

This is a repository copy of *Text Mining in Archaeology: Extracting Information from Archaeological Reports*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/86788/>

Version: Published Version

---

**Book Section:**

Richards, Julian D [orcid.org/0000-0003-3938-899X](https://orcid.org/0000-0003-3938-899X), Tudhope, Douglas and Vlachidis, Andreas (2015) *Text Mining in Archaeology: Extracting Information from Archaeological Reports*. In: Barcelo, Juan and Bogdanovic, Igor, (eds.) *Mathematics in Archaeology*. CRC Press , Florida , pp. 240-254.

<https://doi.org/10.1201/b18530-15>

---

**Reuse**

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

# 12

## Text Mining in Archaeology: Extracting Information from Archaeological Reports

*Julian D. Richards,<sup>1,\*</sup> Douglas Tudhope<sup>2</sup> and Andreas Vlachidis<sup>3</sup>*

---

### Introduction

Archaeologists generate large quantities of text, ranging from unpublished technical fieldwork reports (the ‘grey literature’) to synthetic journal articles. However, the indexing and analysis of these documents can be time consuming and lacks consistency when done by hand. It is also rarely integrated with the wider archaeological information domain, and bibliographic searches have to be undertaken independently of database queries. Text mining offers a means of extracting information from large volumes of text, providing researchers with an easy way of locating relevant texts and also of identifying patterns in the literature. In recent years, techniques of Natural Language Processing (NLP) and its subfield, Information Extraction (IE), have been adopted to allow researchers to find, compare and analyse relevant documents, and to link them to other types of data. This chapter introduces the underpinning mathematics and provides a short presentation of the algorithms used, from the point of view of artificial intelligence and computational logic. It describes the different NLP schools of thought and compares the pros and cons of rule-based vs. machine learning approaches to IE. The role of ontologies and named

---

<sup>1</sup> Department of Archaeology, University of York, The King’s Manor, York, YO1 7EP, UK.  
Email: [julian.richards@york.ac.uk](mailto:julian.richards@york.ac.uk)

<sup>2</sup> Hypermedia Research Unit, Faculty of Computing, Engineering and Science, University of South Wales, Pontypridd, CF37 1DL, Wales, UK.  
Email: [douglas.tudhope@southwales.ac.uk](mailto:douglas.tudhope@southwales.ac.uk)

<sup>3</sup> Hypermedia Research Unit, Faculty of Computing, Engineering and Science, University of South Wales, Pontypridd, CF37 1DL, Wales, UK.  
Email: [andreas.vlachidis@southwales.ac.uk](mailto:andreas.vlachidis@southwales.ac.uk)

\* Corresponding author

entity recognition is discussed and the chapter demonstrates how IE can provide the basis for semantic annotation and how it contributes to the construction of a semantic web for archaeology. The authors have worked on a number of projects that have employed techniques from NLP and IE in Archaeology, including Archaeotools, STAR and STELLAR and draw on these projects to discuss the problems and challenges, as well as the potential benefits of employing text mining in the archaeological domain.

## Background

Easy access to the information locked within texts is a significant problem for the archaeological domain, in all countries. In the UK, there is an average of 6,000 interventions per annum, and there are equivalent figures for other European countries, varying only according to the extent of the legal requirement for intervention prior to development, and whether it is undertaken as a state-led operation (as in France, Germany, Greece and Italy) or commercial enterprise (as in the UK and Netherlands). In the US, in the order of 50,000 field projects a year are carried out by federal agencies under these mandates, with another 50,000 federal undertakings requiring record searches or other inquiries that do not result in fieldwork. However, there is no legal requirement to publish the outcome of all this activity, either in the USA or Europe.

On both sides of the Atlantic, therefore, this activity generates vast numbers of reports that together constitute the unpublished 'grey literature' whose inaccessibility has long been an issue of major concern. With so much work being performed and so much data being generated, it is not surprising that archaeologists working in the same region do not know of each others' work, let alone archaeologists working in different continents. Decisions about whether to preserve particular sites, how many sites of specific types to excavate, and how much more work needs to be done are frequently made in an informational vacuum. Furthermore new data is not fed into the research cycle and academic researchers may be dealing with information which is at least 10 years out-of-date. Nonetheless, the fact that such reports are not fully published should not be taken to suggest that the value of the archaeological data or interpretation is not significant enough for publication (Falkingham 2005).

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 recognised by the academic community. Bradley (2006) has questioned why it is not more widely available, and several research projects have been undertaken specifically to attempt to synthesise the outcomes of development control archaeology from the grey literature (Fulford and Holbrook 2011). Digital collection and online delivery of both newly created (i.e., 'born digital') and legacy material could provide a solution to addressing these access issues. However, good access is predicated on good discovery mechanisms and these rely, amongst other things, on good data about data, or metadata.

In the UK the Archaeology Data Service (ADS) actively gathers digital versions of grey literature fieldwork reports as part of the OASIS project <<http://www.oasis.ac.uk/>> (Hardman and Richards 2003, Richards and Hardman 2008). The ADS grey literature library currently (as of June 2014) comprises over 26,000 reports although it is increasing at the rate of 200 per month. The Dutch e-depot for Archaeology, managed by DANS, also holds over 20,000 reports. In the UK all reports can be downloaded free of charge and there is a high level of usage. In collaboration with the British Library and Datacite each report is assigned a Digital Object Identifier, ensuring a permanent means of citation.

Each of the reports also has manually generated resource discovery metadata covering such attributes as author, publisher, and temporal and geospatial coverage, adhering to the Dublin Core metadata standard <<http://dublincore.org/>>. Generating metadata this way may be feasible, if time-consuming, where it is created simultaneously with the report's deposit with the ADS but it is not a feasible means of dealing with the tens of thousands of legacy reports known to exist. For any attempt to digitise these disparate and distributed sets of records to facilitate broader access, the key in terms of both cost and time would be automated metadata generation.

