Barriers and opportunities for Linked Open Data use in archaeology and cultural heritage

Keith May, Ceri Binding, Doug Tudhope

Zusammenfassung – Archäologen wie auch Einrichtungen des Kulturerbes und der Denkmalpflege sind gegenwärtig zunehmend bemüht, ihre Datenbestände, die bislang nur einem kleinen Kreis von Spezialisten zugänglich waren, auch für eine breite akademische Forschung und die interessierte Öffentlichkeit zu öffnen. Um dieses Ziel möglichst effektiv zu erreichen, werden vernetzte Infrastrukturen und Softwaretools benötigt, die Nutzer bei der Suche und Auswertung von gefundenen Begriffen und Konzepten unterstützen – insbesondere auch deshalb, weil diese in heterogenen Datensammlungen unterschiedlich verwendet werden. Unterschiedliche Personen und Fachdisziplinen können für dasselbe Konzept verschiedene Wörter benutzen oder sie arbeiten mit voneinander abweichenden Vorstellungen. Diese terminologisch-konzeptuelle Vielfalt stellt unausweichlich eine Hürde dar, die den Datenzugriff für Forscher oder die Allgemeinheit erheblich erschwert.

Ein praktischer Ansatz zur Lösung dieses Problems, der in den Projekten STAR, STELLAR und SENESCHAL verfolgt wurde, beruht auf der Anwendung des W3C SKOS Standards zur Integration von kontrollierten Vokabularen und dem Referenzmodell CIDOC-CRM (siehe Referenzen). Als Ergebnis liegen nun mehrere nationale Vokabularien zum Kulturerbe als SKOS-basierte Versionen vor, die über die Webseite HeritageData.org (s. Ref.) aufgerufen werden können.Der Beitrag diskutiert einige Barrieren und Herausforderungen, die während der Entwicklung einer modernen Linked Open Data Ressource auftraten, und zeigt die Potentiale für künftige Entwicklungen in diesem Bereich auf.

Schlüsselwörter – archäologische Daten, archäologische Erfassungssysteme, Linked Open Data, LOD, SKOS, semantische Technologien

Abstract – Archaeologists, along with cultural heritage and memory institutions generally are seeking to open up databases, and repositories of digitised items, previously confined to specialists, for a wider academic and general audience. But to do so most effectively requires joined up infrastructures and tools to help formulate and refine searches and navigate through the information space of concepts used to describe different collections. Different people and domains use different words for the same concept or may employ slightly different concepts and this 'vocabulary problem' is inevitably a barrier to broadening scholarly, let alone wider access.

Practical work in tackling such issues has used the W3C SKOS standard for incorporating controlled terminologies along with the CIDOC CRM in the STAR, STELLAR and SENESCHAL projects (see ref.), leading to SKOS based versions of national cultural heritage domain controlled vocabularies and the publishing of these as Linked Open Data via the HeritageData.org web site (see ref.). This paper will discuss some of the barriers and issues encountered while developing some current 'state of the art' Linked Open Data resources for cultural heritage and consider important opportunities for the development of such LOD resources for the future.

Key words – archaeological data, archaeological recording systems, Heritage Data, Linked Open Data, LOD, SKOS, semantic technologies

Introduction and background to Linked Open Data (LOD) research

The background to the work discussed in this paper has been a series of projects undertaken over the last ten years since 2004. The first research work began as part of a larger project called Revelation (May, Attewell, Cripps et al., 2004) to consider and plan for the development of a new recording system for the English Heritage archaeological research teams. One primary requirement for this project was to investigate methods and technologies for developing a system with much better integration of existing data sets which were created as part of the broad archaeological research associated with a fieldwork project (e.g. geophysics , excavation data, survey data, post-excavation analysis data, environmental analysis, finds objects analysis, etc.).

A deliberate attempt was made early on in the planning to avoid simply re-inventing the existing relational database models that had led to the continued proliferation of separate databases for each fieldwork project undertaken, along with associated analysis and post-excavation work. It was decided to develop a high level ontological model of the main archaeological processes and concepts involved in the creation of data associated with a fieldwork project. This work led to the creation of an ontological model based upon the existing CIDOC-CRM (ISO 21127: 2006) standard for broader Cultural Heritage material (Crofts, et al. 2008) but which covered more domain specific (i.e. archaeological) concepts such as Archaeological Sites, Excavation events, Stratigraphic Contexts, Finds Objects, Sampling, etc. These domain specific extensions that matched English Heritage archaeological practices and processes became known in short-hand as the CRM-EH. These domain specific extensions and scope note definitions were designed to represent specific archaeological entities that mapped to CIDOC CRM concepts. So for example an archaeological finds object was modelled in CRM-EH as a ContextFind EHE0009 which mapped to E19 Physical Object in the CIDOC CRM. The CRM-

Received: 13. Nov. 2014
accepted: 27. Dez. 2014
published online: 4. Feb. 2015

EH archaeological extensions focus on common 'core' concepts of our archaeological processes and the relationships between those core concepts (Cripps, Greenhalgh, Fellows et al., 2004).

An advantage of using such modelling is that it can enable the use of a range of technologies known generally as 'semantic technologies'. In particular these technologies enable the incorporation of structured vocabularies for more refined indexing of datasets and provide tools to help formulate and refine searches and 'navigate through the information space of concepts used to describe archaeological data'. More technical details of these semantic technologies as applied to archaeological data and reports are published elsewhere as the outputs of the STAR and STELLAR projects (Binding, Tudhope, & May 2008; Binding, May, Souza et al., 2010).

