PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

The Archaeotools project: faceted classification and natural language processing in an archaeological context

S. Jeffrey, J. Richards, F. Ciravegna, S. Waller, S. Chapman and Z. Zhang

Phil. Trans. R. Soc. A 2009 **367**, 2507-2519
doi: 10.1098/rsta.2009.0038

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2009) **367**, 2507–2519
doi:10.1098/rsta.2009.0038

The Archaeotools project: faceted classification and natural language processing in an archaeological context

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

[1]*Archaeology Data Service, Department of Archaeology, The King's Manor, University of York, York Y01 7EP, UK*
[2]*Web Intelligence Technologies Laboratory, Natural Language Processing Group, Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK*

This paper describes 'Archaeotools', a major e-Science project in archaeology. The aim of the project is to use faceted classification and natural language processing to create an advanced infrastructure for archaeological research. The project aims to integrate over 1×10^6 structured database records referring to archaeological sites and monuments in the UK, with information extracted from semi-structured grey literature reports, and unstructured antiquarian journal accounts, in a single faceted browser interface. The project has illuminated the variable level of vocabulary control and standardization that currently exists within national and local monument inventories. Nonetheless, it has demonstrated that the relatively well-defined ontologies and thesauri that exist in archaeology mean that a high level of success can be achieved using information extraction techniques. This has great potential for unlocking and making accessible the information held in grey literature and antiquarian accounts, and has lessons for allied disciplines.

Keywords: archaeology; grey literature; faceted classification; information extraction; natural language processing

1. Introduction

During 2004–2005, the Archaeology Data Service (ADS) and Adiuri Systems developed a proof of concept archaeological faceted classification demonstrator on behalf of the Common Information Environment Working Group (now the Strategic Content Alliance; http://jisc/whatwedo/themes/content/contentalliance.aspx). The success of this project in demonstrating the power of a faceted browse interface to archaeological data led the ADS and the Natural Language Processing (NLP) Research Group at the University of Sheffield to embark on the Archaeotools project, funded under the UK's Arts and Humanities e-Science Initiative, itself a collaboration between three major funding

* Author for correspondence (sj523@york.ac.uk).

One contribution of 16 to a Theme Issue 'Crossing boundaries: computational science, e-Science and global e-Infrastructure I. Selected papers from the UK e-Science All Hands Meeting 2008'.

bodies: the Arts and Humanities Research Council (AHRC); the Engineering and Physical Sciences Research Council; and the Joint Information Systems Committee.

The project addresses both practical issues in applying advanced information extraction (IE) techniques to datasets generated in the arts and humanities and two key problems that have emerged in the field of archaeological informatics. These problems are the creation of search mechanisms that go beyond the naive text string searching approach of the classic search engine's search box and the automated creation of the resource discovery metadata required to underpin these more sophisticated searches.

Archaeology, as a discipline, has a long history of active fieldwork and, as a result, there is 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 via traditional library services. However, the majority of equivalent fieldwork reports generated in the last 20 years are either published via short-run journals or not published at all beyond a typescript report lodged with the local planning authority. In the case of this unpublished material, often referred to as 'grey literature', the fact that it is not fully published should not be taken to suggest that the value of the archaeological data or interpretation is not significant enough for publication. 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 (Falkingham 2005).

Historically, archaeological fieldwork was carried out by relatively few academic researchers with specific and targeted research agendas. From the mid-twentieth century, this changed and archaeological work began to take place in response to perceived threats to known archaeological sites. This 'rescue' archaeology, where a site is excavated in order to extract archaeological information before its destruction, eventually evolved into 'development control' archaeology. Often referred to simply as 'commercial archaeology', this work is carried out by numerous small-to-medium-sized charitable or commercial archaeological enterprises. Changes in UK heritage protection legislation during the 1990s precipitated an explosion of archaeological work carried out in the commercial sector. It is from this sector, which does not always have the same academic or financial imperatives for full publication as purely research-led fieldwork, that the vast majority of grey literature comes. The potentially revolutionary impact of this body of material has recently been widely acknowledged by academic archaeologists, but the problem of access remains a thorny one. In recent years, the detrimental effect of inaccessibility and difficulty of discovery of the large amounts of archaeological information represented by this material has begun to be recognized by the academic community. Prominent researchers, such as Bradley (2006) and Lock (2008), have questioned why it is not more widely available. Digital hosting and online delivery of this material, both of newly created material (i.e. 'born digital') and digitized versions of legacy documents, would seem a logical approach to addressing these access issues. However, good access is predicated on good discovery mechanisms and these rely, among other things, on good metadata.