Within the ADS digital library there are also electronic versions of more conventional journals and reports, including a complete back run of the *Proceedings of the Society of Antiquaries of Scotland* (PSAS) going back to 1851 <<http://dx.doi.org/10.5284/1000184>>. Many of the same indexing issues arise with reference to digitised versions of such early or short run published material. As an increasing number of journal back-runs are digitised, and held within large online libraries such as JSTOR, or by smaller discipline-specific repositories such as the ADS, providing deeper and richer access to these resources becomes an increasing priority. Whilst the ADS repository is accessible to Google and other automated search engines these provide only free text indexing, regardless of any domain-specific controlled vocabularies, and they do not allow researchers to situate a specific term within the wider set of concepts implicit within a hierarchical thesaurus, identifying 'round barrow' and 'long barrow' as sub-types of barrow, for example, and even situating them as specific types of funerary monument. Such literal string match searches are also susceptible to large numbers of false positives, recovering 'Barrow' as a place name, or barrow as a wheelbarrow, for instance. Several research projects are now undertaking text mining on large quantities of published text in order to identify intellectual trends (Michel et al. 2011). Furthermore, in archaeology, there is the potential to provide joined-up access to published and unpublished literature within a single interface, allowing users to cross-search both types of resource. However, indexing of journal back-runs 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, whilst 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. 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.

## Mathematical Methods of Natural Language Processing

### *Statistical Information Retrieval*

Information Retrieval (IR) is the activity of finding relevant information resources to satisfy specific user queries originating from generic information needs. The automatic definition of representative document abstractions (metadata or index terms) and ranking of search results is used by many different retrieval models that have been introduced in the last decades, including the Boolean model, vector space models and probabilistic models (Baeza-Yates and Ribeiro-Neto 1999). The Boolean model enables users to seek for information using precise semantics that are joined by the Boolean operators. Thus, a query is conventional Boolean expression composed of index terms linked by three operators: AND, OR, NOT, for example  $[q = K_a \wedge (K_b \vee \neg K_c)]$ . The model predicts that the document  $d_j$  is relevant to the query  $q$  if  $\text{sim}(d_j, q) = 1$ . The model supports clean formalisms but the exact match

condition may lead to retrieval of too few or too many, relevant to the Boolean expression but irrelevant to user information needs, documents.

The vector model acknowledges the fact that binary weights are limited and does not provide a means for partial matching of user queries (Moens 2006). To overcome this, the vector model calculates the degree of similarity between document and query vector. The similarity between the query vector  $\vec{q} = (w_{1,q}, w_{2,q}, \dots, w_{t,q})$  where  $t$  is the total number of index terms in the system and the vector for a document  $d_j$ , represented by  $\vec{d}_j = (w_{1,j}, w_{2,j}, \dots, w_{t,q})$  is based on the quantification of the cosine of the angle between those two vectors. Since  $w_{i,j} \geq 0$  and  $w_{i,q} \geq 0$  the similarity  $sim(q, d_j)$  varies from 0 to +1 and the model ranks the documents according to their degree of similarity to the query. An established threshold could dictate the retrieved matches based on the degree of similarity between query and document which might be equal to or over a given cut-off point. In this way partial match retrieval is achieved and results can be presented in a ranked order.

The vector model employs two distinct factors namely, the (*tf*) and the (*idf*) factor to provide a means of assigning weight to indexed terms. The frequency that a term  $ki$  appears within a document is known as intra-cluster similarity or *tf* and determines how well a term  $i$  is representative of the document contents. Inter-cluster dissimilarity or *idf* is a measurement of the inverse frequency of term  $i$  among the documents in collection. A term appearing in almost every document in a collection is not very useful for distinguishing a relevant document from a non-relevant one.

The probabilistic IR model is based upon the assumption that there is an ‘ideal answer set’ which contains exactly the relevant documents for a given user query. Knowing the description and attributes of an ‘ideal answer set’ will lead us to successful retrieval results. The probabilistic model defines a query  $q$  as a subset of index terms and  $R$  the set of documents known (or initially guessed) to be relevant. If  $\sim R$  is the compliment of  $R$  and the set of non-relevant documents, then the probability  $P$  of the document  $Dj$  being relevant to the query  $q$  is defined as  $P(R|Dj)$  (Baeza-Yates and Ribeiro-Neto 1999).

A number of variations to these IR models aim to improve and enhance their performance. The above models operate on the assumption that index terms are mutually independent and none of them acknowledge dependencies between index terms. This may result in poor retrieval performance since relevant documents not indexed by any of the query terms are not retrieved and irrelevant documents indexed with the query terms are retrieved (Smeaton 1997). The *Latent Semantic Indexing* model proposes a solution to this problem by enhancing the vector based model and matching each document and query vector to a lower dimensional space of concepts enabling concept based matching.

## Information Extraction

Information Extraction (IE) is a specific NLP technique defined as a text analysis task which extracts targeted information from context (Cowie and Lehnert 1996, Gaizauskas and Wilks 1998, Moens 2006). It is a process whereby a textual input is analysed to form a textual output capable of further manipulation. Information extraction systems fall into two distinct categories; Rule-Based (hand-crafted) and Machine Learning systems (Feldman et al. 2002). During the seven Machine Understanding Conferences (MUC), the involvement of rule-based information extraction systems has been influential. The issue of information systems portability quickly gained attention and during MUC-4 the Machine Learning applications introduced a semi-automatic technique for defining information extraction patterns as a way of improving a system’s portability to new domains and scenarios (Soderland et al. 1997).

