When Ontology and Reality Collide: The Archaeotools Project, Faceted Classification and Natural Language Processing in an Archaeological Context

Stuart Jeffrey[1,3] – Julian Richards[1] – Fabio Ciravegna[2] – Stewart Waller[1] – Sam Chapman[2] – Ziqi Zhang[2] – Anthony Austin[1]

[1]Archaeology Data Service, University of York, UK (ADS)
[2] Web Intelligence Technologies Group, University of Sheffield, UK
[3]sj523@york.ac.uk

Abstract

As a direct result of a successful proof of concept demonstrator for an archaeological faceted classification browsing system, "Archaeobrowser" the Archaeological Data Service (ADS) and the Natural Language Processing Research Group at the University of Sheffield have embarked on a further project, named 'Archaeotools'. Archaeotools is funded by the UK's e-Science Research Grants Scheme which itself is a collaboration between three major funding bodies, the AHRC, the EPSRC and the JISC. This project represents the first UK implementation of a faceted classification system in an archaeological context, specifically to enhance the ADS's ArchSearch facility moving the search paradigm away from the search box approach towards a more intuitive and informative faceted browser system. Archaeotools is also using natural language processing to tackle the problem of unstructured, but highly valuable, archaeological data by automatically extracting metadata from legacy datasets such as 'grey literature'.

Keywords

Ontology, Faceted Classification, Natural Language Processing

1. Introduction

Archaeology as a discipline has a long history and as a result there is a large corpus of relevant material in printed form dating back, in some cases, to the middle of the nineteenth century or even earlier. Much of this is fully published and is accessible via traditional library services. However, even more of this material is either published via very short run academic journals or not published fully at all. In the case of unpublished material or 'grey literature' this in no way implies that the value of the archaeological information contained therein is not worthy of publication. In fact the explosion of grey literature in the UK and other countries in recent years, a direct result of planning control legislation, has led to very significant archaeological material, both in terms of quality and in terms of volume, being relegated to this less than satisfactory form of publication. In recent years despite the detrimental effect of this inaccessibility and difficulty of discovery, large amounts of archaeological information has begun to be recognised by the academic community (for a good example see Bradley 2006). The Archaeology Data Service (ADS) based at the University of York, supports research, learning and teaching with high quality and dependable digital resources (ADS website 2008). As part of this role it has gathered together digital versions of over 2000 grey literature publications, each with manually generated metadata, although this should be compared to estimates as to the number of paper grey literature reports which go as high as 30,000. For any attempt to digitis this disparate and distributed set of records to facilitate broader access, the key in terms of both cost and time would be metadata generation. An aspiration of the ADS would be to develop a methodology that would allow automated metadata generation form digital versions of grey literature (and other forms of legacy literature, such as historic journals). This is one of the main objectives of the Archaeotools project and it is an exploration of the potential of Natural Language Processing technologies to solve this problem that forms the heart of two of the project's work packages. With this goal in mind the Natural Language Processing Research Group based at the Department of Computer Science, University of Sheffield are natural partners and collaborators with the ADS on the Archaeotools project (University of Sheffield website 2008).

The discoverability and accessibility of grey literature and legacy literature is only part of the problem. The ADS aggregates over 1,000,000 resource discovery metadata records from a number of large and significant sources including National Monuments Records, Historic Environment Records and Sites and Monuments Records as well as it's own archive holdings. As a result of the developmental history of these various datasets, the terminology used, the record structure and the record metadata are all extremely variable. This can cause significant problems for researchers trying to conduct analysis of these data that relies on completeness or is predicated on the records adhering to agreed terminological norms. These difficulties are accentuated by the now common 'Google' search paradigm where a user is presented with an empty search box and invited to think of the most appropriate search terms, often referred to as a 'type and hope' approach. This is a long way from being the only possible search paradigm for structured and semi-structured datasets, such as those aggregated by the ADS. Previous work at the ADS has demonstrated clearly that a faceted classification approach to large datasets and the associated facet browser results in a significantly more intuitive, usable, complete and reliable searching. The Archaeotools project intends to take the experience garnered by the ADS during its previous Common Information Environment (CIE) Archaeobrowser project and roll this out to service with a full geospatially enabled facet classification browsing system enhancing access to the large aggregated dataset (Common Information Environment website 2007).

The two broad project objectives outlined above are in fact extremely complementary and it is the Archeotools project's implementation of these approaches together that offers such potential. Not only is it intended that the faceted classification browser system works as an interface to the aggregated datasets hosted by the ADS, but it is also intended that the grey literature holdings, and even historic literature holdings, will be integrated into these datasets making them discoverable and searchable via the same faceted browsing interface. In short, the objective of the project can be summed up as being to allow archaeologists to discover, share and analyse datasets and legacy publications which have hitherto been very difficult to integrate into existing digital frameworks.

