A Pilot Investigation of Information Extraction in the Semantic Annotation of Archaeological Reports

Andreas Vlachidis[1], Douglas Tudhope[1]

[1] Hypermedia Research Unit, Faculty of Advanced Technology, University of Glamorgan, Pontypridd, CF37 1DL, Wales, UK
avlachid@glam.ac.uk, dstudhope@glam.ac.uk

Andreas Vlachidis is a Research Assistant in the Hypermedia Research unit at the University of Glamorgan and is currently writing up his PhD in the area of natural language processing, producing semantic annotations conforming to the CIDOC CRM ontology. He was responsible for producing the semantic annotations of grey literature in the AHRC funded STAR project (Semantic Tools for Archaeological Resources). He has previously worked as Division Leader for Computing Courses at North College in Thessaloniki and as ICT Developments Manager in the eCommerce domain.

Douglas Tudhope is Professor in the Faculty of Advanced Technology, University of Glamorgan and leads the Hypermedia Research Unit. His main current research interests are the intersecting areas of information science, digital libraries and hypermedia and the semantic web. He was PI on the AHRC funded STAR and STELLAR projects and the EPSRC funded FACET project investigating thesaurus-based query expansion. Since 1977, he has been Editor of the journal, New Review of Hypermedia and Multimedia. He serves as a reviewer for various journals and international programme committees and is active in the Networked Knowledge Organisation Systems/Services (NKOS) network.

Abstract. The paper discusses a prototype investigation of semantic annotation, a form of metadata assigning conceptual entities to textual instances, in this case archaeological grey literature. The use of Information Extraction (IE), a Natural Language Processing (NLP) technique, is central to the annotation process while the use of Knowledge Organization System (KOS) is explored for the association of semantic annotation with both ontological and terminological references. The annotation process follows a rule-based information extraction approach using the GATE NLP toolkit, together with the CIDOC CRM ontology, its CRM-EH archaeological extension and English Heritage thesauri and glossaries. Results are reported from an initial evaluation, which suggest that these information extraction techniques can be applied to archaeological grey literature reports. Further work is discussed drawing on the evaluation and consideration of the characteristics of the archaeology domain.

Keywords: Natural Language Processing, Knowledge Organization Systems, Semantic Annotation, Information Extraction, GATE, Digital Archaeology, Grey Literature, CIDOC CRM ontology,

1 Introduction

Today we witness an increasing awareness of the potential of semantic contextualisation, employing conceptual models and sharing as much semantic context as possible via open data architectures. In particular, the field of Cultural Heritage has benefited from projects such as Europeana, which delivers significant semantic enrichment to more than 6 million cross-Europe digital items (Gradmann 2010). The UK digital archaeology domain has been supported by projects such as Archaeotools (Jeffrey et.al 2009) and STAR - Semantic Technologies for Archaeological Resources (Tudhope et al. 2011), which have provided semantic methods for linking archaeological resources.

The paper discusses a prototype development and evaluation of a specialised form of metadata known as semantic annotation, which carries the potential to support semantic indexing of natural language context. The discussion reveals the role of Knowledge Organization Systems (KOS) such as thesauri and ontologies for supporting Natural Language Processing techniques, in particular Information Extraction, employed for the identification and association of textual snippets with conceptual and terminological references. Such references can be exploited further by Information Retrieval mechanisms for supporting complex and semantic-aware information seeking activities. The first part of the paper presents background information and relevant literature, the second part discusses the development process while the third part reveals the evaluation method and concludes with the results and future developments.

1.1 Aims of the IE Prototype Development

The prototype stage is part of a larger project (section 2.1), investigating the use of NLP techniques in combination with KOS resources. The main aim of the prototype development is to explore the potential of rule-based IE techniques to deliver semantic-aware abstractions of the free text information in archaeological reports (grey-literature) which can be exploited further by retrieval application, such as STAR. The KOS employed are the CIDOC CRM ontology (Crofts et al. 2011) and the CRM-EH extension for archaeology (Cripps et al. 2009), together with English Heritage terminological resources (EH 2006).

The prototype employs the GATE language engineering architecture (Cunningham et al. 2002) to accommodate the task of IE with respect to the above ontologies and terminological resources, using hand-crafted IE rules targeted at archaeological concepts. This paper presents advances on earlier experience in the archaeology domain over a smaller corpus (Vlachidis et al. 2010) by discussing pre-processing and lessons from an initial evaluation of the prototype information extraction system, following established evaluation measurements for assessing the performance of semantic annotation systems.