The ADS is tasked with supporting research, learning and teaching with high quality and dependable digital resources. In fulfilling this role and as part of the Online Access to the Index of Archaeological Investigations project, it actively

gathers digital versions of grey literature fieldwork reports and currently holds approximately 2300 (Hardman & Richards 2003; Richards & Hardman 2008). This total grows by approximately 50–100 reports a month; all reports can be downloaded, free of charge, from the ADS (http://ads.ahds.ac.uk/catalogue/library/greylit/index.cfm). Each of the reports has manually generated resource discovery metadata covering attributes such as author, publisher, temporal coverage and geospatial coverage, and adhering to the Dublin Core metadata standard. Generating metadata this way may be feasible where it is created simultaneously with the report's deposit with the ADS. It would not be feasible for dealing with the tens of thousands of legacy reports known to exist. For any attempt to digitize these disparate and distributed sets of records to facilitate broader access, the key in terms of both cost and time would be metadata generation. An aspiration of the ADS is the development of a methodology that would allow automated metadata generation from digital versions of grey literature.

In addition to the pressing issue of grey literature, many of the same issues arise with reference to digitized versions of early or very short-run published material. In the nineteenth and early twentieth centuries, antiquarian research was generally published in annual journals and proceedings of learned societies. Indexing of this material rarely goes beyond author and title. This is generally inadequate for the scholar wishing to investigate previous research on a particular site or artefact class. Furthermore, while modern fieldwork reports generally provide Ordnance Survey grid references for site locations, antiquarian reports use a variety of non-standard and historic place names, making it impossible to integrate this sort of information in modern geospatial interfaces. The ADS holds significant amounts of digital versions of 'legacy' material including, for example, the annual *Proceedings of the Society of Antiquaries of Scotland* (PSAS) from 1851 to 1999. Ideally, a methodology to automatically generate metadata for grey literature should be flexible enough to be applicable to this additional dataset with the minimum of reworking.

The Archaeotools approach to the automated generation of resource discovery metadata for both grey literature and legacy literature is based on work carried out by the NLP group at the University of Sheffield. It builds upon work undertaken at Sheffield with Prof. Mark Greengrass of the Department of History. The Armadillo project performed data mining on historical court records from the Old Bailey in the City of London (http://www.hrionline.ac.uk/armadillo/). This project was highly successful in extracting names, locations and trial details from these records and mapping them to a predefined ontology, and also in allowing the discovery of previously unknown relationships between witnesses and defendants in different cases (Greengrass *et al.* 2008). The Archaeotools procedure applies the same general technique, but targeted at 'semi-structured' archaeological documents and with IE rules specific to this target corpus. The ADS grey literature holdings mentioned above were the initial set of documents selected for this project. The ADS also holds manually generated metadata for this corpus, which can be used to evaluate the success of the NLP IE. The PSAS provided the second target dataset to allow investigation of the feasibility of extending data mining to antiquarian literature.

The discoverability and accessibility of grey literature and legacy literature is only part of the challenge facing organizations, such as the ADS, in delivering digitized material. The ADS also aggregates over 1×10^6 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 its own archive holdings. As a result of the developmental history of these various datasets, the terminology used, the record structure and the record metadata all display significant heterogeneity. This can cause non-trivial problems for researchers trying to conduct any analysis, which 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, sometimes referred to as a 'type-and-hope' approach. This is far from being an optimal search paradigm for structured and semi-structured datasets, such as those aggregated by the ADS. Previous work on the Archaeobrowser demonstrated that a faceted classification approach to large datasets and the associated facet classification browser result in significantly more intuitive, usable, complete and reliable searching. The Archaeotools project delivers the first UK service implementation of a faceted classification tree and associated browser in archaeology. This is specifically intended to enhance the ADS's ArchSearch facility with richer data resources and to transform our users' primary search approach away from the vagaries of a Google style type-and-hope free text search model towards a more intuitive and informative system.

The solutions to the two broad issues outlined above, automatic metadata extraction and browsing by facet, are, in fact, extremely complementary. It is the Archaeotools implementation of these solutions together that offers such potential. Not only it is intended that the faceted classification browser 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 that, despite their importance, have hitherto been either impossible or very difficult to integrate into existing digital frameworks.