### **Rule-based Information Extraction Systems**

Rule-based systems consist of cascaded finite state transducers that process input in successive stages. Dictated by a pattern matching mechanism, such systems are targeted at building abstractions that correspond to specific IE scenarios. Hand-crafted rules make use of domain knowledge and domain-independent linguistic syntax, in order to negotiate semantics in context and to extract information for a defined problem. It is reported that rule-based systems can achieve high levels of precision of between 80–90 percent when identifying general purpose entities such as ‘Person’, ‘Location’, and ‘Organisation’ from financial news documents (Feldman et al. 2002, Lin 1995).

The definition of hand-crafted rules is a labour intensive task that requires domain knowledge and good understanding of the IE problem. For this reason rule-based systems have been criticised by some as being costly and inflexible, having limited portability and adaptability to new IE scenarios (Feldman et al. 2002). However, developers of rule-based systems claim that, depending on the IE task, the linguistic complexity can be bypassed and a small number of rules can be used to extract large sets of variant information (Hobbs et al. 1993). In addition, rule-based systems do not require training for delivering results in contrast to the supervised machine learning system discussed below.

### **Machine Learning Information Extraction Systems**

The use of machine learning has been envisaged to provide a solution capable of overcoming the potential domain-dependencies of rule-based IE systems (Moens 2006, Ciravegna and Lavelli 2004). Learning in the Artificial Intelligence context describes the condition where a computer programme is able to alter its ‘behaviour’, that is, to alter structure, data or algorithmic behaviour in response to an input or to external information (Nilsson 2005).

Machine learning strategies can support supervised and unsupervised learning activities. The supervised learning process is based upon a training dataset annotated by human experts, which is used by the machine learning process to deliver generalisations of the extraction rules, which are then able to perform a large scale exercise over a larger corpus. It is argued that it is easier to annotate a small corpus of training documents than to create hand-crafted extraction rules, since the latter requires programming expertise and domain knowledge (Moens 2006). On the other hand, the size of the training set may depend on the range and complexity of the desired annotations and the characteristic language use in the domain.

During unsupervised learning, human intervention is not present and the output of the training data set is not characterised by any desired label. Instead a probabilistic clustering technique is employed to partition the training dataset and to describe the output result, with subsequent generalisation to a larger collection (Nilsson 2005). Unsupervised IE is very challenging and so far such systems have been unable to perform at an operational level (Uren et al. 2006, Wilks and Brewster 2009).

## **Information Extraction Evaluation**

The evaluation of IE systems was established by the Machine Understanding Conference, MUC-2. Two primary measurements adopted by the conference, *Precision* and *Recall*, originated from the domain of Information Retrieval but were adjusted for the task of IE (template filling). According to the MUC definition, when the answer key is  $N_{key}$  and the

system delivers  $N_{\text{correct}}$  responses correctly and  $N_{\text{incorrect}}$  incorrectly then  $Recall = \frac{N_{\text{correct}}}{N_{\text{key}}}$  and  $Precision = \frac{N_{\text{correct}}}{N_{\text{correct}} + N_{\text{incorrect}}}$ .

The formulas examine a system's response in terms of correct or incorrect matches. This binary approach does not provide enough flexibility to address partially correct answers. A slightly scalar approach can be adopted to incorporate the partial matches. In this case, the above formulas can be defined as

$$Recall = \frac{N_{\text{correct}} + \frac{1}{2} \text{Partial Matches}}{N_{\text{key}}}, \quad Precision = \frac{N_{\text{correct}} + \frac{1}{2} \text{Partial Matches}}{N_{\text{correct}} + N_{\text{incorrect}} + \text{Partial Matches}}$$

Partial matches are shown weighted as 'half' matches. The value of the weight can change if partial matches seem more or less important.

The weighted average of Precision and Recall is reflected by a third metric, the F-measure score. When both Precision and Recall are deemed equally important then we can use the equation:  $F_1 = 2 \frac{Precision * Recall}{Precision + Recall}$ . Attempts to improve Recall will usually cause Precision to drop and vice versa. High scoring of  $F_1$  is desirable since the measure can be used to test the overall accuracy of the system (Maynard et al. 2006).

### Gold Standard Measures

The Gold Standard (GS) is a test set of human annotated documents describing the desirable system outcome. An erroneous GS definition could distort the results of the evaluation and lead to false conclusions. Problematic and erroneous GS definition is addressed by enabling multiple annotations per document. The technique allows more than one person to annotate the same text in order to address discrepancies between different annotators. To calculate the agreement level between individual annotators the technique employs the Inter Annotator Agreement (IAA) metric (Maynard et al. 2006).

Manual annotation of archaeological documents is influenced by domain characteristics and embedded language ambiguities that challenge IAA scores. Such ambiguities concern the definition of domain entities; for example, the fine distinction between physical object and material entities, application of annotation boundaries and inclusion of lexical moderators. Typical IAA scores in an archaeological context range from between 60 and 80 percent (Byrne 2007, Vlachidis and Tudhope 2012, Zhang et al. 2010). In the case of a low IAA score, a final and explicit GS set is proposed by a human 'Super Annotator' who acts as a conciliator between individual annotation sets, reviewing the cases of disagreement and choosing the correct annotation (Savary et al. 2010). Normally the Super Annotator is a field expert with the experience and knowledge to reconcile individual annotation discrepancies, although it must be remembered that there may be underlying variation in language use and terminology within a domain.