The Archaeotools project is funded by a joint eScience research grants scheme in the UK, the co-funders of this scheme are the Arts and Humanities Research Council, the Engineering and Physical Sciences Research Council and the Joint Information Systems Committee. Started in 2007 the project is due to finish in 2009, this short paper concentrates on detailing the approaches adopted by the project in reaching its objectives and presenting the first glimpse into the levels of mismatch that exist between existing (legacy) archaeological datasets and the relatively newly developed and implemented thesauri. The existence of these thesauri is integral to both the automated metadata extraction and to the implementation of a faceted classification browsing system (Archaeotools website 2008).

2. A faceted classification browser for archaeology

The initial work package has involved the construction of an advanced faceted classification browser for UK archaeology, including a fully interactive geospatial search interface. A good introduction to the concept of faceted classification is available from The Knowledge Management Connection website (2008). The ADS holds aggregated (resource discovery) datasets from the National Monuments Records of England, Scotland and Wales, Historic Environment Records from around 30 local authorities as well as Sites and Monuments Records from organisations like the National Trust and the Museum of London Archive. Taken together these resource discovery metadata records total over 1,000,000 and are currently searchable via the ADS's ArchSearch interface which uses a traditional 'type and hope' search box.

Previous work carried out by the ADS with regard to faceted classification as a search mechanism demonstrated that the most appropriate search facets for this datasets are:

- What – what does the record refer to?
- When – what is the archaeological date (range) of interest?
- Where – what is the location or region of interest?
- Media – what is the form of the record you are ultimately interested in?

Of course any number of other facets are possible and even desirable (e.g. Who – to whom does the record relate?), but for a practical system these four are the initial facets that are deemed to deliver the greatest utility for the researcher. An as yet undeveloped component of the project is the investigation of how additional facets might be specified and whether user-generated facets are feasible; this is not dealt with in this paper.

These facets need to have a hierarchical structure underpinning them in order for them to be browsed, and we are fortunate in the UK that we have hierarchical thesauri deployed or under development that allow us to populate a browsing structure for each facet. These thesauri, or word lists, are standards that have been generated via a number of sources by they are key to the project that each has a controlling body, are recognised as *de jure* or *de facto* standards and are either already being broadly used or are in the process of being adopted. For the above facets these thesauri were selected:

- What – The Thesaurus of Monuments Types (TMT, English Heritage 2008)
- When – MIDAS Period list (MIDAS website 2008; FISH website 2008)
- Where – County, District, Parish (UK Government list of administrative areas)

A fully controlled word list for media type is not currently implemented. An example of how the hierarchical structure works with a quite detailed record, of monument type 'Cattle Stall' might be:

What →

Agriculture and Subsistence →

Animal Stall →

Cattle Stall

As can be seen for the above example this structure lends it's self well to a 'point and click' browsing approach where each level of the hierarchy is expanded in turn. Again, informed by the experience of the CIE Archaeobrowser project it was this clickable facet tree approach that has been adopted for the Archaeotools browser interface. Just how powerful this approach is on a normalised dataset is demonstrated by a user's ability to drill down to a specific (and complete) set of records with the minimum of clicks. In tests on the Archaeobrowser system it was possible to go from the maximum number of records (c. 1 million) to a selected set representing Bronze Age funerary monuments within 5km of a specific location in North Yorkshire (only 16 sites) with just three clicks of the mouse. Not only does this compare very favourably to traditional search box based techniques that fact that the data has been normalised (i.e. mapped to the terms of the thesaurus) means that the user has a much higher level of confidence in the completeness of the returned results and is much less troubled by the return of false positive results. The classification mechanism adopted by the Archaeotools project was Solr (Solr website 2008), an open source enterprise search server based on the Lucene Java search library. It is this process of recasting the data sets that the ADS delivers in order to retrospectively allocate them to the facets in the facet tree structure that threw up a number of interesting results which are dealt with in the following section.