Although this project focuses on a specifically archaeological context, it builds on NLP work on historical sources and biomedical informatics and uses software approaches developed for engineering purposes with Rolls Royce (http://nlp.shef.ac.uk/wig/research/IPAS.html; Ciravegna *et al.* 2006). Many other disciplines have problems either directly analogous or very similar to the problems that exist with grey literature in archaeology, i.e. a body of literature that is hard to access, but which represents a significant resource. An indication of how broad this problem is can be gleaned from, for example, the American College and Library Association's listing of Internet resources for 'grey' literature (http://www.ala.org/ala/mgrps/divs/acrl/publications/crlnews/2004/mar/graylit.cfm; Mathews 2004). It is likely that a successful approach to tackling this problem developed via e-Science funding in an archaeological context will have much broader application in arts, humanities and sciences.

2. Summary of Archaeotools objectives

The Archaeotools project is following a trajectory that should allow it to reach the goals outlined above in three more or less discrete stages:

(i) The creation of an advanced faceted classification and geospatial browser. The underlying dataset comprises over 1×10^6 records (held in an Oracle Relational Database Management System) aggregated from the National Monuments Records of Scotland, Wales and England, as well as Historic Environment Records from numerous local authorities and the ADS's own archive holdings. The facets selected will be standard hierarchical 'What', 'Where' and 'When' facets plus a 'Media' facet to allow the selection of particular subsets of resources. The facets are populated from existing thesauri (e.g. the Thesaurus of Monument types) in extensible markup 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-based (http://www.w3.org/2004/02/skos/) Web services developed by the AHRC-funded Semantic Tools for Archaeology project (http://hypermedia.research.glam.ac.uk/kos/star/) based at the University of Glamorgan.

(ii) The creation of a reusable NLP system that will automatically extract resource discovery metadata (and other facet types) from unpublished archaeological reports.

(iii) The extension of the NLP systems to capture metadata from legacy historical documents, using the PSAS as an exemplar corpus and using the University of Edinburgh's geoXwalk service to recast place names and locations extracted from text as national grid references (NGRs), allowing enhanced geospatial searching of the data (EDINA 2008).

3. The faceted classification browser: ArchSearch III

The current search mechanism and interface to the ADS's aggregated datasets is called ArchSearch II, having evolved from the ADS's original ArchSearch mechanism developed in the late 1990s. The Archaeotools project is designed to develop this search mechanism into a faceted classification browser and associated interactive geospatial search. The faceted classification approach to presenting structured datasets is increasingly common in the commercial Web, but clearly lends itself to the discovery of any structured dataset (Denton 2003). A faceted query engine has been employed by a team at Columbia University to provide an interface to archaeological finds datasets (Ross *et al.* 2005, 2007), and in the Open Context system at the Alexandria Archive Institute (http://www.opencontext.org/), but applications of faceted classification are still rare in archaeology.

Previous work carried out on faceted classification by the ADS in the Archaeobrowser project (Jeffrey *et al.* 2008) demonstrated that the most appropriate search facets for archaeological datasets are as follows:

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?
Media—what is the form of the record you are ultimately interested in?

These are far from being the only possible facets and some others can be seen as highly desirable (e.g. Who—to whom does the record relate?), but as a matter of practicality, these four are the facets that are expected to offer the greatest use for the archaeological researcher. An investigation of how additional facets might be specified and whether user-generated facets are either desirable or feasible is included within the Archaeotools project.

In order to facilitate browsing, each facet needs to have an associated ontology, expressed as a hierarchy of terms. Fortunately, in the historic environment sector, there are hierarchical thesauri deployed or under development that allow a browsing structure to be populated for each facet, apart from Media. These thesauri, or controlled word lists, 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, and each is either already being broadly used or is in the process of being adopted. For the above facets, the following three thesauri were selected:

What—the Thesaurus of Monuments types (TMT; English Heritage 2008).
When—Monument Inventory Data Standard (MIDAS) period list (http://www.midas-heritage.info) and Forum for Information Standards in Heritage (http://www.fish-forum.info/index.htm).
Where—county, district, parish (UK Government list of administrative areas).