### Previous Work

Archaeology has excellent potential for the deployment of text mining because, despite its humanities focus, it has a relatively well-controlled vocabulary. Significant effort has been put into the development of controlled word lists or thesauri, including the UK MIDAS

data standard (English Heritage 2007, Newman and Gilman 2007). However the nature of archaeological vocabulary poses some challenges in that, unlike highly specialised scientific domains with a unique vocabulary (e.g., biological or medical terms), much archaeological terminology consists of common words in an everyday sense, for example ‘pit’, ‘well’. There is also the distinction between descriptions of the present and the archaeological past (for example, the term ‘road’ has much more significance if it is a ‘roman road’).

Within the last ten years a number of projects have attempted to deploy text mining on archaeological texts, with a specific focus on the grey literature. Amrani et al. (2008) reported on a pilot application in a relatively specialised area of archaeology; the *OpenBook* project experimented with Memory Based Learning in extracting chronological and geographical terms from Dutch archaeological texts (Pajmans and Wubben 2008) and Byrne has also explored the application of NLP to extract event information from archaeological texts (Byrne and Klein 2010). The present authors have worked on two major projects that employed different methods of IE and these provide useful case studies of text mining in Archaeology. Archaeotools largely adopted a machine learning approach, whilst OPTIMA (which provided the basis of the STAR and STELLAR projects) adopted a rule-based approach. Both are described below.

## The Archaeotools Project

In the UK, the Archaeotools project, a collaboration between the ADS and the University of Sheffield Computer Science OAK group, provided a major opportunity to deploy text mining in Archaeology (Jeffrey et al. 2009, Richards et al. 2011). The first objective of Archaeotools was to extract several types of information from a corpus of over 1000 unstructured archaeological grey literature reports, so that this corpus could be indexed and searched by a number of attributes, including subject, location, and period. These support the standard ‘What’, ‘Where’ and ‘When’ queries that underlie a broad range of archaeological research questions. The project employed a combination of a rule-based (KE) and an Automatic Training (AT) approach. The rule-based approach was applied to information that matched simple patterns, or occurred in regular contexts, such as national grid references and bibliographies. In order to deploy the AT approach the ADS staff, all of whom are archaeologically trained, carried out extensive annotation exercises on a subset (c.150 reports) of the grey literature corpus. The AT approach was applied to information that occurred in irregular contexts and could not be captured by simple rules, such as place names, temporal information, event dates, and subjects. In addition, both approaches were combined to identify report title, creator, publisher, publication dates and publisher contacts.

### *Text Mining Applied to Grey Literature*

Relatively high levels of success were achieved when the above techniques were applied to the sample of 1000 semi-structured grey literature reports. By removing files which could not be converted to machine readable documents due to file formatting issues, this left a working sample of 906 reports.

The greatest problem encountered was that of distinguishing between ‘actual’ and ‘reference’ terms. As well as the ‘actual’ place name referring to the location of the archaeological intervention, most grey literature reports also refer to comparative information from other sites, here called ‘reference’ terms. The IE software returned all place names in the document, masking the place name for the actual site amongst large



numbers of other names. However this was solved by adopting the simple rule that the primary place name would appear within the ‘summary’ section of the report. If it was not possible to identify a summary then the first ten percent of the document was used instead. Out of 1000 reports, this left 162 documents where it was not possible to identify a place name in the summary or first ten percent of the report.

**Table 1.** ‘Actual’ identifications for 1000 grey literature reports.

	No data	
What	159	17.5%
Where	162	17.9%
When	263	29.0%

However, for the documents as a whole there were only 17 where it was not possible to identify the ‘What’ facet, 20 with no ‘Where’ information, and 40 where it was impossible to identify a ‘When’ term.

**Table 2.** ‘Reference’ identifications for 1000 grey literature reports.

	No data	
What	17	1.9%
Where	20	2.2%
When	40	4.4%

Although these figures do not guarantee that the terms identified were meaningful, so long as users are shown why a document has been classified according to those terms they represent acceptable levels of classification.

### ***Text Mining Applied to Historic Literature***

Another strand of the Archaeotools project was to focus the NLP automated metadata extraction on the PSAS. Despite the highly unstructured nature of the text and the antiquated use of language we were surprised to find that once trained on the grey literature reports the IE software achieved comparable levels of success with the antiquarian literature. Problems were encountered with more synthetic papers, but where the primary subject of the article was a fieldwork report then it was possible to identify the key ‘What’, ‘When’ and ‘Where’ index terms.

After discounting prefatory papers, such as financial accounts, the PSAS corpus was reduced to 3991 papers referring to archaeological discoveries. By applying the rule that the actual ‘What’, ‘Where’ and ‘When’ would appear in the first ten percent of the paper it was possible to identify a subject term for all but 277 of the papers, although there was less success with a geospatial location (627 papers with no location), and least success with period terms (2056 papers with no When term).

However, these results could be improved somewhat by looking at the ‘Reference’ terms; although less certain to provide the primary identification of the key What, Where and ‘When’ for each paper these left far fewer papers unclassified:

**Table 3.** ‘Actual’ identifications for 3991 PSAS papers.

	No data	
What	277	6.9%
Where	627	15.7%
When	2056	51.5%

**Table 4.** ‘Reference’ identifications for 3991 PSAS papers.

	No data	
What	123	3.1%
Where	238	6.0%
When	1049	26.3%

Determining place names within the County-District-Parish (CDP) place name thesaurus proved a challenge, particularly given the number of historic names used in older accounts, but the geo-gazetteer web service hosted by EDINA at the University of Edinburgh was used to resolve many of the outstanding names. Extracted place names were sent directly to this service and the GeoXwalk automatically returned NGRs for the place name (centred) or in the case of some urban areas an actual polygon definition. This allowed the relevant place name from PSAS to be mapped in the Archaeotools geo-spatial interface and therefore made them as discoverable and searchable as standard monument inventory datasets.