Some Barriers Encountered

This paper will discuss some of the barriers encountered and addressed using the semantic technologies for archaeological resources and in particular highlights more recent work as part of the ARIADNE project (see ref.). ARIADNE is an EU funded FP7 project that aims to better integrate and cross-reference existing archaeological research data infrastructures to improve the usability of the various distributed digital datasets with new and powerful technologies and help researchers make such resources an integral component of their archaeological research methodology. Within ARIADNE there is specific development work to look at how various approaches adopted in STAR and STELLAR could be further developed to incorporate a wider European perspective on archaeological recording methodologies. The first section of this paper will focus on where those barriers are more about the human and social aspects of archaeological recording rather than semantic web technological issues, while the following sections will go on to discuss more recent work using semantic technologies for overcoming some of those barriers and publishing archaeological terminologies as Linked Open Data (LOD). Linked Open Data uses the existing technologies of the World Wide Web. But more than linking together pages and documents to browse on the web, LOD is a method of publishing detailed data as a series of inter-linked and inter-related data statements. LOD makes the actual data items held within databases, or published within text based resources, more searchable and enables more complex reasoning about the semantic relationships between those different data statements.

Different recording systems - the UK experience

One of the major issues that arise in attempting to work with archaeological data from different projects that have been recorded by different archaeological organisations is that they often have differing recording systems and often use differing terminologies to make those records. These recording systems may share considerable commonalities in their general structure, but more often than not the actual terms and fields used in the databases and systems that hold the data can vary quite considerably. This variance in the way records are held can be an immediate barrier to making searches or analyses across the data contained in those different records.

In the UK the situation is helped by the fact that there is generally one main recording methodology most commonly used which is usually referred to as 'single context recording'. This method is based on the principle that each individual unit of stratigraphy – usually referred to as an archaeological 'Context' – is given a separate number and recorded separately, often with its own single record sheet of descriptive data and a single drawing in plan.

However although most archaeological organisations in the UK will use some recognisable version of this methodology, that does not necessarily make their resulting digital data sets so easy to integrate. More often than not each organisation has its own computer database system to hold their data, often running on different software platforms. Quite often different projects carried out even on geographically adjacent sites by different organisations may be recorded on quite differing database systems. Also the pick-lists of terminologies used within those differing database systems may be 'controlled' to varying degrees and are not usually cross-referenced to any standardised vocabularies used by other organisations.

Even where the same organisation uses relatively common database software such as MS Access, the changes in versions over time can make data from a project that was recorded a few years ago, no longer easily integrated with the current version of a newer database system.

Different recording systems across Europe and beyond

Recent work as part of the ARIADNE FP7 project has allowed some wider comparison of archaeological recording methodologies across Europe and the Mediterranean area. While variations on the single context recording system used in the UK are widely used elsewhere in Europe, there is still the same issue on a magnified scale that most of the different archaeological organisations have their own database systems with many variations of file structure and software platforms.

In addition these issues for cross-search and interoperability are exacerbated by the fact that different countries may also use quite differing recording methodologies. In many parts of Germany and the Netherlands a system known commonly as the '*Planum*' system is used (based upon the excavation of a series of regular '*Schnitt*'), which relies on excavating and recording (plan drawing) recognised features at a series of spatially defined levels or horizons of excavation (e.g. a new plan made after excavating every 10 cm in depth). In some regions, especially around the Mediterranean, a system of Locus and Basket numbers which derives from the 'Wheeler box excavation methodology', is used to distinguish and record the units of excavation (Locus) that are excavated along with the different 'baskets' of soil/deposits containing objects (finds) from that 'Locus'. At some projects, such as Çatalhöyük in Turkey, a version of single context recording is used, although the recording system records at Çatalhöyük refer to single units of stratigraphy as "Units" rather than "Contexts" (Hodder 2000).

Although the ARIADNE project has only been working in Europe, the understanding is that similar issues of differing recording systems and methodologies are also common in North America and elsewhere (Pavel 2010).

Different excavation methodologies bring differing documentation with differing vocabularies

"An archaeological deposit is a three dimensional artefact, only seen once, and never seen whole" (Carver 2009, p. 123). Carver's key message about archaeological methodologies is that it is part of the role of the excavator to assess and adopt the appropriate methodology, and thereby recording system, to tackle the particular archaeological remains that they encounter. Until the currently excavated archaeological deposit is fully removed it is never *entirely* certain what its full extents and identity will be, by which point it is no longer extant.

The main point to this very brief outline of differing methodologies is that different recording methodologies bring further variations in records and documentation of what is investigated. These variations do not usually create problems within individual projects or organisations as they are generally able to adapt their own systems to compare data recorded by variations in methodology within single projects. For example English Heritage recording forms are primarily based upon single context recording, but in places it is acknowledged that some stratigraphic deposits (e.g. deep well fills that can only be excavated safely by supporting or removing the well sides) may need to be dug in fixed levels as 'spits' (comparable to the German 'Schnitt' mentioned above).

However the issue becomes more considerable if we want to try and compare data using online and semantic technologies from a range of sites where we do not necessarily know the details of what methodology was used for the excavation. This situation is further compounded by the fact that the different records from multiple organisations can be made on a plethora of differing recording sheets/systems (**Fig. 1**). Catalin Pavel has made a very useful analysis of different recording systems and record sheets from Europe and America (Pavel 2010) which gives just a flavour of the degree of variation that can be introduced by variations to the recording methodology.

Problems of semi-controlled vocabularies

We have plenty of controlled vocabularies in the cultural heritage domain, but there are tensions when using them in the field between wanting to be as descriptive as possible about what is being recorded, versus wanting to have controlled indexing to make data retrieval as accurate as possible. In practice during fieldwork data entry is often not restricted to controlled vocabularies and at a practical level, while excavating, often only hand written records are made so there is even more scope for spelling errors or other mistakes to occur.