An example of how the hierarchical structure looks for a detailed record of the monument type 'Tower Keep' might be

$$\text{What} \rightarrow \text{Defence} \rightarrow \text{Castle} \rightarrow \text{Keep} \rightarrow \text{Tower Keep}.$$

This example shows that the hierarchical structure lends itself to a 'point-and-click' browsing approach, such that each level of the hierarchy can be expanded or collapsed by a mouse click. Each record in the target dataset is assigned a What, When and Where value from the selected thesauri. The power of this approach for a normalized 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 1×10^6 or so records to a selected set of 16 records representing Bronze Age funerary monuments, within 5 km of a specific location in North Yorkshire with just three or four clicks of the mouse. Not only does this compare very favourably to traditional search box-based techniques, but the fact that the data have been mapped to the terms of the thesaurus means that the user can have a much higher level of confidence in the completeness of the returned results and is

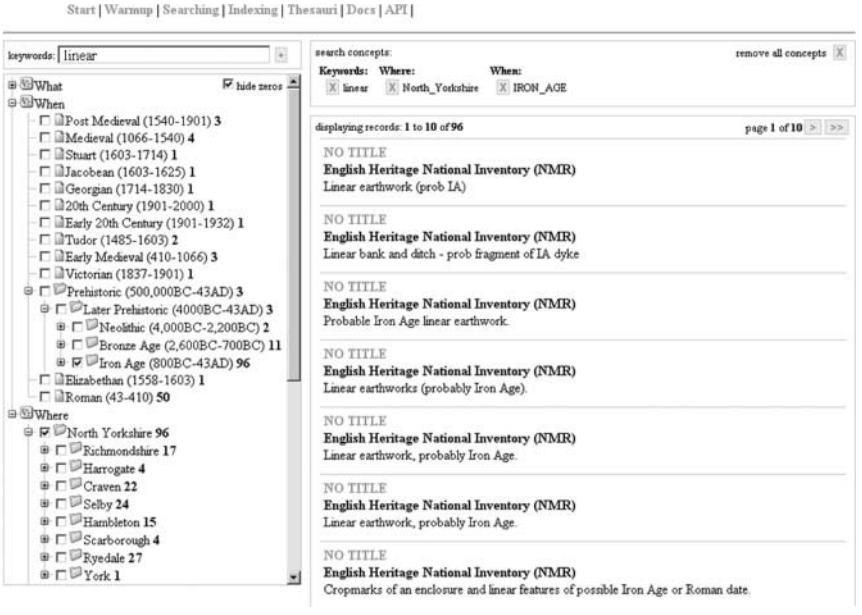

Figure 1. A screen shot of the prototype faceted classification browsing tree for the Archaeotools project. The window on the left shows the What, When and Where ontologies, with When and Where trees expanded to the first level. Iron Age linear features in North Yorkshire have been selected. The numbers in bold after each term indicate the number of records classified according to that facet. The window on the right displays the first 10 records out of the 96 returned by this query.

much less troubled by the return of false positive results. The indexing mechanism adopted by the Archaeotools project was built on top of Solr, an open-source enterprise search server based on the Lucene Java search library (http://lucene.apache.org/solr/).

Figure 1 is a screen shot of the draft faceted classification browsing tree for the Archaeotools project as it is currently implemented. The browser interface shows the number of records associated with each facet, as well as allowing the user to hide nodes that have no associated records. This feature facilitates easier navigation by cutting down screen clutter through the hiding of negative results.

Any large monument inventory, indeed any large dataset, especially one that has developed over a number of years is unlikely to conform perfectly to any rigid terminology standards, especially if these were created subsequent to the inception of the dataset. The Archaeotools project is the first instance of any archaeological project in the UK that has both generated metrics on these mismatches and mitigated the problem via a combined automated and manual approach. This mitigation generated interesting statistics that are summarized in table 1. It is true to say that all datasets contributed to these mismatches more or less equally, and that there was no obvious dataset where the terminology used diverged more radically from the thesauri than all the others.

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

Table 1. Ontology mismatches between record entries and the appropriate thesauri (from Jeffrey *et al.* in press).

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 (6442 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%)

The figures for Where with terms not found in the county/district/parish (CDP) list (24.5%) can be safely ignored, as these figures were generated prior to the integration of the Scottish CDP list into the thesauri set; this comfortably accounts for the majority of these missing terms.