1.2 NLP for Advancing Information Retrieval

The complexity of human language results in a challenging environment for computations to provide solutions for the whole range of language related processes. Language ambiguities are part of language itself and concern a number of lexical, syntactic and semantic ambiguities which can considerably influence the performance of Information Retrieval (IR) systems. Polysemous words and synonyms generate ambiguity, which statistical methods are ill-equipped to address (Lewis and Jones, 1996; Moens, 2006).

Information Extraction (IE) is a Natural Language Processing (NLP) technique that analyses a textual input and produces a structured textual output that is suitable for further manipulation. IE tasks do not involve finding relevant documents from a collection but they are rather text analysis tasks aimed at extracting specific information snippets from documents. The output of IE can be directed towards automatic database population, machine translation, term indexing analysis and text summary generation. (Gaizauskas and Wilks, 1998; Lewis and Jones, 1996; Moens, 2006)

The fundamentally different role of IE does not compete with IR; on the contrary the potential combination of the two technologies promises the creation of new powerful tools in text processing. In particular, IR could benefit from the construction of sensitive indices closer related to the “actual meaning” of a given text (Cowie and Lehnert 1996).

1.3 Simple Knowledge Organization Systems

Simple Knowledge Organization System (SKOS) is a standard formal representation of structured controlled vocabulary systems, such as thesauri (Isaac and Summers 2009). SKOS is intended to enable easy publication of controlled structured vocabularies for the Semantic Web via standard RDF(S)/XML W3C technologies. The encoding of information in RDF allows distribution and decentralisation of knowledge organization systems to computer applications in an interoperable way.

SKOS representations are lightweight, capable of expressing semantic structures that can be employed in search and browsing applications. They allow usage of unique identifiers (URIs) for each concept as well as enabling linking between concepts. The intra scheme relationships, such as *skos:Narrower* and *skos:Broader*, supports linking between semantically narrower (hyponym) and broader concepts (hypernym) concepts. In addition, mapping relationships such as *skos:exactMatch* and *skos:closerMatch*, enable linking between concepts of different KOS according to the degree of match.

1.4 Ontologies for Information Extraction

Ontology in philosophical terms is the study about the nature of existence of 'things' (Guarino, 1998; Wilks, 2008). In its simplest form, a computer science ontology might be described as a taxonomy and a set of inference rules. Ontologies can be

understood as conceptual structures that formally describe a given domain by defining classes and sub-classes of interest and by imposing rules and relationships among them, in order to determine a formal structure of 'things' (Berners-Lee, Hendler, and Lassila 2001; Wilks 2008). The size and the scope defines, whether an ontology is called light-weight, core or upper level but all ontologies are based on the basic assumption, that there is a single reality supporting the cohesion of their structure.

Ontologies can be incorporated in both rule-based and adaptive tools to enhance system operation and to describe the conceptual arrangements of semantic annotations. Such IE systems can be described as ontology based (OBIE) or ontology oriented (OOIE) depending on the level of ontology engagement (Bontcheva, Li, and Cunningham, 2007).

1.5 The Role of Ontology-Based Semantic Annotation

Semantic annotation is the process of tying ontological definitions to natural language text by providing class information for textual instances (Bontcheva, 2006). Described as a mediator platform between concepts and their worded representations, semantic annotation as metadata can automate the identification of concepts and their relationships in documents. It is proposed as a mechanism for connecting natural language and formal conceptual structures to enable new information access methods and to enhance existing ones.

The annotation process enriches documents and enables access on the basis of a conceptual structure. This aids information retrieval from heterogeneous data sources, empowering users to search across resources for entities and relations instead of words. As evident from a number of IE projects, semantic annotation has the potential to bridge the gap between natural language text and formal knowledge expressed in ontologies (Uren et.al 2006).

1.6 Application Domains and Semantic Annotation

The work described here is targeted particularly at the archaeology domain. Before considering archaeology, we briefly place it in context of other domains. Different application domains and typical use cases have their own individual characteristics which may pose challenges for semantic annotation. However, it is not always easy to separate the inherent features of an application domain from the particular aims and constraints of individual research projects, which may also have consequences for semantic annotation. Thus the availability of annotated corpora and vocabulary resources are not necessarily inherent to a domain but may have a strong influence on research at a particular time. Similarly a project may choose to focus on a particular IE research objective.

Much of the existing semantic annotation work is influenced by the aims of the MUC series (Grishman and Sundheim 1996) and involves, for example, projects targeting Named Entity Recognition (NER) of business related Web resources (eg financial news documents). Typical entities are often proper nouns and might include Persons, Organizations, Places and Dates. For example, Bontcheva et al. (2004)

describe a technique of ontology engagement in semantic annotation using the GATE OWLIM-Lite processing resource, which associates ontological classes with one or more vocabulary listings (gazetteers). Lists contain entries which help populate ontological classes with instances, for example the class *Location* can be associated with the list *Cities* containing entries, such as *London*, *Paris*, *Athens* etc.

