Integrating archaeological literature into resource discovery interfaces using natural language processing and name authority services.

S.Jeffrey[1], J.Richards[1], F.Ciravegna[2], S.Waller[1], S.Chapman[2], Z.Zhang[2]

[1] *Archaeology Data Service, Department of Archaeology, The King's Manor, University of York, Y01 7EP. Sj523@york.ac.uk*

[2] *Web Intelligence Technologies Lab, Organizations, Information and Knowledge (OAK) Group, Department of Computer Science, University of Sheffield, S1 4DP.*

Abstract

There exists a large and underutilized resource of archaeological literature, both formal, such as scholarly journals and less formal in the form of 'grey literature'. In the archaeological domain the vast majority of this literature contains some geo-spatial element as well as the expected temporal information and therefore its ease of discovery would be greatly enhanced were it accessible via a geo-spatially enabled search mechanism. As a result of this, geo-referencing these types of material and integrating them with other resources, such as monument inventories, is seen as a desirable enhancement for digital archives serving the archaeological research community. This paper provides an overview of a number of the approaches to the integration of such legacy literature into geospatial search mechanisms in an archaeological context. In particular efforts to achieve this via the Archaeotools e-Science project and its use of natural language processing and a geo-spatial cross-walk service are discussed as well as potential future enhancements to the process.

1. Introduction

The Archaeology Data Service (ADS)[1] has a core role as a digital archive for archaeological outputs originating in UK's higher education domain, in practice this role has extended to touch on every sector of archeology in the UK. In addition the ADS has been providing free online access to it's digital resources for teaching, learning and research for over a decade. Primarily this access is facilitated through its main search interface, 'ArchSearch'. This catalogue and interface now provides access to over one million metadata records covering the archaeology of the British Isles and beyond. These records are variously derived from the National Monuments Records (NMRs) of England, Scotland and Wales, as well as local authority Historic Environment Records (HERs) and the ADS's own archive holdings.

Archaeology as a discipline has generated a large corpus of printed material dating back to the nineteenth century and earlier. Much of this is fully published as monographs or journal articles and is accessible through traditional academic routes. However, as a result of policy and legislative changes, the majority of fieldwork reports generated in the last twenty years are not published at all beyond a typescript report lodged with the local planning authority. In the case of this unpublished material, or 'grey literature', the fact that it is not fully published is not an indication of its value to the research community. It is a well recognized problem in archaeology that there is a large volume of grey literature that is simply not as accessible as published material, despite the high quality of the work and the results it describes [5].

Recently the detrimental effect of the difficulty of discovery and access to the large amounts of archaeological information has begun to be recognized by the academic community. Prominent archaeologists such as Bradley [2] and Lock [10] have asked why it is not more widely available, given it's potential to

[1] http://ads.ahds.ac.uk , accessed September 2009.

change archaeological interpretations. The online delivery of the material, both born digital and legacy, would seem like a logical solution to addressing these access issues. The ultimate goal would be the integration of archaeological monument data, archaeological event data and the literature referring to these monuments and events all being searchable and accessible via the same geospatial interface.

The ADS and the Computer Science department at the University of Sheffield have been working together on an AHRC, EPSRC, NERC funded e-Science project, 'Archaeotools', with the intention of facilitating this integration. Based on research carried out by the Natural Language Processing (NLP) group at the University of Sheffield, it builds upon work undertaken at Sheffield with Professor Mark Greengrass on the Armadillo project. This performed data mining on historical court records from the Old Bailey in the City of London[2] [6]. The Archaeotools project applies the same general technique, but targeted at the ADS grey literature holdings and digital versions of the Proceedings of the Society of Antiquaries of Scotland (PSAS).

The Archaeotools project

The Archaeotools project followed a strategy that allowed it to tackle the issue of geo-spatial integration of monument inventories, grey literature and legacy literature through the creation of an advanced faceted classification and geospatial browser. The underlying dataset comprises over 1,000,000 records aggregated from the NMRs of Scotland, Wales and England, HERs and the ADS's own archive holdings. Previous work carried out on faceted classification by the ADS in the Archaeobrowser project [7] demonstrated that the most appropriate search facets for archaeological datasets are:

> What - what subject(s) does the record refer to?
> When - what is the archaeological date range of interest, and exact singular temporal point?
> Where - what is the location(s) or region(s) of interest?

These facets can be populated from existing thesauri (e.g. the Thesaurus of Monument types) in Extensible Mark-up Language (XML) format and extended/integrated to allow for geographical differences, such as terminological differences in monument and period types between Scotland and England. The Archaeotools project also integrates thesauri served in XML by Simple Knowledge Organization Systems (SKOS[3]) based web services developed by the AHRC-funded Semantic Tools for Archaeology project (STAR[4]) based at the University of Glamorgan. Facets were further populated after the NLP system had automatically extracted resource discovery metadata (and other facet types) from unpublished archaeological reports. This process was then extended to capture metadata from legacy historical documents, using the PSAS as an exemplar corpus. In the case of the grey literature actual grid references could often be extracted directly from the text. Finally the University of Edinburgh's geoXwalk service was used to recast place names and locations extracted from text (such as PSAS place names) as grid references, to allow geospatial searching of the data. Rather than cover the NLP natural language processing itself, which is described in detail elsewhere [8] [9], it is the faceted classification using place names extracted by NLP, the faceted browser and its complimentary geo-spatial interface that are of interest here.