Contrary to our original expectations, it has proven possible to completely map these record sets to the thesauri, and therefore the facets, by a combination of automatic rule-based expressions and manual techniques. The When facet provides an example for the success of this combined approach. There is a large 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. Most of these were mapped directly to MIDAS-defined date ranges. Analysis recovered 457 types of irresolvable dates. After automated processing using regular expressions, however, this ultimately equated to only approximately 700 records. This is a perfectly manageable number to expect to be corrected by manual intervention.

4. NLP and IE

NLP is the branch of artificial intelligence concerned with extracting meaning from human speech and text. Ontology-based text annotation is seen as making a key contribution to the development of the Semantic Web (Uren *et al.* 2006). However, work has focused on commercial applications; there have been few research-driven projects, and fewer still in the arts and humanities. Archaeology has some potential as a test bed in this field because, despite its humanities-based focus, it has a relatively well-controlled vocabulary. Amrani *et al.* (2008) have reported on a pilot application in a relatively specialized area; the Archaeotools project aims to employ NLP across a range of archaeological texts.

Figure 2 shows the process architecture adopted for the Archaeotools project. In brief, selected fields are extracted from the ADS Oracle database in MIDAS XML format data, converted to a resource description framework (RDF) format.

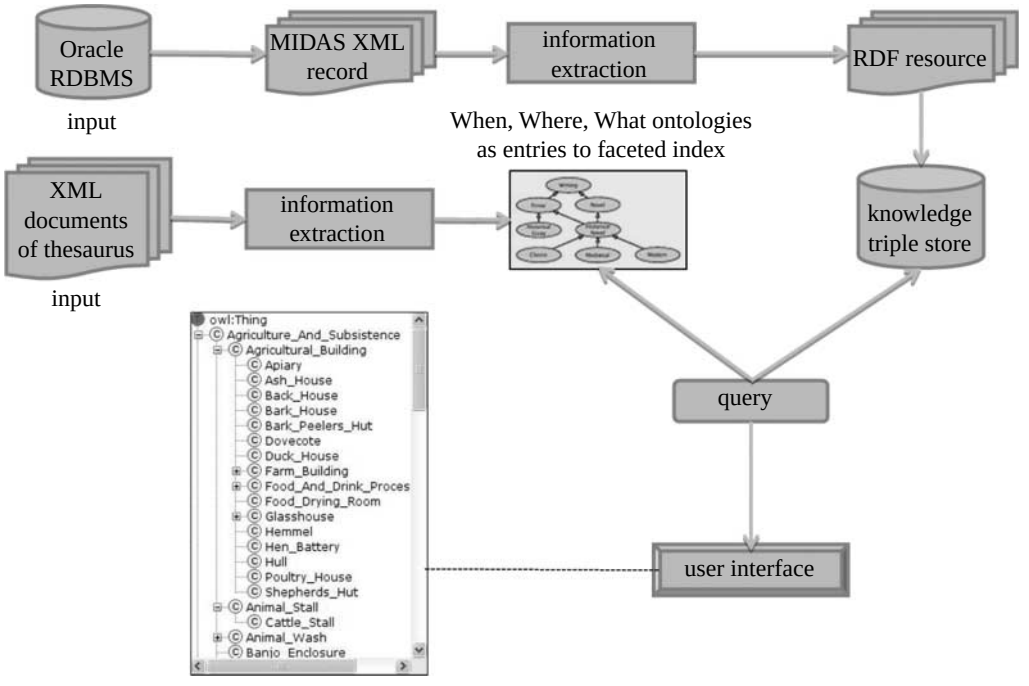

Figure 2. The Archaeotools process architecture RDBMS is the Relational Database Management System.

XML (Ontology Web Language) versions of the thesaurus are extracted to create workable ontologies, and these in tandem with the RDF knowledge triple store are queried to classify the records (http://www.w3.org/TR/owl-features/).

IE is the process of automatically extracting structured information from unstructured natural language texts (Cowie & Wilks 2000). One of the processes key to IE is the application of NLP technologies. NLP is the analysis of human language to enable computers to discern semantic meaning in natural languages. The outputs of NLP are linguistic data that are crucial to IE tasks; including sentence boundaries, part-of-speech tags and grammar parsing. Conversely, IE usually requires human input to define (via templates) the general form of the information to be extracted; these templates then guide the extraction process.