General purpose platforms such as KIM (Kiryakov et al. 2004) make use of domain-independent, upper level ontologies, in this case KIMO, for supporting NER and beyond to the level of semantic document retrieval. KIMO is now superseded by PROTON (Mascardi et al. 2006).

Domain-specific projects include KMP h-TechSight (Maynard et al. 2005) for automatic monitoring of Web information resources. Semantic annotation of job advertisements is based on rules working with an ontology consisting of 9 classes, such as Location, Sector, Job Title, Salary, Expertise, etc. This results in text strings that are annotated as instances of a class in the ontology. In some cases this requires more complex contextual rules. Adoption of the KMP application to the area of chemical technologies, required construction of a new ontology consisting of 13 concepts such as Corrosion, Thermodynamics, Optimization, Reaction, Equipment, etc. populated from 181 vocabulary lists.

On the other hand, the biology and biomedical domain poses somewhat different challenges (Ananiadou et al. 2004, 2005). Here we see a large volume of constantly growing literature and an interest in text mining to supplement traditional retrieval applications, due to the volume of potential results. There is also a potential interest in the integration of literature search with experimental and factual databases.

The vocabulary is a highly specific scientific terminology (genes, proteins, drugs, etc.) and is constantly evolving; new technical terms are dynamically appearing. Basic term identification and appropriate terminological association is one of the key problems - the process is very context dependent. The organization of technical terms in Knowledge Organization Systems with hierarchies and association between terms is an important part of the work. The problem of annotating biomedical entities, which may be indicated by descriptors rather than a controlled vocabulary, can differ from the identification of good subject description index terms. For example, the frequency of occurrence may be less important.

In the domain of molecular biology the GENIA ontology, described as a “formal model of cell signalling reactions in human”, consists of 45 classes which classify substances by their chemical structure (Ohta et al. 2002). The choice to classify on chemical structure and not on the biological role of substances is explained by the behaviour of the substances to change biological role depending on biological contexts but to maintain their chemical structure regardless of context.

According to Tsujii and Ananiadou (2005), unlike some domains such as e-business, the tendency to make logical semantic connections between a (formal) ontology and specific text strings can be problematic in the biomedical domain. For example if a substance is an enzyme or whether a protein contains certain properties or not depends on contextual factors. They argue that contextual dependencies strongly influence the semantic annotation process: “*... relationships among concepts as well as the concepts themselves* r*emain implicit in text, waiting to be discovered*”. Thus general language ambiguity and domain-dependent inferences and knowledge

are barriers to comprehensive encoding in ontological structures. New discoveries may change the understanding of a particular concept.

The biomedical domain poses problems for purely logical deduction. Different communities within the same broad field have evolved their particular vocabularies and language uses. Interpretation of context is important for the selection of relevant facts from the literature, where inevitably language is ambiguous. They argue that terminological thesauri, as language oriented structures, are more appropriate for supporting implicit definition of semantics in text

1.7 Semantic Annotation of Archaeological Grey Literature

The application domain of archaeological grey literature has some similarities with both the general business oriented Web IE and the biomedical domain, while retaining particular features of its own. It has, however, a particular concern with contextual issues, although they differ in some respects from the biomedical contextual issues discussed above.

Archaeological grey literature is considered a valuable but under utilised resource in the field. As with the biomedical domain, there is a desire to integrate both the published archaeological literature and grey literature with excavation datasets (Tudhope et al. 2011), although since grey literature is not formally published its style can sometimes be less formal. While archaeological terminology has some specialised technical vocabulary, it is distinguished by use of common terms, some of which are employed in archaeologically specific ways, for example ‘cut’, ‘context’, ‘deposit’, ‘find’. Other common terms have particular significance when associated with archaeological events in the past. Thus a ‘Roman road’ will probably be of interest but not a contemporary road (which might be useful for spatially describing an excavation site but is probably not a focus of inquiry in itself). This is aptly illustrated by the importance of the very term ‘context’, denoting a place that holds the context for archaeological ‘finds’. Thus consideration of the context for terms matched in the text is vital; there is no necessary connection between a term and an ontology instance.

In addition, the scientific vocabulary of the archaeology domain is not as heavily specialised as some domains. For example, in biology ‘Glucocorticoid’ is a hormone; this term will rarely be used outside the biology domain and it is nearly impossible to refer to something different than the known hormone. On the other hand, in archaeology the term ‘pit’ is very frequently used to describe an archaeological context. Although, this term clearly refers to a place it is not as specialised as ‘Glucocorticoid’ and it cannot be assumed that every instance of ‘pit’ refers to an archaeological context, even when the term occurs within the content of an archaeological grey literature report. Thus entity specialisation cannot be inferred solely by a specialised vocabulary but must be derived by a combination of vocabulary and contextual evidence.