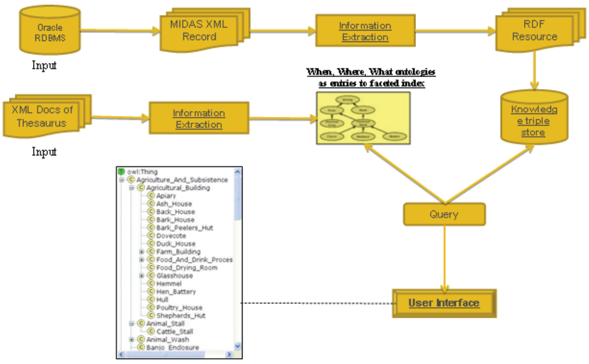
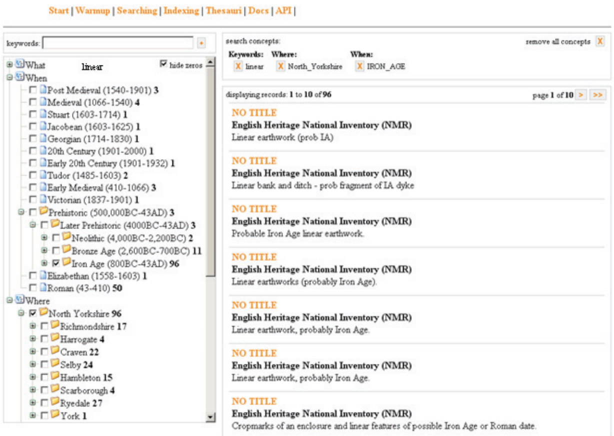


Fig. 1.

Fig. 1 shows the process architecture adopted for the Archaeotools project. In brief, selected fields (parish, period, type etc.) are extracted from the ADS Oracle database in MIDAS XML format data, then extracted to Resource Description Framework (RDF) format. XML (OWL) versions of the thesaurus are extracted to create workable ontologys, and these in tandem with the RDF knowledge triple store are queried to classify the records (Solr).

Above (*Fig. 2* is a screen shot of the prototype faceted classification browsing tree for the Archaeotools project. Whilst the underlying technology will not change the final implemented version, due online autumn 2008, it is expected to have a different 'look and feel' in order to sit within the general ADS website house style. It is important to note that the browser interface shows the number of records associated with each facet and also allows the user to hide nodes that have no records. This last feature makes the facet tree much easier to navigate as only nodes that have records associated with them will be shown. For example there would not be any 'henge' monument


Start | Warmup | Searching | Indexing | Thesauri | Docs | API |
keywords:
☑ hide zeros
remove all concepts X
> >>

What linear

When
- □ Post Medieval (1540-1901) **3**
- □ Medieval (1066-1540) **4**
- □ Stuart (1603-1714) **1**
- □ Jacobean (1603-1625) **1**
- □ Georgian (1714-1830) **1**
- □ 20th Century (1901-2000) **1**
- □ Early 20th Century (1901-1932) **1**
- □ Tudor (1485-1603) **2**
- □ Early Medieval (410-1066) **3**
- □ Victorian (1837-1901) **1**
- □ Prehistoric (500,000BC-43AD) **3**
  - □ Later Prehistoric (4000BC-43AD) **3**
    - □ Neolithic (4,000BC-2,200BC) **2**
    - □ Bronze Age (2,600BC-700BC) **11**
    - ☑ Iron Age (800BC-43AD) **96**
- □ Elizabethan (1558-1603) **1**
- □ Roman (43-410) **50**

Where
- ☑ North Yorkshire **96**
  - □ Richmondshire **17**
  - □ Harrogate **4**
  - □ Craven **22**
  - □ Selby **24**
  - □ Hambleton **15**
  - □ Scarborough **4**
  - □ Ryedale **27**
  - □ York **1**

search concepts:
Keywords: X linear **Where:** X North_Yorkshire **When:** X IRON_AGE

displaying records: **1** to **10** of **96** page **1** of **10**

NO TITLE
English Heritage National Inventory (NMR)
Linear earthwork (prob IA)

NO TITLE
English Heritage National Inventory (NMR)
Linear bank and ditch - prob fragment of IA dyke

NO TITLE
English Heritage National Inventory (NMR)
Probable Iron Age linear earthwork.

NO TITLE
English Heritage National Inventory (NMR)
Linear earthworks (probably Iron Age).

NO TITLE
English Heritage National Inventory (NMR)
Linear earthwork, probably Iron Age.

NO TITLE
English Heritage National Inventory (NMR)
Linear earthwork, probably Iron Age.

NO TITLE
English Heritage National Inventory (NMR)
Cropmarks of an enclosure and linear features of possible Iron Age or Roman date.

Fig. 2.

types appearing in a facet tree where 'medieval' is a selected period, as henges are an exclusively prehistoric monument form.

