.

The GeoNames[10] service provides an online gazetteer of millions of places on earth and assigns each one a URI. Places can be searched for by name or category or by using a GoogleMaps-style interface. Early attempts at fully-automated location assignment using GeoNames proved to have an unacceptably high level of inaccuracy due

[10] http://www.geonames.org/

to the large number of topographic homonyms. From a computational perspective, the term 'Athens' is just as likely to refer to the American city as the Greek one. Fortunately, GeoNames also provides a webservice that can return potential matches based on a selection of criteria. This has been incorporated into the workflow so that a user only has to type in the ancient or modern placename they wish to use and a drop down list will present a limited range of options (Fig. 6). As GeoNames is also a community-based service, if the Place does not exist in the gazetteer, it is even possible to add it.

Figure 6. Data Inspector Wizard. Location mapping

Automated Mapping

On completion of the Data Inspector Wizard, an XML configuration file is generated. This contains all the mappings between the local dataset and the ontology and is sufficient for a fully-automated mapping process to be undertaken at any point in future.

A second Java tool, the Data Importer, automatically generates RDF from the database in conjunction with the configuration file. Minor database changes, such as new records using the same local classification terms, can be handled without any changes to the file being necessary. Structural changes, or the introduction of new local terms, can easily be managed by editing the configuration file within the Data Inspector Wizard or by hand (the XML file is human-readable). In either case, maintenance is minimal.

The RDF generated is in two forms. The basic form is an RDF/XML document which is immediately available to the data providers themselves. If they have provided a domain name to the Data Inspector Wizard which hosts their own website, then they can post the document just as they would a webpage. This makes it instantly accessible to other researchers who can then refer to the URIs for each context or find. For the benefit of the project, the RDF is additionally imported into a central triplestore, providing enhanced performance, security and querying functionality. Each dataset is also given an individual URI which is used to tag every triple. This makes updates simply a case of deleting all the triples in one such 'subgraph' and replacing it with another.

2 Conclusions & Further Work

The prototype tools have proven remarkably successful against a broad range of sample datasets from four different countries (UK, Spain, France, Italy). The most important achievement has been to enable domain experts to provide data derived in different contexts as ontology-compliant Linked Data extremely quickly and sustainably. Previous attempts to produce homogeneous RDF have generally required a lengthy and expensive mapping process against one or two large resources. We feel that making it possible for 'the long tail' of archaeological data is a vital task in the Linked Data revolution. We also draw some of the following conclusions:

We believe it is important for the Semantic Web not to be perceived as intending to replace or substitute conventional data archiving. Its principal advantage lies in the ability to ask broad questions across many small and diverse datasets, thus current development work ought to focus on data and processes which support that goal. We are especially interested to see whether aggregating data with uncertain levels of precision will enable us to tackle the problem of uncertainty in new

ways. We envisage the production of data point clouds and histograms which show probability distribution patterns inaccessible at datum level.

The use of a two-level ontology for classification and instance data greatly simplified the process. Open services which provide classification and singleton URIs, such as GeoNames, have made datalinking possible without instance data providers having to be aware of each other's existence. Naturally, developments which help to 'canonicalize' these classificatory or singleton concepts greatly aid the process. We would like to see a service similar to GeoNames for temporal periods.

We found that predictive mapping and a multi-step classification workflow greatly increased the speed and ease with which mapping could be undertaken of large, un-normalized datasets. It was also helpful to show how often terms are used in order to pick out probable anomalies. These are vital benefits in making mappings between relational and RDF datastores possible for non-IT professionals.

Acknowledgements

We gratefully acknowledge the assistance of all the Port Networks Project partners in developing the ontology and software described. We would also like to thank Dr. Hugh Glaser for kindly making available the http://archaeology.rkbexplorer.com domain available for this work.

Bibliography

Addis, Matthew, Shahbaz Hafeez, Daniel Prideaux, Richard Lowe, Paul Lewis, Kirk Martinez, and Patrick Sinclair. "The eCHASE System for Cross-border Use of European Multimedia Cultural Heritage Content in Education and Publishing." In *AXMEDIS 2006: 2nd International Conference on Automated Production of Cross Media Content for Multi-channel Distribution*. Leeds, UK, 2006.

Astudillo, Hernan, Pablo Inostroza, and Claudia Andrea López Moncada. "Contexta/SR: A multi-institutional semantic integration platform." In *Museums and the Web 2008*. Montreal, Canada, 2008.

Doerr, Martin, Kurt Schaller, and Maria Theodoridou. "Integration of complementary archaeological sources." In *Proceedings of Computer Applications and Quantitative Methods in Archaeology Conference 2004*. Prato, Italy, 2004.

Hyvönen, Eero, Eetu Mäkelä, Mirva Salminen, Arttu Valo, Kim Viljanen, Samppa Saarela, Miikka Junnila, and Suvi Kettula. "MuseumFinland—Finnish museums on the semantic web." *Web Semantics: Science, Services and Agents on the World Wide Web 3, no. 2-3* (October 2005): 224-241.

van Ossenbruggen, Jacco, Alia Amin, Lynda Hardman, Michiel Hildebrand, Mark van Assem, Borys Omelayenko, Guus Schreiber, et al. "Searching and Annotating Virtual Heritage Collections with Semantic-Web Techniques." In *Museums and the Web 2007*. San Fransisco, USA, 2007.