Future research should aim to investigate further the above observation regarding the limited discriminatory power of archaeological terms in isolation. Since domain vocabularies are available, one potential approach could examine the percentage of overlapping terms between archaeology vocabulary and a general purpose English dictionary. The analysis should include a determination of whether the correct sense

of a term is included in the dictionary. For example, for the term, archaeological ‘context’, it is likely that the broad sense would fall within a general dictionary but not the specific archaeological sense. This requires fine distinctions in the results with perhaps a degree of match, or even a semantic mapping relationship between the vocabulary term and dictionary sense (the SKOS mapping relationships could be employed). In order to give a basis for comparison, the results could be compared with the percentage of term overlap between the dictionary and another domain vocabulary, for example biology. However, it may be necessary to consider some principle for selecting roughly equivalent terms from the two domains. For example, proper nouns (such as ‘Glucocorticoid’) may be more common in one domain than another. Another element of the comparison might analyse the relative frequency of proper nouns (and other parts of speech) in relevant vocabularies from different domains. An empirical approach could also compare and contrast the volume of terms from different domain-specific vocabularies (or even the actual texts) that also occur within a general purpose corpus, such as the Brown corpus (Francis and Kucera 1964).

Since the assignment of (usually approximate) dates is one of the broad aims of archaeological investigation, there is a fine grained terminology relating to historical time periods, often with moderators. (eg ‘later phase of postmedieval period’) . Given the limited amount of available evidence in many excavations, findings are often qualified and detection of moderators and qualified assertions is important (e.g. 'occasional charcoal', 'mould decorated beaker', ‘randomly coursed bricks’). Negation detection is also potentially important given the prevalence of negative findings in archaeological reports (e.g. ‘no traces of a Roman settlement’).

As discussed earlier, differences in approach can be in part a matter of differing emphasis in the research goals. STAR’s focus is on detailed consideration of archaeological features. However, it is also possible for archaeological IE to focus on NER of People and Locations and prominent objects. To some extent, this is the approach followed by the Archaeotools project (Jeffrey et.al 2009). Information extraction is focused on the four facets of the classification; *What* (what subject does the record refer to), *Where* (where, location, region of interest), *When* (archaeological date of interest) and *Media* (form of the record). Byrne (2007) also focused on NER of relatively high level entities from historical archive texts, originating from the Royal Commission on the Ancient and Historical Monuments of Scotland (RCAHMS). The NER task employed 11 classes; Organisation, Person-name, Role, Site-type, Artefact, Place, Sitename, Address, Period, Date, and Event. Similarly Grover et al. (2008) applied NER techniques over historical texts from the House of Lords, dating to the 18th century. The project employed a rule-based approach supported by lexicons (gazetteers) for the identification of person and place names.

STAR’s objectives required a finer granularity of detail, in order to integrate the semantic annotations with the information extracted from excavation datasets. As described below, entities include archaeological objects (finds) and materials, time periods and the specific places (contexts) that finds are associated with. Spatial coordinates of contexts are not usually appropriate at the level of detail of a grey literature report (though may be included in an accompanying database), while the site’s location is usually part of the meta-data.

For STAR's purposes, it is desirable to annotate 'rich phrases' of semantically associated entities, as opposed to isolated entities, for example associating a find with its appropriate context and time period, if that information is provided. Some initial results in this direction are reported below, although full exploration of relation extraction remains for future work.

Ambiguity is inherent in much of the archaeological vocabulary, for example object-material, place-object. Sometimes the distinction reflects the focus of scholarly inquiry rather than any absolute semantics. An entity may be treated as either a place/context or as an object, depending on the archaeological objective, for example both 'vessel' and 'skeleton' can be objects of inquiry in their own right, or be treated as the contexts for archaeological finds. The underlying aims of the ontological modelling are also relevant; different ontologies model different concerns. The CRM-EH is concerned to model the processes involved in archaeological excavation recording and analysis. Thus a particular pottery fragment becomes a 'find', as a result of being extracted from a context and recorded separately on site. Ambiguity is an ever present feature, one that may not be possible to resolve during the IE process.

Thus similarly to the biomedicine domain, there is a strong concern with context but the semantic annotation of archaeological texts imposes its own highly contextual challenges. Contextual factors dictate if, for example, a particular place is an archaeological 'context', or if a physical object constitutes an archaeological 'find'. This also has consequences for evaluation – see section 4.2's discussion of a user-centred approach to the methodology.