Quite often 'semi-controlled vocabularies' get adopted which seem to represent a useful compromise somewhere between allowing the excavator to be descriptive while still using a more closely defined set of terms. However for data retrieval this proliferation of terms is a major problem when trying to search for specific types of records. The problem comes from trying

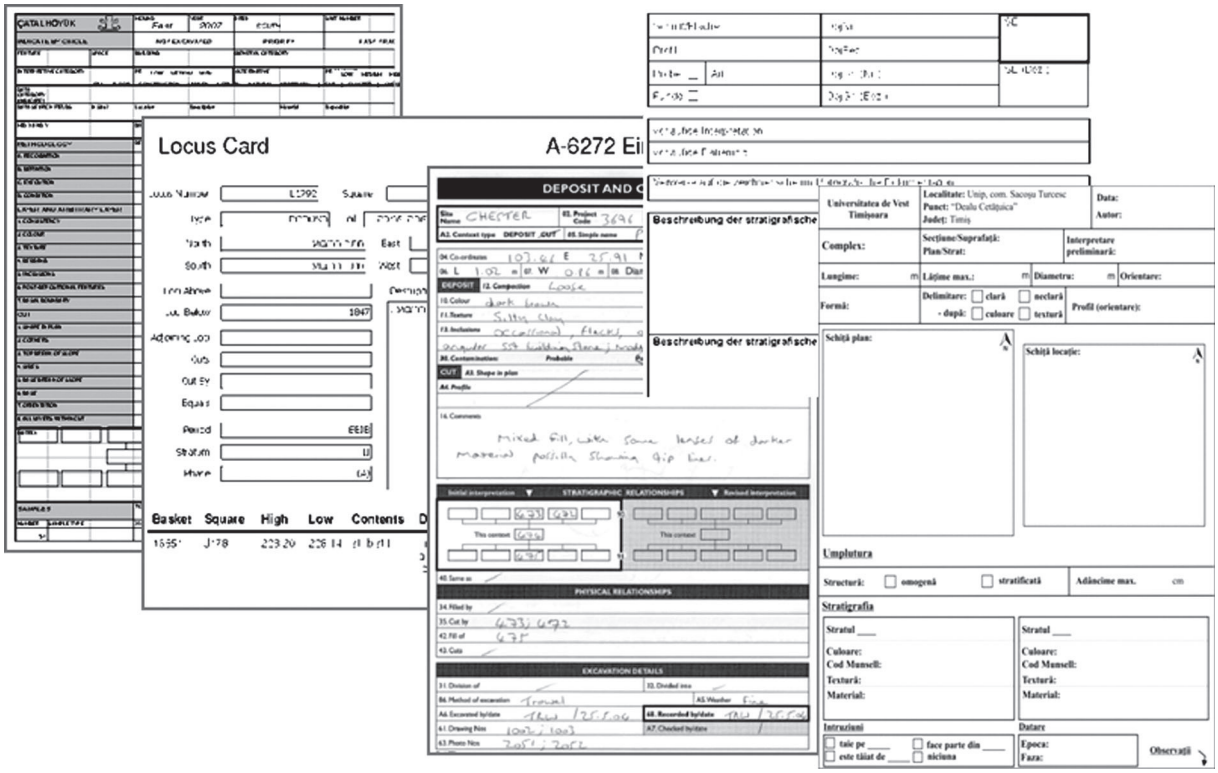

Fig. 1 Examples of a variety of archaeological recording sheets from Europe.

to achieve two different things with one single record or field in a database. Instead recorders should be using free text to describe what they want to record and then index that record using controlled index fields with controlled vocabulary terms.

Unlocking Some Barriers

This section of the paper will outline some of the work that has been undertaken to try and overcome some of the barriers presented in the previous section.

Archaeological Terms represented as Concepts with Relationships

Words are ambiguous, and when using them for metadata keyword indexing it can often result in the return of diverse, inaccurate and often conflicting search results. The problem often originates from people using the same words to refer to different underlying conceptual meanings.

For example, in Scotland the Royal Commission on the Ancient and Historical Monuments of Scotland (RCAHMS) define a "Tenement" as "*a large building containing a number of rooms or flats, access to which is usually gained via a common stairway*". In England the term "Tenement" is defined as "*a parcel of land*". A search for "tenement" using just the text will not disambiguate between search results from both Scottish and English sources. But the meanings of the concepts which are referred to in the Scottish and English thesauri scope notes are clearly two different concepts. If we can refer to the two distinct concepts of "Tenement" with two different *concept identifiers* and we supplement our data where relevant with those different concept identifiers, it should become clear (especially to computers) which meaning we are using; either Scottish or English, and we can also express the differences in our search criteria to search engines more accurately.

So one solution to the problem of ambiguous use of words in data fields can be to use concept-based controlled vocabularies, which enable the computer search systems to disambiguate between the same digital text strings that may carry two separate conceptual meanings.

SKOS – Simple Knowledge Organisation System

One significant approach to enabling the wider use of concept based terminologies has been to develop methods and tools to try and better enable archaeologists and those working with archaeological data to express the terms that they use to record and index their data in more consistent and accurate ways. A key approach to this has been to make such controlled vocabularies available online in a form that can be used consistently and in ways that can reduce some of the ambiguities by making use of the W3C standard for controlled vocabularies known as Simple Knowledge Organisation System (SKOS) format (Miles, Matthews & Wilson 2005). The SKOS format enables existing structured vocabularies such as thesauri to be converted into Linked Data thereby enabling use of the semantics of the relationships between the various terms in the vocabularies. In particular the use of SKOS has enabled many of the existing standardised national heritage terminologies used in the UK to be transformed into a concept based format for cross-reference searching of data sets using semantic technologies that don't just search using keyword 'text string matching', but also make use of the relationships between the concepts that are inherent (but not always explicit) in the various controlled vocabularies.