3. Record and thesauri mismatches

It has long been understood that any large monument inventory, especially one that has developed over a number of years (indeed starting in the 19th century in the case of some) is unlikely to confirm perfectly to any schema or controlled terminology sets, especially as these have been developed relatively recently. The Archaeotools project is the first instance of any archaeological project in the UK that has tried to both generate metrics on this mismatch (see *Table 1* below) and to mitigate the problem via a combined automated and manual attack.

This process generated a large amount of statistics which are summarised in the table below. For the purposes of this paper only the 'headline' figures are shown, i.e. they are not broken down by data set. In fact it is true to say that all datasets contributed to these mismatches more or less equally and that there was no obvious data set where the terminology used diverged more radically from the thesauri than all the others.

The numbers given below are derived from a total aggregated record set of 1,001,107 records and all percentages represent a percentage of this number.

The figures for 'where' with terms not found in the CDP (24.5%) can be safely ignored as these figures were generated prior to the integration of the Scottish CPD into the thesauri set; this comfortably accounts for the majority of these missing terms.

The results of this analysis showed that, contrary to the expectations, it was possible to fully map these record sets to the thesauri (and therefore facets) by combined automatic, say by the use of regular expressions, and manual techniques. This is best exemplified by the 'when' facet. There is a huge number of ways in which archaeological dates and date ranges can be written, e.g. 1066, 1001–1100, 11th Centuary(sic), C11, 11C, Eleventh Century. The vast majority of these can be mapped directly to date ranges. In the case of Archaeotools this was MIDAS defined date ranges, with regional variations. Our analysis recovered 457 types of irresolvable dates, but in practice this ultimately equated to only c.700 records. A figure of 700 is perfectly manageable in terms of manual intervention, especially in comparison to the scale of the original dataset, and this was how the final mapping for this facet was complete.

In general the most fruitful form of enhanced automated classification was to extend the search for thesauri terms to monument 'description' fields. Although not entirely problematic it allowed for relevant dates and date ranges that were not in the appropriate date/period field to be extracted. This process is not reflected in the above figures or explained in detail here.

What	
Records that have **no** subject information	19,269 records (2%)
Records that use terms **not** found in TMT, so these records cannot be indexed (**6,442** unique terms)	101,507 records (10.1%)
When	
Records that have **no** temporal information	292,793 records (29.2%)
Records that use period terms **not** found in MIDAS so these records cannot be indexed (**457** types of irresolvable dates)	114,505 (11.4%)
Where	
Records that have **no** spatial information	11,126(1.1%)
Records that use terms **not** found in CDP, so these records cannot be indexed.	245,601 records (24.5%)

Table 1.

4. 'Grey literature', legacy texts and automated metadata extraction

This Archaeotools work package, the extraction of metadata from grey literature using natural language processing, is at an early stage. However this phase of the Archaeotools project is based on work that was carried out by the Natural Language Processing (NLP) group at the University of Sheffield and Professor Mark Greengrass at the Department of history at the same university. The Armadillo project (Armadillo website, 2007) used to perform data mining on historical records from the Old Bailey (law courts) in the City of London. This project was highly successful in extracting names, locations and trial details from these records and mapping them to a pre-defined ontology and also in allowing the discovery of previously unknown relationships between witness and defendants in different cases (Greengrass et al 2008). Archaeotools is applying the same general technique, but applying it to 'semi-structured' archaeological documents. The documents selected are the ADS 'grey literature' holdings, a primary reason for this is that the ADS also has manually generated metadata for this corpus that can be used to evaluate the success of the NLP automated extraction. For an example of grey literature see Conolly 2003 or the ADS Library of Unpublished Fieldwork Reports (ADS website 2008).

The ultimate objective of this metadata extraction is to allow grey literature to be discoverable and searchable in the faceted browsing interface by classifying their metadata using the same process as described above for monument inventories.

The proposed final phase of the Archaeotools project is to refocus the experience of the NLP automated metadata extraction on 'semi-structured' grey literature to the almost entirely unstructured digitised version of the Proceeding of the Society of Antiquaries of Scotland (PSAS). Going back to 1851, these extremely valuable journals are archived and disseminated by the ADS (in digital form, ADS website, Library 2008). It is hoped that a meaningful metadata dataset can be generated from this resource using a similar technique to that for the grey literature. An example of what is hoped to be extracted can be seen here:

Here is section of original PSAS text:
"*The bronze ring inscribed with runic characters, presented to the Society, was found in the year 1849, in the Abbey park, in the immediate neighbourhood of St Andrew. It is a large bronze finger ring inscribed on two faces in Anglo-Saxon runes, and is of peculiar interest, as being, it is believed, only example of the Paleography of our Anglo-Saxon forefathers hitherto found in Scotland, with the single, but most important exception of the noble monument at Ruthwell, Dumfrieshire*" (Wilson 1851)