2 Archaeological background to the Prototype Development

This section discusses some of the main implications of the archaeological characteristics discussed in the previous section for the work described in this paper. These characteristics also influence the discussion of the evaluation in section 4.3.

Semantic annotation aims to support the STAR project in the semantic discovery and interoperability of archaeological information. The immediate application is to support cross search of excavation datasets and grey literature reports integrated via an umbrella ontology. The use of ontological and terminological resources is influenced by the characteristics of archaeological vocabulary discussed above and various project specific issues. To support semantic interoperability, STAR employs the CIDOC CRM (and its archaeological extension CRM-EH) ontology, while also utilising a range of English Heritage terminological resources. The semantic annotation effort is targeted at delivering semantic indexes which will support information retrieval at the level of *concepts*. Thus, the prototype system is not concerned with the annotation of unique individuals *(post-hole A, post-hole B)* but with the annotation of concepts (the concept of *post-hole)*. However a concept may have term variants *(post hole).*

There is no commonly agreed glossary of terms although English Heritage thesauri and glossaries are influential within the field. However different archaeological teams may use different terms for the same concept and in some cases a typology may be the outcome of an investigation rather than the starting point. Thus language use is

fluid and sometimes in flux. The scope of the thesauri is broader than STAR's focus on search and inquiry over excavation datasets, requiring also the more detailed terminology within the glossaries. However the thesauri and glossaries were developed independently from each other and from the ontology; the connections are highly context dependent. The EH thesauri are information retrieval thesauri, which share some features with the terminological thesauri discussed by Tsujii and Ananiadou (2005) but are more oriented to supporting retrieval than linguistic processing (ISO 25964). Some terminology work was done within the project to develop specialised time period glossaries, together with period-specific moderators, due to their importance within the domain.

The CRM does not afford an uncomplicated context-independent association between ontological classes and the available English Heritage vocabulary, neither directly associated vocabularies nor individual instances. Semantic annotation at the CRM-EH level cannot be reached by specialised vocabulary alone. The archaeology domain vocabulary does not contain heavily specialised scientific terms and archaeological vocabulary usage is highly contextual. In addition, as discussed above, there is ambiguity surrounding some uses of archaeological terminology, as far as the ontology is concerned.

Therefore, a simple association of thesauri resources with CRM-EH classes does not answer the contextual dependencies of the archaeology domain as described above. A complementary use of terminological and ontological resources was adopted, combining the CRM-EH with a variety of archaeological vocabularies, together with grammar and contextual rules.

Using complementary ontological and terminological (thesaurus and glossary) resources empowers dual semantic annotations, both expressed as URIs (Figure 1). Extracting CRM ontology elements supports data integration and potentially logical inferencing if that is appropriate. Extracting SKOS concepts supports retrieval applications of browsing and semantic search. The distinction is similar to that drawn by the W3C Library Linked Data Incubator Group Report (2011) in its discussion of Value Vocabularies and Metadata element sets.

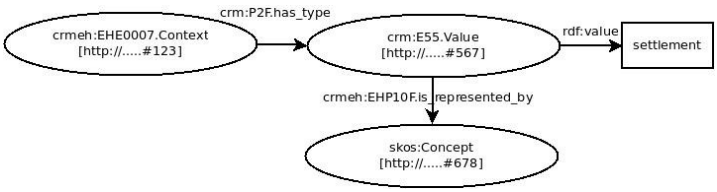

Fig. 1: A Semantic Annotation example containing a terminological and an ontological reference. The textual value of the annotation is "settlement" which is defined as an instance of the CRM-EH class EHE0007.Context. The value is linked (is_represented_by) a SKOS definition.

The prototype development has an experimental focus aimed at obtaining practical experience and results to inform a large scale semantic annotation effort. This includes initial experiments regarding moderators, relation extraction and negation detection, which are important techniques for the archaeology domain as discussed above. These will be further developed in future work (see section 4.3). Another focus

for future work is whether the instance relationship is the most appropriate connection between an ontology class and a textual occurrence, given the contextual discussion above. Consideration should be given to the modelling of the provenance and reliability of semantic annotations within an ontological framework, where they may form the basis of logical inferencing.

2.1 Semantic Technologies for Archaeological Resources (STAR) project

The Semantic Technologies for Archaeological Resources (STAR) project aims to develop new methods for linking digital archive databases, vocabularies and associated unpublished on-line documents. The project supports the efforts of English Heritage (EH) in trying to integrate the data from various archaeological projects and their associated activities and seeks to exploit the potential of semantic technologies and natural language processing techniques, for enabling semantically defined queries over archaeological digital resources (Tudhope et al. 2011).