Faceted Classification and the Geo-spatial interface.

The faceted classification approach to presenting structured datasets is increasingly common in the commercial web, but clearly lends itself to the discovery of structured datasets [3]. A faceted query engine has been deployed by a group at Columbia University which provides an interface into archaeological finds datasets [11] [12]; and also in the Open Context system at the Alexandria Archive Institute[5], but applications of faceted classification are still rare in archaeology.

In order to facilitate browsing each facet needs to have an associated ontology, expressed as a hierarchy of terms. The thesauri, or controlled word lists used in this project, have been generated via a number of sources but it is key to their usefulness and sustainability that each has a controlling body, each is recognized as a *de jure* or *de facto* standard. In an English context the 'where' facets key list is the hierarchical County, District, Parish list. This hierarchical structure lends itself neatly to a 'point and click' browsing approach such that each level of the hierarchy can be expanded or collapsed and whole

[2] http://www.hrionline.ac.uk/armadillo/ , accessed September 2009.

[3] http://www.w3.org/2004/02/skos/ , accessed September 2009.

[4] http://hypermedia.research.glam.ac.uk/kos/star/ , accessed September 2009.

[5] http://www.opencontext.org/ , accessed September 2009.

layers of the hierarchy or individual elements selected by a mouse click.

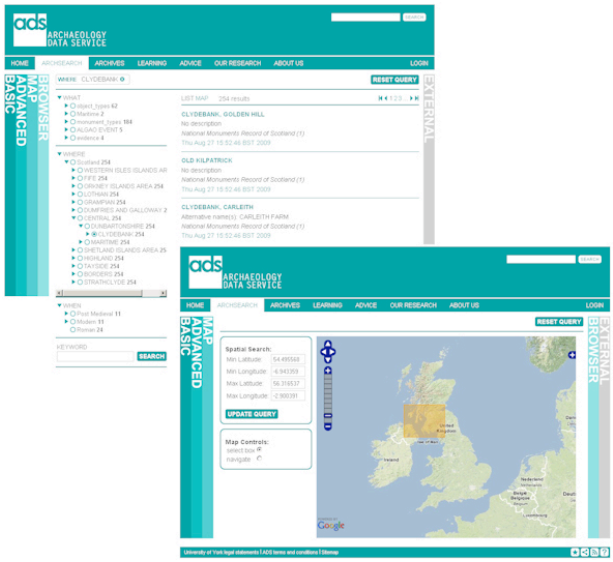

Figure 1, A screenshot of the prototype search interface.

Although we did not originally expect it to be possible, it has proven feasible to completely map the NMR/HER record sets to the thesauri, and therefore to the facets, by a combination of automatic rule-based expressions and manual techniques [9]. Similarly the combined attack on the data using Knowledge Extraction (KT) and Automatic Training (AT) [1] on the literature has already been successful to the extent that we are confident that automated resource discovery metadata extraction has been achieved for grey literature at least. Returning solely to the 'where' facet, once this element of the record (whether a monument inventory or literature record) has been identified, the next stage is to generate grid-references to allow it to be searchable via a mapping interface as well as via the browser tree. In the case of grey literature grid references or latitude and longitude are generally part of the structure of the document and, as noted above can often be extracted directly, however this was not the case for the PSAS corpus. Here it was necessary to adopt a more sophisticated approach such as the utilization of a crosswalk (look up) service linking place names with either geographic points or geo-referenced polygons, although polygons were not used in this project.

Of the total of 3991 PSAS papers, it was simply not possible to find any form of explicit grid reference in the text for 3388 (85%) using KT/AT methods, compared to a figure of just 185 (20%) for the grey literature. This can be traced to the fact that older reports, for example those from the 19th century, do not tend to use precise geospatial references to refer their geographical elements. However, by using the GeoXwalk service, a gazetteer web service hosted at EDINA, University of Edinburgh [4] it was possible to resolve all reports with a recognizable place name entity into a grid reference except for 268 reports (6.7%) – 238 of which actually had no 'Where' term at all, leaving just 30 for which a place name had been identified that could not be geo-referenced by the EDINA web service. Close manual checking uncovered that the majority of these were where a county name was the most precise spatial location that had been used in the published paper. This result demonstrates that the GeoXwalk service allows us to assign geospatial coordinates to, and thus display in a geo-spatial interface, archaeological (and antiquarian) literature in a way that would have been entirely impractical without using these automated methods.