Of the total of 3991 PSAS papers, it was initially impossible to find an Ordnance Survey grid reference for 3388 (85 percent), compared to a figure of just 185 (20 percent) for the grey literature. This reflects the fact that older reports did not tend to use precise geospatial references to refer to site or find locations. However, by using the GeoXwalk service it was possible to resolve a place name into a grid reference for all but 268 reports (6.7 percent)—for which there was no ‘Where’ term for 238 reports, leaving just 30 for which a place name had been identified that could not be geo-referenced by the EDINA web service. Manual checking revealed that the majority of these were instances where a county name was the most precise spatial location that had been used in the published paper.

The analysis of the PSAS also provided some tantalising glimpses into the potential of using IE tools to research the development of the use of more controlled and standardised vocabulary through time. In the process of generating the frequency counts used to identify the primary focus of each paper, the Archaeotools project produced frequency counts for each set of named entities for each article in the entire run of the PSAS available from the ADS. These represent the actual frequency of place names, period and monument types within these journals year on year, from 1851 to 1999. A superficial examination of these counts made it apparent that they detailed, metrically, what, when and where was being written about in each year and therefore what was considered significant at that time. It was clear that this could offer significant potential in the longitudinal consideration of changes in archaeological practice and thought. Bateman and Jeffrey (2011) were therefore able to give a more concrete basis to the presumed biases in subject and area believed to exist in the literature. For example, the usage of period terms varies in a non-random fashion both in the actual periods used and the number of different period terms themselves. The Roman period term was shown to dominate early articles and it is not until the 1970s that what we would recognise as the broad modern range of terms reflected in the MIDAS Heritage data standard came into use.

## OPTIMA

By contrast, OPTIMA is an example of a rule-based semantic annotation system that performs the Natural Language Processing (NLP) tasks of Named Entity Recognition, Relation Extraction, Negation Detection and Word-Sense disambiguation using hand-crafted rules and terminological resources (Vlachidis 2012). Semantic Annotation refers to specific metadata which are usually generated with respect to a given ontology and are aimed to automate identification of concepts and their relationships in documents. The system associates contextual abstractions from grey literature documents with classes of the ISO Standard (ISO 21127:2006) CIDOC Conceptual Reference Model (CRM) ontology for cultural heritage and its archaeological extension, CRM-EH. The CRM entities **Physical Object**, **Place**, **Time Appellation** and **Material** are at the core of the system's semantic annotation process and form the basis of the system's acronym. In addition to the four main CIDOC-CRM entities, the system delivers a range of CRM-EH archaeology specific entities and relationships, which are expressed as annotations of 'rich' contextual phases connecting two or more individual entities. The hand-crafted rules of the system are expressed as JAPE grammars which are responsible for the delivery of the semantic annotations in context. JAPE (Java Annotation Pattern Engine) is a finite state transducer, which uses regular expressions for handling pattern-matching rules (Cunningham et al. 2000). The rules are developed and deployed within the NLP framework GATE (Cunningham and Scott 2004) and enable a cascading mechanism of matching conditions.

OPTIMA contributed to the Semantic Technologies for Archaeological Research (STAR) project (Vlachidis et al. 2010, Tudhope et al. 2011), which explored the potential of semantic technologies in cross search and integration of archaeological digital resources. STAR and the follow-on STELLAR projects were collaborations between the Hypermedia Research Unit at the University of South Wales (then the University of Glamorgan) with English Heritage and the ADS, funded by the UK Arts and Humanities Research Council (AHRC). STAR developed new methods for linking digital archive databases, vocabularies and associated unpublished on-line documents originating from OASIS (see above). The project supported the efforts of English Heritage in trying to integrate data from various archaeological projects, exploiting the potential of semantic technologies and NLP techniques to enable complex and semantically defined queries of archaeological digital resources. STAR developed a CRM-EH based search demonstrator which cross searches over five different excavation datasets, together with a subset of archaeological reports from the OASIS grey literature library (examples can be seen in Tudhope et al. 2011).

### ***Named Entity Recognition***

Named Entity Recognition (NER) is a particular IE subtask aimed at the recognition and classification of units of information within predefined categories, such as names of person, location, organisation, and expressions of time, etc. (Nadeau and Sekine 2007). The NER phase of OPTIMA employed hand-crafted rules, glossaries and thesauri to support identification of the four CRM entities (Place, Physical Object, Material and Time Appellation). Specialised vocabulary was also utilised by hand-crafted rules to support word-sense disambiguation and negation detection.

The NER phase introduced a novel approach of Semantic Expansion of the terminology-based resources contributing to the task. This invokes a controlled semantic expansion technique, which exploits synonym and hierarchical relationships of

terminological resources for assigning distinct terminological and ontological definitions to the extracted results. The mechanism is capable of selective exploitation of gazetteer listings via synonyms, narrower and broader concepts relationships. Hence, the system can be configured to a range of different modes of semantic expansion depending on the aims of an IE task, i.e., being lenient and applying a generous semantic expansion or being strict and applying a limited semantic expansion.

A word-sense disambiguation module is invoked by the NER phase to resolve ambiguity between physical object and material terms by assigning appropriate terminological (SKOS) references (Isaac and Summers 2009). For example, when the term ‘brick’ is disambiguated as a material, a terminological reference from the *Material* thesaurus is assigned to the annotation. When the same term is resolved as a physical object, a terminological reference from the *Object Type* thesaurus is assigned instead.