To achieve semantic interoperability over diverse information resources and to support complex and semantically defined queries, STAR adopted the English Heritage extension of the CIDOC Conceptual Reference Model (CRM-EH). The project demonstrator cross searches disparate datasets[1] and a subset of archaeological reports of the OASIS grey literature corpus. Also the project produced a set of web services for accessing the SKOS terminological references and relationships of the domain thesauri and glossaries.

2.2 OASIS Grey Literature Reports

The term grey literature is used by librarians and research scholars to describe a range of documents and source materials that cannot be found through the conventional means of publication. Preprints, meeting reports, technical reports, working papers, white papers are just a few examples of grey literature documents which are not always published by conventional means. The need for solutions targeted at accessing information held by available grey literature documents was identified as early as 1995 (Debachere 1995) and is still a major research issue today.

A considerable volume of grey literature documents falls within the scope of the STAR project. Some grey literature documents hold information relative to archaeological datasets that have been produced during archaeological excavations and summarise sampling data and excavation activities. Some grey literature may be concerned with other types of investigation that fall short of an excavation but may hold useful information. Integration of grey literature in STAR is intended for enabling cross-searching capabilities between datasets and grey literature documents, with respect to the semantics defined by the adopted CRM-EH ontology.

The collection of grey literature documents (corpus) that concerns the prototype development originates from the Online AccesS to the Index of archaeological investigations (OASIS) project. The OASIS project is a joint effort of UK

1 Raunds Roman, Raunds Prehistoric, Museum of London, Silchester Roman and Stanwick sampling

archaeology research groups, institutions, and organizations, coordinated by the Archaeology Data Service (ADS), University of York, aiming to provide an online index to archaeological grey literature documents.

2.3 The CIDOC CRM – EH ontology

The CIDOC CRM core ontology for cultural heritage information aims to enable information exchange between heterogeneous resources by providing the required semantic definitions and clarifications. The CRM is the result of 10 years effort by the CIDOC Documentation Standards Working Group and has become an ISO Standard (ISO 21127:2006). It is a comprehensive semantic framework aimed at promoting shared understanding of cultural heritage information and is particularly relevant to projects relating to archaeological cross domain research. Since the CRM is an ISO standard within the cultural heritage domain, the resulting semantic annotations should have potential for interoperability with CRM-based metadata more generally.

The central concepts of the CIDOC CRM ontology are Temporal Entities of spatio-temporal boundaries involving Time-Spans and Places putting events at the main focus of the model. Such events involve Persistent Items, such as Physical Things and Actors and immaterial objects like Conceptual Objects. Any instance of a class can be identified by Appellations like labels, names, or whatever else used in context. In addition, Types allow further detailed classification of any class instance supporting additional distinction and property engagement. The latest version of the CIDOC CRM comprises 90 classes and 148 properties (Crofts et al. 2011).

EH plays a major role in the dissemination of standards in cultural heritage domain, both at national and international level. EH attempted an initial modelling exercise of the archaeological domain to the existing CIDOC CRM ontology. After consultation with the CIDOC CRM-SIG, the modelling exercise concluded that extension of the CRM ontology to the archaeological domain entities was necessary due to the complexity and specificity required in representing the broader archaeological processes.

The extended CRM-EH ontology (Cripps et al. 2004) comprises 125 extension sub-classes and 4 extension sub-properties. Based on the archaeological notion of context, which is modelled as a subclass of place, the CRM-EH describes entities and relationships relating to a series of archaeological events, including stratigraphic relationships and phasing information, finds recording and environmental sampling.

2.4 English Heritage Terminological Resources

EH made available a large number of terminology resources (glossaries and thesauri) to the STAR project for supporting its aims for widening access to digital archaeology resources. The available glossaries of recording manuals (EH 2006) and EH thesauri were previously converted from their original format (recording manuals and relational databases) to controlled terminology SKOS resources using XSL transformation techniques (Binding, Tudhope and May 2008).

The terminology resources adopted by the prototype were; the *Simple Names for Deposits and Cuts* glossary, which provides a controlled vocabulary for recording archaeological context (taken here to also include broader interpretive groupings); the *MDA Archaeological Object Type* thesaurus which contains physical evidence that can be recovered from archaeological fieldwork such as objects and environmental remains; and The *Timeline thesaurus* which, contains dates and periods under 6 categories; artistic period, cultural period, geological period, historic period, political period and religious period. *Simple Names for Deposits and Cuts* contained both archaeological contexts (eg cut) and broader semantic groupings of basic contexts (eg ditch). Since grey literature reports tend to reflect a higher level of generality in reporting excavations, after consultation with archaeologists it was decided that the IE pipeline should yield a composite group-context entity that reflected the meaning of a group as a collection of contexts.