The actual interface itself (*Figure 1.*) uses the facet index file created from both the monument inventory data and the research literature to allow searching and selection by two complementary interface mechanisms. The first is a clickable facet browser tree that expands to allow the user to search down a hierarchy, such as country, region, county, district, parish, to select the area of interest. Similar trees exist for both the 'what' facet (based on the English Heritage Thesaurus of Monument Types[6]) and the 'when' facet (based on the MIDAS[7] period list), thus the user can build up relatively complex queries using a straight forward interface. The query can be further narrowed (or expanded) by directly interacting with the map interface, in this iteration of the interface the map window is based on Open Layers[8]. Using a selection box to capture, if desired, monuments and the grey literature and legacy literature associated with them. A text search box is also always available to help build searches. The results from each searched dataset, monument inventories and archive, grey literature and historical literature (PSAS), can be tabbed through individually, rather then have them appear as a potentially confusing aggregated results list. More sophisticated analysis of search results is facilitated by

[6] http://thesaurus.english-heritage.org.uk/thesaurus.asp?thes_no=1 , accessed September 2009.
[7] http://www.english-heritage.org.uk/server/show/nav.00100300b006001 , accessed September 2009.
[8] http://openlayers.org/ , accessed September 2009.

the porting of results sets to downloadable formats such as KML[9].

Conclusion

The approach of implementing a faceted classification browsing and geospatial interface system in the context of aggregated archaeological records promises to greatly enhance the resource discovery capabilities of archaeological researchers. The service version of this new interface is scheduled to be deployed by the ADS in the winter of 2009 initially with only the monument inventory data, the geo-referenced literature being added to the live systems by spring 2010. Whilst we are confident that this general approach in itself will be beneficial we will be carrying out user testing to hone the details of interface design and to ensure usability. The possibility of using alternative browsing approaches such as one that uses a much 'flatter' ontological structure, similar to Umeå University's environmental archaeology project (SEAD[10]), will be investigated. User testing also gives us the opportunity to investigate the desirability of the extension of the facets integrated into the search interface or indeed integration of other datasets, potentially including user generated content. Additional facets such as a 'media' type, allowing the narrowing of searches by media or file type, might also be complemented by facets dealing with entirely different aspects of the content of data rather than its form, such as 'who', e.g. to whom does the data relate, which archaeologist or which archaeological organization? Our success in leveraging highly usable results from geo-spatial cross walk services suggests that Union Name List services might be utilized in the same way in the future.

10. References

[1] Appelt, D. E. & Israel, D. 1999 Introduction to information extraction technology. IJCAI-99 tutorial, Stockholm.
See http://www.ai.sri.com/wappelt/ie-tutorial/IJCAI99.pdf.

[2] Bradley, R. 2006 Bridging the two cultures. Commercial archaeology and the study of prehistoric Britain. Antiq. J. 86, 1–13. (doi:10.1017/S0003581500000032)

[3] Denton, W. 2003 How to make a faceted classification and put it on the Web. See http://www.miskatonic.org/library/facet-web-howto.html

[4] EDINA 2008 geoXwalk. See http://www.geoxwalk.ac.uk.

[5] Falkingham, G. 2005 A whiter shade of grey: a new approach to archaeological grey literature using the XML version of the TEI guidelines. Internet Archaeol. 17.
See
http://intarch.ac.uk/journal/issue17/falkingham_index.html

[6] Greengrass, M., Chapman, S., McLaughlin, J., Bhagdev, R. & Ciravegna, F. 2008 Finding needles in haystacks. Data-mining in distributed historical datasets. In The virtual representation of the past (eds M. Greengrass & L. Hughes), pp. 301–324. London, UK: Ashgate.

[7] Jeffrey, S., Kilbride, W., Richards, J. & Waller, S. 2008 Thinking outside the search box: the Common Information Environment and Archaeobrowser. In Layers of perception, Proc. 35th Int. Conf. on Computer Applications and Quantative Methods in Archaeology (CAA), Berlin, Germany, 2007 (eds A. Posluschny, K. Lambers & I. Herzog), pp. 206–211.

[8] Jeffrey, S., Richards, J., Ciravegna, F., Waller, S., Chapman, S. & Zhang, Z. In press. When ontology and reality collide: the Archaeotools project, faceted classification and natural language processing in an archaeological context. In On the road to reconstructing the past, Proc. 36th Int. Conf. on Computer Applications and Quantative Methods in Archaeology (CAA), Budapest, Hungary, 2008 (eds E. Jerem & V. Szevere´nyi).

[9] Jeffrey, S., Richards, J., Ciravegna, F., Waller, S., Chapman, S. & Zhang, Z., The Archaeotools project: faceted classification and natural language processing
in an archaeological context, Philosophical Transactions of the Royal Society. (A) (2009) 367, 2507–2519 doi:10.1098/rsta.2009.0038

[10] Lock, G. 2008 A professional mockery. British Archaeology. 101, 36–37.

[11] Ross, K. A., Janevski, A. & Stoyanovich, J. 2005 A faceted query engine applied to archaeology. In Proc. 31st Int. Conf. on Very Large Data Bases, pp. 1334–1337. Trondheim, Norway: ACM.

[12] Ross, K. A., Janevski, A. & Stoyanovich, J. 2007 A faceted query engine applied to archaeology.Internet Archaeol. 21.
See
http://intarch.ac.uk/journal/issue21/stoyanovich_index.html

[9] http://code.google.com/apis/kml/documentation/ , accessed September 2009.
[10] http://www.sead.se , accessed October 2009.