The methodology for the systematic conversion of standardised heritage terminologies to SKOS format was utilized and refined as part of an Arts and Humanities Research Council (AHRC) funded project called STELLAR (see ref.) which developed a template tool and associated guidance documentation to enable non-specialists to convert their controlled terminologies to the SKOS format.

Converting the terminologies to SKOS format is a very useful step in making them more cross-searchable by computers, but it does not necessarily make the terms accessible in an open way that people can use online. Further work under a project called SENESCHAL has built upon the SKOS conversion tools and made resulting national standard controlled vocabularies available online as Linked Open Data (LOD).

Vocabularies as Linked Open Data

A number of standard vocabularies have been developed by national heritage agencies and are used by many historic environment organisations and practitioners across the UK, but until now we have seen a number of issues about making them easily and readily available for use online. Often the size of the vocabularies has meant difficulties in including them as reference terms in online data entry forms. Creating versions of the vocabularies as Linked Open Data is a way of addressing that problem and means the vocabularies can be incorporated more easily as drop-down lists in online forms thus greatly aiding the search, selection, speed and accurate entry of controlled terms without the errors associated with hand-typed free text data inputting.

The creation of Linked Open Data has been covered by a number of authors, not least Tim Berners-Lee (Bizer, Heath & Berners-Lee 2009). A fundamental step is that the data is in Resource Description Format (RDF) via the web. This is one aspect that the SKOS conversion process provides, so that the concept identifiers for the terms and the relationships between them are expressed as a series of simple triple statements in the form:

Subject <predicate> Object
such as
"skos:Concept" <skos: inScheme> "skos: ConceptScheme"
or e.g.
TENEMENT <is in scheme> "Monument Type (EH)"

Persistent globally unique identifiers (URIs) for every concept

To enable the linked data to be used consistently online a key requirement is that the reference used for each data item is a persistent URI. This simply means that the online hyperlink used to reference the data item always resolves to the same persistent identifier for that data item (you can think of it as a unique online name for any item). When we convert our controlled terminologies to SKOS we are creating an identifier for each concept in the vocabulary and it is these concept identifiers (amongst other items in the schema) which get represented as persistent URIs when they are made available as Linked Open Data online. In the case of the SENESCHAL project we implemented an organisation neutral base URI in the format http://purl.org/heritagedata/ – which then becomes the base URI for all scheme and concept identifiers that it references. The organisation neutral choice took some deliberations but it is grounded on the general guidance provided by UK government information principles (Cabinet Office 2011) to try and avoid any use of website domain names (e. g. www.english-heritage.org.uk) that are likely to change over time. Given that English Heritage and RCAHMS are both

in the process of changing organisation names, structures and websites subsequently, it has already proved a sound strategy for maintaining a persistent URI.

Open Access

To enable users to search and browse the vocabularies online we have set up a website at www.heritagedata.org/ with browsable HTML 'landing pages' where we can also give guidance on use of the vocabularies and provide other related tools which have been made available (see sections on widgets and web services below) . The website acts as the landing page for human users, rather than just making the 'raw' Linked Data URIs available in the SKOS format which, although readable by humans, is primarily intended for direct interpretation by computers.

The decision has also been taken to make the vocabularies available under an Open Access CC-By licence (sometimes referred to as an attribution licence) so that widest possible re-use of the vocabularies can be made as long as they are attributed to the originating agency. This means they could be open to commercial re-use, but more importantly they are accessible for the widest range of re-use in other applications. Again we followed Government advice to public sector agencies in the UK which endorses this approach and indeed the Scottish thesauri are actually licenced under an Open Government licence which is equivalent to CC-By.

The attribution is seen as significant in two respects. Firstly it seems good practice to acknowledge the work that has been put into these resources, but perhaps more significantly we felt in an Open Access environment it helps provide some authority and validity to the origination of the Linked Open Data items. It was felt that the community of users of HeritageData.org would be doing so because they wanted to be using a recognised and verified national standard and therefore attribution should be included.

Web services to facilitate concept searching, browsing, suggestion, and validation

To enable use of the vocabularies by others in their own applications a set of web services have been made available (these are explained in more detail at http://www.heritagedata.org/blog/services/. For programmers, the web services consist of a series of REST URI calls with a number of associated parameters which return the vocabulary data strings in the form of a JSON structured string. The web services are designed so they can be easily used in all commonly available browser based applications.

Tools to use controlled vocabularies: 'widget' user interface controls

One factor that is important to making the terminologies more (re)usable is to provide them in a form that can be easily and readily used by others within their own web pages and data entry forms. The approach taken to this is to provide 'widgets' which are a suite of predefined user interface controls that can be inserted into a web page and dynamically obtain the required vocabulary information using the available web services. The widgets function in any current browser on PC, Mac, smartphone, tablet, console. The controls provide vocabulary navigation, search and selection functionality that can be embedded directly within other peoples own web pages. More information on their use is available from the HeritageData site (http://www.heritagedata.org/blog/term-suggestion-in-a-widget/), including a set of associated demonstration pages that show how to configure and use each widget control, and how to combine them to create functionally rich user interfaces. The widget source code is also available as Open Source from https://github.com/cbinding/SENESCHAL, under CC-BY licence.

Downloadable data files and listings

Although the main innovation in creating the Heritage Data website was to provide the vocabularies in a Linked Open Data format, it was also decided to make several downloadable versions available, primarily to aid people who might be considering using any of the thesauri to get an 'overview' of a whole thesauri or scheme, and to help with any considerations of cross vocabulary alignment with other thesauri. Each complete thesaurus is therefore also available as a download in SKOS (RDF), as an alphabetical listing (PDF) and in an hierarchical structure (PDF) the last two of which are similar to the more conventional printed thesauri outputs.