The OPTIMA NER phase also implements a negation detection mechanism targeted at matching phrases which negate any of the four CRM entities (Vlachidis and Tudhope 2013). The implemented mechanism enhances the NegEx algorithm (Chapman et al. 2001) addressing known limitations and domain related issues. The primary aim of the negation module is to strengthen Precision by discarding negated matches that could reduce the validity of results (e.g., delivering a match on ‘*Roman settlement*’ when it originates from the negated phrase ‘*No evidence of Roman settlement*’).

### **Named Entity Recognition Results**

The NER system’s performance was conducted on summary extracts of archaeological fieldwork reports, originating from a range of archaeological contractors. The summaries present some significant advantages over other document sections as they are brief and contain rich discussion which reflects the main findings. The manual annotation task for the purposes of Gold Standard definition was conducted at the ADS by 12 staff and post-graduate students. Table 5 presents the full set of results. The Hypernym mode of semantic expansion, which exploits narrower terms of the vocabulary, delivers the best F-measure rates. However, there is a difference in the system performance between the different entity types.

**Table 5.** F-measure score of four CRM entities (E19.Physical Object, E49.Time Appellation, E53.Place and E57.Material) for the five modes of semantic expansion.

	<b>E19</b>	<b>E49</b>	<b>E53</b>	<b>E57</b>
<b>Only-Glossary</b>	0.63	0.98	0.69	0.50
<b>Synonym</b>	0.76	0.98	0.77	0.52
<b>Hyponym</b>	0.77	0.98	0.82	0.54
<b>Hypernym</b>	0.81	0.98	0.85	0.63
<b>All-Available</b>	0.73	0.98	0.83	0.57

The system performs best (98 percent) for the Time Appellation entity type (E49). The performance is the same across all 5 modes of semantic expansion because the entity is not affected by the expansion modes. This very good performance is based on the completeness of the Timeline thesaurus with its non-ambiguous terms. The Timeline thesaurus is the only terminological resource which contributes to the NER that does not have any overlapping terms with other terminological resources. The results of Physical Object (E19) and Place (E53) entities range from 63 percent to 85 percent depending on the semantic expansion

mode. Places include archaeological contexts and larger groupings of contexts (locations are not the focus of the semantic annotation). The highest score for both entities is delivered by the Hypernym expansion mode reaching 81 percent and 85 percent for the Physical Object and the Place entity respectively. The system delivers the lowest F-measure score (50 percent) in the recognition of Material (E57), which can be ambiguous. For example the same concept ('iron', 'pottery', etc.) could be treated by archaeologists as a find (i.e., physical object) or as the material of an object. Although disambiguation is performed, it can still be challenging to identify. Whether the distinction is worth making might depend on the use cases for the information extraction.

### **Relation Extraction**

Extraction of semantic relations between entities is a significant step towards the development of sophisticated NLP applications that explore the potential of natural language understanding. The OPTIMA pipeline can be configured to detect 'rich' textual phrases that connect CRM entity types in a meaningful way. The aim of the pipeline is to detect and to annotate such phrases as CRM-EH event or property entities. The pipeline uses hand-crafted rules that employ shallow parsing syntactical structures for the detection of 'rich' textual phrases. Other projects have also found shallow parsing useful for tackling the task of relation extraction (Zelenko et al. 2003).

The pair of entities that participate in an event phrase are the arguments of the event. For example the phrase '{ditch contains [pottery] of the Roman period}' delivers two CRM-EH events. One event connects 'ditch' and 'pottery' and another event connects the same 'pottery' with the 'Roman period', both events having 'pottery' as a common argument. The first event can be modelled in CRM-EH terms as a *deposition* event (EHE1004.ContextFindDepositionEvent) while the second event can be modelled as a *production* event (EHE1002.ContextFindProductionEvent). The pipeline detects contextual binary relationships, for example "*pit dates to Prehistoric period*", "*pottery dates to Prehistoric period*", "*ditch contains pottery*" and "*sherds of pottery*", assigning the appropriate CRM-EH ontological annotations (Vlachidis 2012).

## **Conclusions and Future Work**

Mathematical approaches to archaeology have generally been employed in rather esoteric research areas and have tended to lose popularity with the decline in interest in deterministic explanations and the rise of post-processual archaeology. By contrast in the last decade text-mining has increased in importance and has been employed to resolve problems of the inaccessibility of the results of day-to-day archaeological practice, previously locked up in the grey literature. It provides a good example of mathematical techniques serving the needs of the profession, to some extent bridging the gap between academic research and field practice. This has implications for the future structuring of reports in order to facilitate information extraction, for example through the provision of summaries, and the value of using controlled vocabularies. Above all it emphasises that those undertaking scanning projects of unpublished reports and back-runs of printed journals must always plan to produce machine-readable text in order to facilitate easy information extraction.

This chapter has described the underpinning algorithms and has provided an overview of the techniques employed, highlighting their strengths and weaknesses. It has highlighted two projects: Archaeotools, and OPTIMA (which underpins STAR and STELLAR).

Machine Learning and Rule based techniques are sometimes seen as competing NLP paradigms with different strengths and weaknesses. Which works best often depends on the specifics of the entities to be extracted and the language style of the text. It may also depend on the future use cases for the information extraction outputs and the applications that will consume the output.

However, the two methods can be combined, either in a complementary fashion, or sequentially in a pipeline. Archaeotools combined the two methods for different types of entities; OPTIMA employed rule-based techniques for very specific annotations involving ‘rich phrases’ that combined different types of entities in a meaningful way, while retaining the semantics of each entity in the ontological output. Thus the more specific OPTIMA annotations can be seen as complementary to the broader Archaeotools classifications of the main focus of the documents in question. Each can be seen as tending to serve a different use case, thus perhaps Archaeotools in classification for browsing and OPTIMA in providing more detailed annotations for semantic searching.