Using NLP the following data is potentially be extracted from it.
What – Bronze Ring, Runic Inscription
Where – Abbey Park, St Andrews (not Ruthwell)
When – Anglo-Saxon (found 1849)
Who – Wilson, D.
Media – PSAS(PDF)

Clearly this type of extracted data would mesh perfectly with the already implemented faceted browsing interface discussed in earlier sections. There is the obvious potential to aggregate resource discovery metadata relating to the PSAS directly with the other datasets that have been made searchable in this way. It should be noted that the highly unstructured nature of the text and the antiquated use of language raises the possibility that this will be a very challenging phase of the project.

One of the most exciting prospects, as yet unrealised, is that place names extracted from PSAS can be 'cross walked' to an existing gazetteer web service hosted at EDINA, University of Edinburgh (EDINA 2008). It is proposed that extracted place names can be sent directly to this service and the service will automatically return National Grid References for that place name (centred) or in the case of some urban areas an actual polygon definition. This would allow the relevant place name from PSAS to be mapped in the Archaeotools geo-spatial interface and therefore make them as discoverable/searchable as standard monument inventory datasets.

5. Conclusion

The Archaeotools project has successfully implemented a faceted classification browsing system in the context of aggregated archaeological records. This service is due to be released for public access as a replacement for the existing ArchSearch interface in the autumn of 2008. User needs research and user testing that was carried out as part of the

CIE Archaeobrowser project indicates that the browsing interface itself is likely to be very well received. In preparing the datasets for presentation by this interface a deep insight into the condition of archaeological monument inventories has been gained. Despite the apparent mismatch between rather loose terminology of the historical datasets and the rigorous control of word lists, thesauri and ontologies, in practice a combination of automated and manual approaches allowed for the classification process to be comprehensive and meaningful. In addition the classification process is essentially a 'one off' with regard to erroneous or missing classifications as, although there are over 1 million records in the datasets, they increase at a much slower rate (c 5000 per annum) meaning that future mismatches or missing facets are much more likely to be in small and manageable numbers.

The other two major components of the project, automated data and metadata extraction from grey literature and legacy literature, are only now going ahead at full speed. Precedents that have been set, such as the Armadillo project mentioned above, indicate we should be confident that basic metadata can be extracted at the very least, which in itself would be a significant achievement in an archaeological context. However, the ADS and the University of Sheffield are optimistic that at future CAA conferences we will be reporting that much richer levels of data mining can be demonstrated and a major obstacle to tackling digitisation back logs and accessing legacy datasets will be removed.

Bibliography

Archaeology Data Service website (2008). http://ads.ahds.ac.uk

Armadillo project website (2008). http://www.hrionline.ac.uk/armadillo/

Bradley, Richard (2006). Bridging the two cultures. Commercial archaeology and the study of prehistoric Britain *Antiquaries Journal* 86, 1–13.

Common Information Environment project website (2007). http://ads.ahds.ac.uk/project/cie/index.html

Conolly, Richard (2003). Archaeological watching brief at 1–11 Main St, Brampton, Cumbria. Headland Archaeology, Edinburgh (Unpublished Fieldwork Report).

EDINA geoXwalk website (2008). University of Edinburgh: http://www.geoxwalk.ac.uk/

English Heritage, NMR Thesaurus website (2008). http://thesaurus.english-heritage.org.uk/

Forum for Information Standards in Heritage website (2008): http://www.fish-forum.info/

Greengrass, Mark, Sam Chapman, Jamie McLaughlin, Ravish Bhagdev and Fabio Ciravegna (2008). "Finding Needles in Haystacks. Data-Mining in Distributed Historical Datasets". In: *Virtual Histories and Pre-Histories: Finding Meanings*. London: Ashgate.

Jeffrey, Stuart, William Kilbride, Julien Richards and Stewart Waller (2008). Life beyond Google: The CIE Archaeobrowser, CAA2007. *Proceedings of the CAA*. Forthcoming (2008)

MIDAS Heritage website (2008). http://www.english-heritage.org.uk/server/show/nav.18041

The Knowledge Management Connection website (2008). http://www.kmconnection.com/DOC100100.htm

Sheffield University NLP group website (2008). http://nlp.shef.ac.uk/

Solr website (2008). http://lucene.apache.org/solr/

Wilson, Daniel (1851). "Inscribed Ring" in the *Proceedings of the Society of Antiquaries of Scotland*, Volume 1. Edinburgh, 1851–54.