Further Opportunities

Thesaurus to thesaurus alignment

The conversion of the terminologies into SKOS RDF/XML format has considerable potential for enabling alignment of terms that have the same conceptual meaning but which derive from different online vocabularies. This would then

enable cross searches to be based upon the semantic meanings of the terms involved rather than the current exact text string matching. However such alignment does require a concerted initial effort on behalf of vocabulary owners to make the relationships between terms explicit (at least in SKOS format), and thus enable the consequent automated cross-referencing by computers to work.

Legacy data to thesaurus alignment

Another possible opportunity is to carry out alignment between data items already contained in existing data sets to align the data with the newly available SKOS Linked Data vocabularies. This alignment of 'legacy' data is likely to be a more intricate operation, but there is some potential to make a semi-automated bulk alignment process.

One approach taken is to adopt an algorithm that calculates the degree of matching between terms. In the SENESCHAL project the 'Levenshtein edit distance' (Levenshtein 1966) algorithm has been used to explore the feasibility of bulk alignment approaches. The Levenshtein algorithm measures the optimal number of character edits required to change the content of one string of letters into another.

The bulk alignment process makes a comparison between the selected term and all terms from the specified thesaurus that you are trying to align with, to obtain the closest textual match. Because the nature of the algorithm is to *always* find some degree of match it is necessary to introduce suitable thresholds which can flag up and supress low scoring matches. Also, as can be seen in the example (**Fig. 2**), there can often be quite a high degree of matching between terms which have just a negation prefix (e.g. organic / inorganic) so an element of human intervention in checking the matches is certainly still required.

Concept	Best Match	Score
CANDLEHOLDER	CANDLE HOLDER	92%
MANUFACTURING AND PROCESSING	MANUFACTURE AND PROCESSING	89%
CRUSIE	CRUSE	83%
INORGANIC MATERIAL	ORGANIC MATERIAL	88%
PERSONAL ADORNMENT	PERSONAL ORNAMENT	83%
BALANCE	BALANCE	100%

Fig. 2 Thesaurus to Thesaurus alignment: RCAHMS objects to FISH objects

```
<rdf:RDF xmlns:rdf="http://www.
w3.org/1999/02/22-rdf-syntax-ns#" xmlns:
ns0="http://creativecommons.org/ns#"
xmlns:ns1="http://www.w3.org/2004/02/
skos/core#"
xmlns:ns2="http://purl.org/dc/terms/">
<rdf:Description rdf:about="http://purl.org/
heritagedata/schemes/1/concepts/447">
<ns1:prefLabel xml:
lang="en">SCHOOLHOUSE</ns1:
prefLabel>
<ns1:prefLabel xml:lang="gd">TAIGH-
SGOILE</ns1:prefLabel>
<ns1:broader rdf:resource="http://purl.org/
heritagedata/schemes/1/concepts/422"/>
<ns1:broader rdf:resource="http://purl.org/
heritagedata/schemes/1/concepts/537"/>
<ns1:scopeNote xml:lang="en">A dwelling
attached to a school, usually occupied by a
school teacher.</ns1:scopeNote>
<ns1:scopeNote xml:lang="gd">Àite-
còmhnaidh a tha co-cheangailte ri sgoil, mar
is trice bhiodh tidsear na sgoile a' fuireach
ann.</ns1:scopeNote>
```

Fig. 3 RDF snippet including Scottish Gaelic (gd) along with English language (en) terms.

Multi-lingual potential: Schoolhouse example in English & Gaelic

The conversion to SKOS and RDF has also made it possible to incorporate different language versions (labels) of the terms together in the Linked Data concept schema of the thesauri. The main examples of this so far on HeritageData are in the RCAHMS Scottish vocabularies where both English and Scots Gaelic preferred labels and scope notes are provided. An example of this using the term for Schoolhouse (English) and Taigh-sgoile (Scottish Gaelic) can be seen in the snippet of RDF included below (**Fig. 3**):

Other languages could equally be provided if translations are available simply by including the appropriate ISO language tags for the language concerned in the RDF (e.g. for Scottish Gaelic xml:lang="gd"). Clearly there is potential in this approach for cross-referencing of terminologies in different languages using Linked Open Data technologies. Making mappings between different language thesauri is part of the ARIADNE infrastructure work which should also offer opportunities for other providers to make their vocabularies available as Linked Data.