Looking to the future, as part of the EU-funded ARIADNE research e-infrastructure project (Niccolucci and Richards 2013) the authors are collaborating to explore the possibilities for combining rule-based approaches and machine learning sequentially as stages in a pipeline, as well as investigating the generalisation of OPTIMA rule based techniques to other European language grey literature. With the current interest in Big Data it seems that the potential of text mining to address archaeological research questions and some of the grand challenges of our domain (Kintigh et al. 2014) is only just beginning to be explored.

## Acknowledgements

The STAR and STELLAR projects were supported by the Arts and Humanities Research Council [grant numbers AH/D001528/1, AH/H037357/1]. Archaeotools was funded under the AHRC/EPSRC/JISC eScience programme [grant number AH/E006175/1]. Julian Richards would like to thank Professor Fabio Ciravegna, Sam Chapman and Ziqi Zhang of the Sheffield Organisation, Information and Knowledge (OAK) group for their input to that project. Thanks are also due to Stuart Jeffrey and Lei Xia at the Archaeology Data Service, Phil Carlisle (English Heritage) for providing domain thesauri and for helpful input from Renato Souza (Visiting Fellow at University of South Wales).

## References Cited

- Dublin Core metadata standard <<http://dublincore.org/>>.  
 OASIS project <<http://www.oasis.ac.uk/>>  
*Proceedings of the Society of Antiquaries of Scotland* (PSAS) going back to 1851  
 <<http://dx.doi.org/10.5284/1000184>>.  
 Amrani, A., V. Abajian, Y. Kodratoff and O. Matte-Tailliez. 2008. A chain of text-mining to extract information in Archaeology. Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference 1–5.  
 Baeza-Yates, R. and B. Ribeiro-Neto. 1999. Modern Information Retrieval. Boston, Addison-Wesley Longman Publishing Co., Inc.  
 Bateman, J. and S. Jeffrey. 2011. What Matters about the Monument: reconstructing historical classification. Internet Archaeology 29. <http://dx.doi.org/10.11141/ia.29.6>.  
 Bradley, R. 2006. Bridging the two cultures. Commercial archaeology and the study of prehistoric Britain. *Antiquaries Journal* 86: 1–13.

- Byrne K. 2007. Nested named entity recognition in historical archive text. *In*: Proceedings (ICSC 2007) International Conference on Semantic Computing, Irvine, California, pp. 589–596.
- Byrne, K.F. and E. Klein. 2010. Automatic Extraction of Archaeological Events from Text. pp. 48–56. *In*: B. Frischer, J.W. Crawford and D. Koller (eds.). Making History Interactive. Proceedings of the 37th Computer Application in Archaeology Conference, Williamsburg 2009. Archaeopress, Oxford.
- Chapman, W.W., W. Bridewell, P. Hanbury, G.F. Cooper and B.G. Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics* 34(5): 301–310.
- Ciravegna, F. and A. Lavelli. 2004. Learning Pinocchio: adaptive information extraction for real world applications. *Natural Language Engineering* 10(02): 145–165.
- Cowie, J. and W. Lehnert. 1996. Information extraction. *Communications ACM* 39(1): 80–91.
- Cunningham, H., D. Maynard and V. Tablan. 2000. JAPE a Java Annotation Patterns Engine (Second Edition). [online] Technical report CS—00—10, University of Sheffield, Department of Computer Science. Available at <http://www.dcs.shef.ac.uk/intranet/research/resmes/CS0010.pdf>.
- Cunningham, H. and D. Scott. 2004. Software Architecture for Language Engineering. *Natural Language Engineering* 10(3-4): 205–209.
- English Heritage 2007. MIDAS Heritage—The UK Historic Environment Data Standard (Best practice guidelines) <http://www.english-heritage.org.uk/publications/midas-heritage/midasheritagepartone.pdf>.
- Falkingham, G. 2005. A Whiter Shade of Grey: a new approach to archaeological grey literature using the XML version of the TEI Guidelines. *Internet Archaeology* 17. <http://dx.doi.org/10.11141/ia.17.5>.
- Feldman, R., Y. Aumann, M. Finkelstein-Landau, E. Hurvitz, Y. Regev and A. Yaroshevich. 2002. A Comparative Study of Information Extraction Strategies. Proceedings (CICLing-2002) Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico city, Mexico, 17–23 February.
- Fulford, M. and N. Holbrook. 2011. Assessing the contribution of commercial archaeology to the study of the Roman period in England, 1990–2004. *Antiquaries Journal* 91: 323–345.
- Gaizauskas, R. and Y. Wilks. 1998. Information extraction: beyond document retrieval. *Journal of Documentation* 54(1): 70–105.
- Hardman, C. and J.D. Richards. 2003. OASIS: dealing with the digital revolution. pp. 325–328. *In*: M. Doerr and A. Sarris (eds.). CAA2002: The Digital Heritage of Archaeology. Computer Applications and Quantitative Methods in Archaeology 2002. Archive of Monuments and Publications Hellenic Ministry of Culture.
- Hobbs, J.R., D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel and M. Tyson. 1993. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. *In* Proceedings (IJCAI 1993) 13th International Joint Conference on Artificial Intelligence, Chambéry, France, 28 August–3 September.
- Isaac, A. and E. Summers. 2009. SKOS Simple Knowledge Organization System Primer. [Online]. Available at: <http://www.w3.org/TR/skos-primer> (Accessed: 12 June 2012).
- Jeffrey, S., J.D. Richards, F. Ciravegna, S. Waller, S. Chapman and Z. Zhang. 2009. The Archaeotools project: faceted classification and natural language processing in an archaeological context. *In*: P. Coveney (ed.). Crossing Boundaries: Computational Science, E-Science and Global E-Infrastructures, Special Themed Issue of the Philosophical Transactions of the Royal Society A 367: 2507–19.
- Kintigh, K., J. Altschul, M. Beaudry, R. Drennan, A. Kinzig, T. Kohler, W.F. Limp, H. Maschner, W. Michener, T. Pauketat, P. Peregrine, J. Sabloff, T. Wilkinson, H. Wright and M. Zeder. 2014. Grand Challenges for Archaeology. *American Antiquity* 79: 5–24.
- Lin, D. 1995. University of Manitoba: description of the PIE system used for MUC-6. *In* Proceedings (MUC 6) 6th Message Understanding Conference, Columbia, Maryland, 6–8 November.
- Maynard, D., W. Peters and Y. Li. 2006. Metrics for Evaluation of Ontology-based Information Extraction, *In*: Proceedings WWW Conference 2006, Workshop on “Evaluation of Ontologies for the Web”, Edinburgh, Scotland, 23–26 May.
- Michel, J.-B., Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, The Google Books Team, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M.A. Nowak and E. Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331, 176. doi:10.1126/science.1199644.
- Moenz, M.F. 2006. Information Extraction Algorithms and Prospects in a Retrieval Context. Dordrecht, Springer.
- Nadeau, D. and S. Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1): 3–26.