Typical IE tasks include the following:

(i) Terminology extraction—identification of relevant terms for a given corpus, e.g. identifying the most relevant terms for an archaeology corpus or dataset, such as the ADS grey literature holdings.
(ii) Named-entity recognition (NER)—identification of entities in a document, such as archaeological period terms, parish names, district names, archaeological findings and so on.
(iii) Fact extraction—identification of facts, which could be statements of relationships between entities, e.g. link each identified archaeological find spot to identified parish names, thus constructing a relationship of the form ‘artefact-found-at-place’.

The Archaeotools IE tasks fall under NER and fact extraction. The first objective is to extract the following types of information units from a corpus of over 1000 unstructured archaeological grey literature reports, such that this corpus can be indexed and searched by a number of attributes, including the predefined facets:

Subject (topics covered, findings mentioned)—mapped to the What facet.
Location (place names related to events and findings)—mapped to the Where facet.
Temporal (temporal information related to findings)—mapped to the When facet.
Grid reference—mapped to the Where facet.
Report title, creator, publisher, publisher contact, publication date.
Event dates.
Bibliography and references.

In addition, IE aims to discover relationships between certain types of information, for example the relationships between archaeological finds and period terms, thus to enable semantic searches such as 'sites where *Roman pottery was* found'.

There are two basic approaches to the design of IE systems, the knowledge engineering (KE) approach and the automatic training (AT) approach (Appelt & Israel 1999).

In the KE approach, an IE expert and a domain expert manually read through a moderate-size domain corpus, while the domain expert identifies information units to be extracted, and the IE expert identifies and translates the textual patterns into formal programming rules. Next, the rules are applied to several corpora, and the extracted information is examined to see where the rules under- and overgenerate results, revising the rules accordingly. The IE expert's skills play a critical role in building working systems. An example from Archaeotools might be an IE system for extracting information about the publishers of archaeological reports, a sample rule can be as simple as 'the first organization that appears following the report title, and is a registered name on the Institute of Field Archaeologists list'. A disadvantage of this approach is that, in this example, the rule will not work for any unregistered organizations.

In the AT approach, it is not necessary to have detailed knowledge of IE systems and rule formalism. On the contrary, the most difficult rule-induction process is handled by the machine. Typically, domain experts are required to produce adequate volumes of sample annotations—usually a subset of the entire corpus—which are tagged to mark expected information units to 'train' the IE system; and then specify features that are likely to discriminate these sample annotations from unannotated sections of documents. Examples of features could be text units, generic entity types (person, organization, location, etc.), existence in gazetteers or dictionaries, position in the document and so on. Next, an IE algorithm is run on the training corpus, consuming the selected features and producing a model that stores generalized rules to be applied to novel texts.

The advantages of the KE over the AT approach are that there is no need to prepare training data, which, in the case of Archaeotools, has proved to be a time consuming and laborious task. Also, in situations where information occurs in regular and limited patterns and contexts, it is easy to develop systems that perform well. However, the KE approach itself requires an extensive amount of manual input. Porting systems to different domains is difficult, as rules are often context and domain specific, and thus porting usually requires a system rebuild. Furthermore, when information to be extracted is diverse, such as artefacts in archaeology, which may occur anywhere within any context in a document, the task can become extremely difficult. By contrast, the AT approach has better domain portability. Porting IE systems to a different domain is relatively straightforward, only requiring the rebuilding of a training corpus and feature tuning, and costs far less than rebuilding an IE system. In addition, the AT system handles diversity well and can be applied to large-scale datasets. The main drawbacks of the AT approach is that training data can be expensive to build (due to it being time consuming). On the other hand, feature selection is equally as important as the learning algorithm for a system that performs well, although, in many cases, feature tuning can also be time consuming.

Both KT and AT approaches have been employed in this project, depending on the form of the information being tackled. The KE approach is applied to information that matches simple patterns, or occurs in regular contexts, such as NGRs and bibliographies; the AT approach is applied to information that occurs in irregular contexts and cannot be captured by simple rules, such as place names, temporal information, event dates and subjects. In addition, both approaches have been combined to identify report title, creator, publisher, publication dates and publisher contacts. While the development of the NLP aspects of the project are still ongoing, the positive results of the above approaches already demonstrate that the objective of automatically extracting resource discovery metadata from grey literature, and not only making it discoverable, but making it discoverable by the classification of its metadata into previously defined facets, is achievable. The service version of the faceted classification browser, linking directly to grey literature, will be online as an interface option with the ADS from spring 2009.