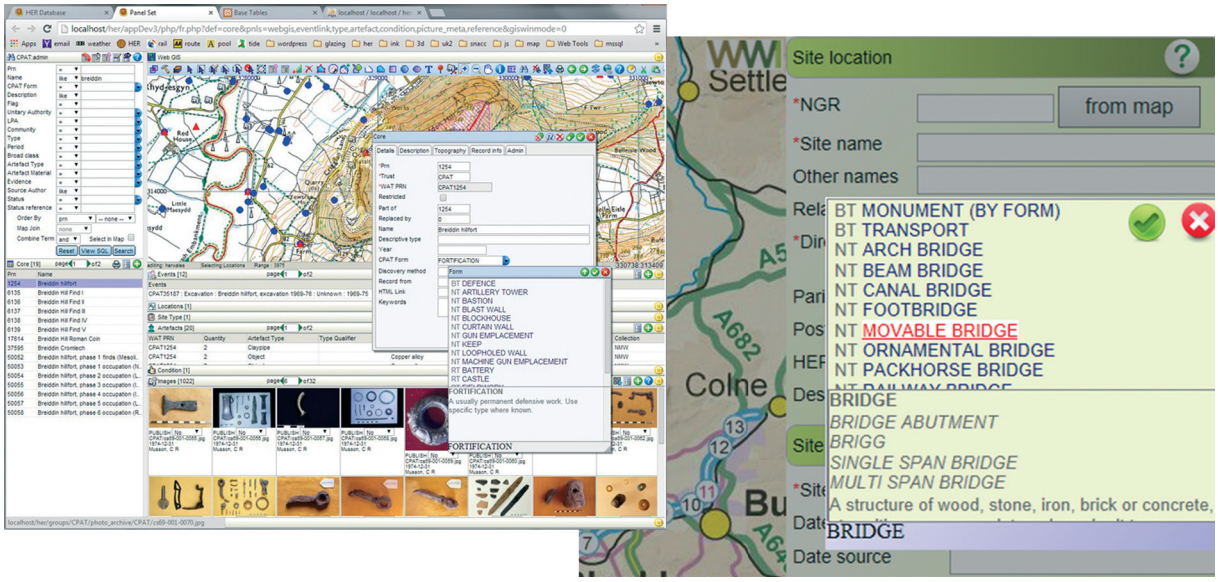

Fig. 4 Clwyd-Powys experimental field recording system with HeritageData Widgets

The hope is that by converting more standardised national thesauri, and aligning other more detailed terminologies, such as used by archaeologists within their data recording systems, then it will be possible to enable far more advanced cross-search of research data online including across data recorded using different languages. But for now that degree of cross-language capability may still be some way off.

Wider use of HeritageData.org vocabularies

A number of encouraging early adoptions of the HeritageData.org web services show the potential opportunities for expanding the use of the LOD vocabularies. Examples include the inclusion of English Heritage thesauri of maritime craft terms by the British Oceanographic Data Centre in their Linked Open Data of oceanographic survey vessels. The Archaeology Data Service, not so surprisingly, has made use of the LOD vocabularies to align key terms in their archive metadata to the LOD terms for national monuments and periods and have described the processes concerned in more detail on their website (http://archaeologydataservice.ac.uk/blog/2013/07/seneschal-value-to-the-ads/).

As a demonstration of the flexibility of the widgets we have also seen the archaeological trust of Clwyd-Powys in Wales adopting use of the terminology widgets for including the RCAHMW Welsh Monuments Type Thesaurus in an experimental field recording mobile app (**Fig. 4**).

NLP Information Extraction (IE) of Concepts from OASIS Grey Literature Reports

A final area where there is still much potential for using the SKOS versions of vocabularies is in the area of Natural Language processing (NLP). As part of the work on the STAR project (Tudhope, May, Binding & Vlachidis 2011; Vlachidis, Binding, May & Tudhope 2013) a corpus of so-called grey literature comprising about 500 archaeological reports were analysed using a 'pipeline' of Natural Language Processing techniques to attempt to develop a semi-automated process for the extraction, or highlighting, of specific archaeological concepts within the free text of the reports.

The pipeline is built up using an NLP toolkit to define a series of related syntactical and semantic rules and the pipeline relies upon the use of a standard ontology to express key concepts to be extracted. STAR used the CIDOC CRM ontology with specific archaeological conceptual extensions (CRM-EH) and also used a number of controlled vocabularies, including earlier versions of the SKOS-ified national thesauri prior to their being made available as LOD.

The outcomes from using the SKOS controlled vocabularies to identify key concepts such as "Places", "Periods" or "Object" types along