2.5 The Platform of the IE Prototype Pipeline

The prototype pipeline was developed within the GATE (General Architecture for Text Engineering) environment, utilising hand crafted JAPE rules and exploiting domain vocabulary that is converted to gazetteer listings. GATE is an infrastructure for processing human language, which provides the architecture and the framework environment for developing and deploying natural language software components (Cunningham et al. 2002).

JAPE (Java Annotation Pattern Engine) grammar is a finite state transducer, which uses regular expressions for handling pattern-matching expressions (Cunningham, Maynard, and Tablan 2000). Such expressions are at the core of all rule-based IE systems aimed at recognising textual snippets that conform to particular patterns. The rules enable a cascading mechanism of matching conditions that is usually referred to as the IE pipeline.

3 The Prototype Development of IE Pipelines

Two separate information extraction pipelines were developed to address particular objectives of the information extraction task. Both contribute to the main aim of the provision of semantic annotation associated with terminological and ontological reference with respect to the EH vocabulary and CRM-EH ontology respectively.

The first pipeline (pre-processing) is intended to reveal commonly occurring section titles of the grey literature documents and to extract the summary sections of grey literature documents. Section titles are isolated from the semantic annotation phase while summaries were identified as being important document sections, containing rich information worth targeting by the semantic annotation phase.

Complementary use of the ontologies and terminological resources is explored by the second, main semantic annotation phase, which is aimed at identifying textual instances of information from grey literature documents. Such instances are associated with CRM and CRM-EH ontological entities that contain links to SKOS terminological definitions Figure 1.

3.1 Transforming to Plain Text

The grey literature documents consisted of 2460 archaeological reports originating from the OASIS corpus. Due to the employment of specific JAPE grammars targeted at identifying document sections and headings, it is important to avoid generation of blank areas and multiple line breaks which occur during that import of PDF and MS word in GATE. For this reason, there was an initial transformation of files to plain text using the Linux shell application "*pdftotext*". The application allows transformation of files in a raw dump format that suppresses page breaks and blank areas, while it enforces specific text encoding (Latin 1) and single carriage return (Unix). File transformation enables the construction of JAPE rules that detect document sections by exploiting the single carriage return.

3.2 Pre-Processing Corpus Collection

The pre-processing phase (Figure 2) employs domain neutral information extraction techniques for the identification of specific document sections, which are either excluded from the semantic annotation phase or used as input at later stages of the prototype IE pipeline.

As discussed in section 1.5, contextual-dependency influences the validity of semantic annotation in archaeological text. As a very simple example, the inclusion of document sections such as heading and table of contents (TOC) in the semantic annotation effort can lead to annotations with limited validity. Headings and TOC might make use of EH vocabulary, however such sections do not use terms in a rich discussion setting but instead use terms in isolation, as in titles. Detection of headings also supports extraction of document sections, such as summary sections, which contain rich discussion worth revealing.

The identification of heading spans is based on a collection of eight different pattern-matching rules. Two rules annotate headings that commence with a numerical prefix followed by a capitalised or upper initial word, which might be followed by more words not necessarily in capital or upper initial case, such as "*3.1 Prehistoric phase*". Another set of rules are targeted at single worded headings that have upper initial or capitalised case and do not commence with numerical prefixes, such as"*Introduction*".

The identification of TOC is based on a pattern that joins four or more previously identified *Heading* annotations together. Similarly the identification of *Summary* sections is also based on a JAPE grammar, which annotates as summary a document section that is wrapped between two previously identified *Heading* annotations. The first Heading annotation must contain any of the words; "summary", "abstract" or "overview" independently of their case and the second *Heading* annotation is the next available heading of the document.