- Newman, M. and P. Gilman. 2007. *Informing the Future of the Past: Guidelines for Historic Environment Records* (2nd edition). ADS, ALGAO UK, English Heritage, Historic Scotland, RCAHMS and RCAHMW.
- Nicolucci, F. and J.D. Richards. 2013. ARIADNE: Advanced Research Infrastructures for Archaeological Dataset Networking in Europe. *International Journal of Humanities and Arts Computing* 7(1-2): 70–88.
- Nilsson, N. 2005. *Introduction to Machine Learning*. [Online]. Nils J Nilsson publications. Available at: <http://robotics.stanford.edu/people/nilsson/mlbook.html>.
- Pajmans, H. and S. Wubben. 2008. Preparing archaeological reports for intelligent retrieval. pp. 212–217. *In: A. Posluschny, K. Lambers and I. Herzog (eds.). Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA) Berlin, Germany, April 2–6, 2007. Kolloquien zur Vor- und Frühgeschichte Band 10, Bonn.*
- Richards, J.D. and C. Hardman. 2008. Stepping back from the trench edge. An archaeological perspective on the development of standards for recording and publication. pp. 101–112. *In: M. Greengrass and L. Hughes (eds.). The Virtual Representation of the Past. Ashgate, London.*
- Richards, J.D., S. Jeffrey, S. Waller, F. Ciravegna, S. Chapman and Z. Zhang. 2011. The Archaeology Data Service and the Archaeotools project: faceted classification and natural language processing. pp. 31–56. *In: S. Whitcher Kansa, E.C. Kansa and E. Watrall (eds.). Archaeology 2.0 and Beyond: New Tools for Collaboration and Communication. Cotsen Institute of Archaeology Press, Los Angeles.*
- Savary, A., J. Waszczuk and A. Przepiórkowski. 2010. Towards the Annotation of Named Entities in the National Corpus of Polish. In *Proceedings (LREC'10) Fourth International Conference on Language Resources and Evaluation, Valletta, Malta, 13–17 May.*
- Smeaton, A.F. 1997. Using NLP and NLP resources for Information Retrieval Tasks. *In: T. Strzalkowski (ed.). Natural Language Information Retrieval, Kluwer Academic Publishers.*
- Soderland, S., D. Fisher and W. Lehnert. 1997. *Automatically Learned vs. Hand-crafted Text Analysis Rules, CIIR Technical Report T44, University of Massachusetts, Amherst.*
- Tudhope, D., K. May, C. Binding and A. Vlachidis. 2011. Connecting Archaeological Data and Grey Literature via Semantic Cross Search, *Internet Archaeology* 30. <http://dx.doi.org/10.11141/ia.30.5>.
- Uren, V., P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta and F. Ciravegna. 2006. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web* 4(1): 14–28.
- Vlachidis, A., C. Binding, K. May and D. Tudhope. 2010. Excavating grey literature: a case study on the rich indexing of archaeological documents via Natural Language Processing techniques and knowledge based resources. *ASLIB Proceedings* 62 (4&5): 466–475.
- Vlachidis, A. and D. Tudhope. 2012. A pilot investigation of information extraction in the semantic annotation of archaeological reports. *International Journal of Metadata, Semantics and Ontologies* 7(3): 222–235. Inderscience.
- Vlachidis, A. 2012. *Semantic Indexing via Knowledge Organization Systems: Applying the CIDOC-CRM to Archaeological Grey Literature*. Unpublished PhD Thesis, University of South Wales (USW).
- Vlachidis, A. and D. Tudhope. 2013. The Semantics of Negation Detection in Archaeological Grey Literature. pp. 188–200. *In: E. Garoufallo and J. Greenberg (eds.). Metadata and Semantics Research. Communications in Computer and Information Science* 390.
- Wilks, Y. and C. Brewster. 2009. *Natural Language Processing as a Foundation of the Semantic Web, Foundations and Trends in Web Science* 1(3-4): 199–327.
- Zhang, Z., S. Chapman and F. Ciravegna. 2010. A Methodology towards Effective and Efficient Manual Document Annotation: Addressing Annotator Discrepancy and Annotation Quality, *Lecture Notes in Computer Science*, 6317 pp. 301–315, Springer-Verlag: London.
- Zelenko, D., C. Aone and A. Richardella. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research* 3: 1083–1106.