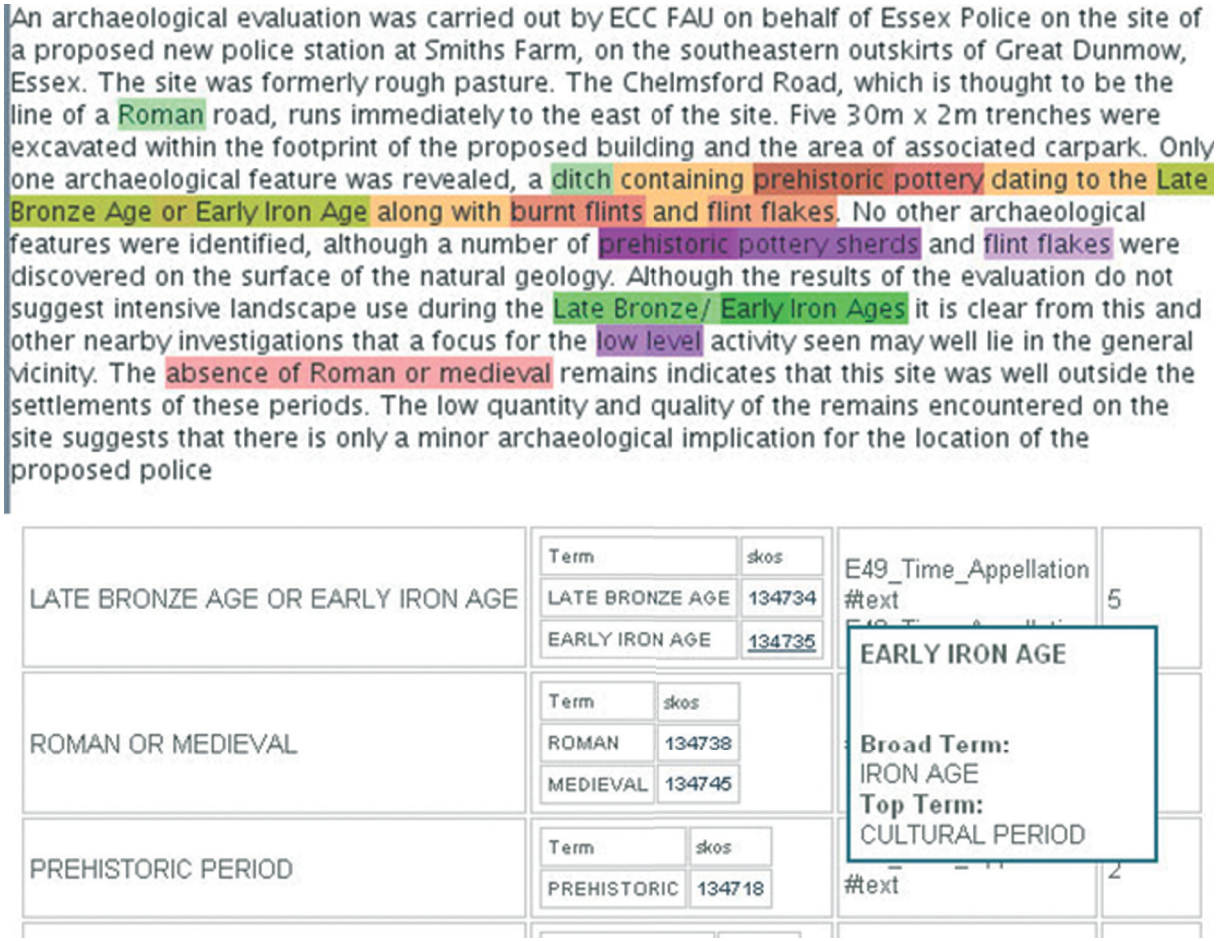

Fig. 5 Natural Language Processing extraction of Concepts from Grey Literature Reports

with the CIDOC CRM ontology suggests there would be potential for further semi-automated indexing of other archaeological grey literature to enable enhanced indexing services, and this is likely to be a developing field for the future if more archaeological terminologies can be made available as LOD (**Fig. 5**).

Conclusions

Different archaeological recording systems share common conceptual frameworks and semantic relationships. By conceptualising common relationships in our different data sets at a broad level and *aligning* vocabularies of shared reference terms we can cross-search data for patterns and broader answers to related research questions. The technologies are being developed in other domains (e.g. biology) but the question remains in archaeology, where there are different traditions and time-scales for publication, whether there is a common will for sharing archaeological data openly and in a timely manner for re-use in the interests of improving research methods?

If archaeological data is made available as Linked Open Data there may also be some blurring of the existing processes and associated boundaries for publication of archaeological results, as (Big) data integration becomes more dynamic between data sets that have been published online from different stages in the archaeological research process. STAR research suggests that there are still four key stages for coherent data integration in the Archaeological Research Cycle:

- Excavation results
- Outcomes of Analysis after excavation is completed
- "Completed" Publication of synthesised results
- Integrated archive for new research

Again, some of these may be related to methodologies for data recording, but the main point is that interpretations of archaeological data can be revised throughout the research process, so it is important to keep track of the processes involved. It will remain important for viable use of open access data that suitable mechanisms are put in place and adopted and advocated by archaeologists for adequate citation of data (e.g. DataCite). In particular the adoption, linking and re-use of data using Linked Open Data technologies, could be greatly supported and more readily adopted and promoted if some further mechanisms could be put in place by the W3C establishing mechanisms for identifying where and how other bodies or systems have made links and co-references to Linked Open Data once it has been published on the web.

Acknowledgements

Many thanks go to Paul Cripps who worked at English Heritage while producing the original CRM-EH modelling and who has contributed on subsequent projects that have used that modelling as such an excellent basis for on-going research. Thanks also to Anja Masur for sharing work on the investigations of European archaeological recording systems made possible through her ARIADNE project collaboration and particularly for the context sheet illustrations. Also thanks to Andreas Vlachidis, a STAR project team member, for his innovative development of the approaches using Natural Language Processing and for the use of his illustration. Particular credit goes to colleagues from the national heritage agencies, Phil Carlisle at English Heritage, Peter McKeague RCAHMS, and David Thomas at RCAHMW. Without their collaboration and work on SENESCHAL, along with Holly Wright from ADS, then the publishing of the LOD vocabularies on Heritagedata.org would not have proved possible. We are most grateful for the contribution of the Arts and Humanities Research Council for funding SENESCHAL which built on previous AHRC funding for STAR and STELLAR. Finally to acknowledge the role of the ARIADNE EU funded FP7 project for enabling a number of workshops and conference sessions which have helped shape the wider European perspective on this research.

References

All cited URLs last visited on 09-01-2015.

ARIADNE: http://www.ariadne-infrastructure.eu/

Binding, C., Tudhope, D. & May K. (2008). Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOCCRM. *Proceedings (ECDL 2008) 12th European Conference on Research and Advanced Technology for Digital Libraries, Aarhus* (pp. 280-290). Lecture Notes in Computer Science, 5173. Berlin: Springer. – Final preprint presentation DOI:10.1007/978-3-540-87599-4_30

Binding, C., May, K., Souza, R., Tudhope, D. & Vlachidis A. (2010). Semantic Technologies for Archaeology Resources: Results from the STAR Project. In F. Contreras, M. Farjas & F. Melero (eds.), *Proceedings 38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology* (pp. 555-561). Granada: Archaeopress.

Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked Data – the story so far. *International Journal on Semantic Web and Information Systems*, 5 (3), 1-22. DOI:10.4018/jswis.2009081901

Cabinet Office (2011). *Information Principles for the UK Public Sector*. http://www.cabinetoffice.gov.uk/sites/default/files/resources/Information_Principles_UK_Public_Sector_final.pdf

Carver, M. (2009). *Archaeological Investigation*. Abingdon: Routledge.