The final challenge for the Archaeotools project is to refocus the NLP-automated metadata extraction process from semi-structured grey literature to the almost entirely unstructured digitized version of the PSAS held as an archive of PDF files by the ADS. Clearly, automatically extracted data from these journals 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.

One exciting prospect 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). Extracted place names can be sent directly to this service and the service will automatically return NGRs for that place name, thus allowing the relevant place name from PSAS to be mapped in the Archaeotools geospatial interface, and therefore make them as discoverable and searchable as standard monument inventory datasets.

5. Conclusion

The Archaeotools project has reached its objective of successfully implementing a faceted classification browsing system in the context of aggregated archaeological records. This service will be released for public access as a replacement for the existing ArchSearch II. During the process of preparing the datasets for classification, useful insights have been gained into the level of vocabulary control within archaeological monument inventories. Although work has had to be done regarding the apparent mismatch between the seemingly loose terminology of the historical datasets and the rigorous word lists, thesauri and ontologies, in practice, a combination of automated and manual approaches allowed for the classification process to be both comprehensive and meaningful. The classification and data cleaning process itself can be seen essentially as a single operation, as existing records are rarely changed. There may be over 1×10^6 records in the datasets, but these are added to at a fairly slow rate (approx. 5000 per annum), meaning that future mismatches or missing facets are much more likely to be in small and manageable volumes. This has also given the ADS the unexpected benefit of being able to report back to donor organizations, not just on the level of data cleansing required, but pinpointing the specific records and problematic fields.

The other two major components of the project, automated data and metadata extraction from grey literature and legacy literature are now where the main focus of our work lies. The combined attack on the data using KE and AT IE has already proved successful to the extent that we are confident that automated resource discovery metadata extraction can be achieved for grey literature at least. This removes a major obstacle to efforts in digitizing the huge backlog of this material, hopefully leading to the unlocking of its potential to significantly influence the development of archaeological theories in the future. If similar levels of IE can be achieved for antiquarian literature, for example the PSAS material, then it will truly be possible to integrate a broad range of data types within a single search interface, greatly empowering the researcher of the future.

References

Amrani, A., Abajian, V., Kodratoff, Y. & Matte-Tailliez, O. 2008 A chain of text-mining to extract information in archaeology. In *Information and communication technologies: from theory to applications, ICTTA 2008, 3rd Int. Conf.*, pp. 1–5.

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

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

Ciravegna, F., Lanfrachi, V., Moore, P., Baghdev, R. & Iria, J. 2006 Automatically annotating jet engine event reports using information extraction. In *Proc. Knowledge and Information Management: the Challenge of Through Life Support Seminar, Institution of Mechanical Engineers, London, 26 September 2006.*

Cowie, J. & Wilks, Y. 2000 Information extraction. In *Handbook of natural language processing* (eds R. Dale, H. Moisl & H. L. Somers), pp. 241–260. Boca Raton, FL: CRC Press.

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.

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

English Heritage 2008 NMR thesaurus browser. See http://thesaurus.english-heritage.org.uk/thesaurus.asp?thes_no=1.

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.

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.

Hardman, C. & Richards, J. D. 2003 OASIS: dealing with the digital revolution. In *Digital heritage of archaeology. Computer Applications and Quantitative Methods in Archaeology 2002* (eds M. Doerr & A. Sarris), pp. 325–328. Athens, Greece: Archive of Monuments and Publications Hellenic Ministry of Culture.

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.

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. Szeverényi).

Lock, G. 2008 A professional mockery. *Br. Archaeol.* **101**, 36–37.

Mathews, B. 2004 Gray literature: resources for locating unpublished research. *Coll. Res. Libr. News* **65**, 198–201.

Richards, J. D. & Hardman, C. S. 2008 Stepping back from the trench edge. An archaeological perspective on the development of standards for recording and publication. In *The virtual representation of the past* (eds M. Greengrass & L. Hughes), pp. 427–445. London, UK: Ashgate.

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.

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.

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E. & Ciravegna, F. 2006 Semantic annotation for knowledge management: requirements and a survey of the state of the art. *J. Web Semant.* **4**, 14–28. (doi:10.1016/j.websem.2005.10.002)