Cripps, P., Greenhalgh, A., Fellows, D., May, K. & Robinson, D.-E. (2004). Ontological Modelling of the work of the Centre for Archaeology**.** *CIDOC CRM Technical Paper*. http://www.cidoc-crm.org/docs/Ontological_Modelling_Project_Report_%20Sep2004.pdf

Crofts, N., Doerr, M., Gill, T., Stead, S. & Stiff, M. (eds.) (2008). *Definition of the CIDOC Conceptual Reference Model*. CIDOC-CRM website – http://cidoc.ics.forth.gr/index.html (now formally ISO 21127:2006).

HeritageData.org: http://www.heritagedata.org/blog/

Hodder, I. (ed.) (2000). *Towards Reflexive Method in Archaeology*. Cambridge: McDonald Institute for Archaeology.

Levenshtein, V. I. (1966). "Binary codes capable of correcting deletions, insertions, and reversals". *Soviet Physics Doklady* 10 (8), 707-710.

May, S., Attewell, B., Cripps, P., Crosby, V., Cromwell, T., Graham, K., Heathcote, J., Jones, C., Lyons, E., May, K., Payne, A., Reilly, S., Robinson, D., Stonell-Walker, K., Schuster, J., Walkden, M. (2004). *Revelation: Phase 1 Assessment.* English Heritage Research Report 78/2004. http://services.english-heritage.org.uk/ResearchReportsPdfs/078-2004.pdf

Miles, A., Matthews, B. & Wilson, M. (2005). SKOS Core: Simple Knowledge Organisation for the Web. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, 5-13.

Pavel, C. (2010). *Describing and Interpreting the Past.* Bucharest: University of Bucharest Press.

SKOS Core: http://www.w3.org/2004/02/skos/

STAR project: http://hypermedia.research.southwales.ac.uk/kos/star/

STELLAR project: http://hypermedia.research.southwales.ac.uk/kos/stellar/

SENESCHAL project: http://hypermedia.research.southwales.ac.uk/kos/SENESCHAL/

Tudhope, D., May, K., Binding, C. & Vlachidis, A. (2011). Connecting archaeological data and grey literature via semantic cross search. *Internet Archaeology, 30:* http://intarch.ac.uk/journal/issue30/tudhope_index.html

Vlachidis, A., Binding, C., May, K. & Tudhope, D. (2013). Automatic Metadata Generation in an Archaeological Digital Library: Semantic Annotation of Grey Literature. In A. Przepiórkowski, M. Piasecki, K. Jassem & P. Fuglewicz (eds.). *Computational Linguistics – Studies in Computational Intelligence 458*, (pp. 187-202). Berlin: Springer.

About the authors

Keith May FSA
Keith May is Information Strategy Advisor for historic environment digital research in the Capacity Building and Knowledge Transfer Team at English Heritage, soon to become Historic England. He joined English Heritage as an Archaeological Information Officer developing, project managing, and providing Quality Assurance for their archaeology project grants budgeting and Management Information Systems. His roles have included project assurance officer, information and systems analyst, field archaeologist in England and internationally, project manager, and project management trainer. Before joining English Heritage he worked as a field archaeologist in London, Hertfordshire and Kent. His research interests and specialisms span the inter-relationships between Archaeology, Digital Information Strategies, Knowledge Management, and Cultural Heritage Ontologies.

Capacity Building Team
Strategic Planning & Management Division
Heritage Protection Department
English Heritage
Fort Cumberland
Fort Cumberland Road
Eastney
Portsmouth PO4 9LD
United Kingdom
Phone: +44 (0)23 9285 6755
Keith.May@english-heritage.org.uk
http://www.english-heritage.org.uk/

Ceri Binding
Ceri Binding is a Research Associate in the Hypermedia Research Group within the Faculty of Computing, Engineering and Science, University of South Wales. He graduated with a BSc in Computer Studies in 1997 whilst working as an Analyst Designer / Programmer in the water industry, before joining the University in 2000. He had responsibility for development work on a number of EPSRC and AHRC funded projects including FACET, STAR, STELLAR & SENESCHAL. He is currently involved with the ARIADNE FP7 Infrastructures project. Related research interests include Knowledge Organisation Systems and Applied Semantic Technologies.

Hypermedia Research Group
School of Computing and Mathematics
Faculty of Computing, Engineering and Science
University of South Wales
Llantwit Road
Pontypridd CF37 1DL
United Kingdom
Phone: +44(0)345 576 0101
ceri.binding@southwales.ac.uk
http://hypermedia.research.southwales.ac.uk/kos/

Prof Douglas Tudhope
Douglas Tudhope is Professor in the Faculty of Computing, Engineering and Science, University of South Wales and leads the Hypermedia Research Group. His main research interests

are the intersecting areas of information science, hypermedia and the semantic web. He directed the AHRC funded STAR, STELLAR and SENESCHAL projects applying semantic techniques in archaeology, in collaboration with English Heritage, the Archaeological Data Service and other partners. He leads the Linking Archaeological Data Work Package for the ARIADNE FP7 Infrastructures Project. Since 1977, he has been Editor of the journal, New Review of Hypermedia and Multimedia. He serves as a reviewer for various journals and international programme committees and is a member of the Networked Knowledge Organisation Systems/ Services (NKOS) network. He was a member ISO TC46/SC9/SC8 (and NISO) working group developing a new thesaurus standard (ISO 25964).

Hypermedia Research Group
School of Computing and Mathematics
Faculty of Computing, Engineering and Science
University of South Wales
Llantwit Road
Pontypridd CF37 1DL
United Kingdom
Phone: +44(0)345 576 0101
douglas.tudhope@southwales.ac.uk
http://hypermedia.research.southwales.ac.uk/kos/