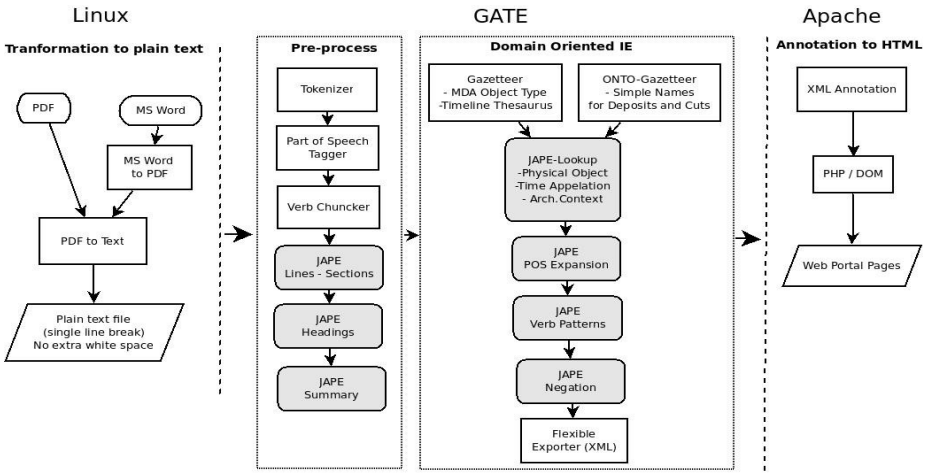

Fig. 2: The Information Extraction pipeline developed in GATE. Two separate exercises are shown here the Pre-processing and the Domain Oriented IE. Bespoke JAPE rules shown in Grey boxes, white boxes are used for GATE modules

3.3 Semantic Annotation Pipeline

The domain-oriented pipeline (Figure 2) extracts specific archaeological information utilising available EH terminology resources and the domain ontologies, CIDOC CRM and CRM-EH. The choice of ontological entities is based on project discussions with EH, specifically with the project collaborator Keith May. After discussion and consideration of available use case scenarios it was decided that the prototype system should focus on the extraction of the following CRM classes:

a) E19.Physical_Object described as "*items having physical boundaries that separate them completely in an objective way from other objects*" such as arrowhead, bone, pot, etc.

b) E49.Time_Appellation described as "*appellation of all forms of names or codes, such as historical periods, and dates, which are characteristically used to refer to a specific temporal extend that has a beginning an end and a duration*" such as Roman, Mediaeval, Bronze Age, etc.

c) E53.Place with emphasis on EHE0007.Context described as "*Spatial elements that constitute an individual archaeological unit of excavation including both primitive contexts and larger groupings of contexts*" such as pit, ditch, post-hole etc.

3.3.1 Transformation of Terminology Resources to GATE Gazetteers

The "Skosified" terminological resources were transformed into GATE gazetteer listings using XSLT transformation templates. In detail the following terminological resources were transformed to GATE gazetteer listing and used by the prototype

system; (i) the Archaeological Object Type thesaurus, (ii) the Time-line thesaurus and (iii) the EH glossary Simple Names for Deposits and Cuts.

GATE gazetteers allow the association of features with gazetteer lists as well as with particular list entries. Features can be accessed by JAPE grammars for the definition of matching expressions. For example a list containing month names might have a primary feature (Major Type) *date*, a secondary feature (Minor Type) *month*, whereas each entry of the list might be associated with a specialised entry for holding the three letter version of each month e.g. Jan for January, Feb for February etc. Similarly another list containing week days might be associated with the same primary feature *Date* but to have a different secondary feature for example *day*. A JAPE grammar can exploit the primary feature (Major Type) of *Date* in order to produce matches of both lists or it can be more specialised and exploit the secondary feature (Minor Type) for producing either month or day matches. Any annotations produced by the gazetteers lists would also be associated with the features specified by the gazetteer listing.

The prototype development experimented with two methods for making possible output annotations available to the JAPE grammars. In the first method, the Major and Minor gazetteer features were associated respectively with an ontological (CRM or CRM-EH) class reference and a *skos:concept* terminological reference (one of the EH glossaries or thesauri). In the second method, the *Simple Names for Deposits and Cuts glossary* was associated directly with the EHE0007.Context CRM-EH class, using the OWLIM GATE resource to represent the CRM-EH ontology. The incorporation of thesauri into GATE gazetteers was an immediate practical solution in the absence of available GATE resource without requiring the representation of thesauri as a formal OWL ontology within GATE. The thesauri employed do not follow a strict class relationship structure (this is common with many widely used thesauri) and asserting such relationships would be false.

3.3.2 JAPE Rules of the Semantic Annotation Pipeline

The prototype pipeline implemented fifteen different JAPE grammars for identifying the three main ontological concepts (Physical Object, Archaeological Context, and Time Appellation). The grammars exploit the *Major Type* gazetteer property for assigning the corresponding ontological reference to the matches, with the exception of Archaeological Context, which instead of the Major Type property used the CRM-EH class property, made via the OWLIM plug-in as discussed above. Additional rules were used for extending the initial Lookup annotations to include meaningful moderators. In the case of Time Appellation, two gazetteer listings were used for expanding over prefix terms (Earlier, Later, etc.) and suffix terms (Period, Century, etc.)

The following grammar matches three different cases of Time Appellation expansion. a) Expansion towards prefix and suffix i.e. “Early Roman Period”, b) Expansion only towards prefix i.e. “Early Mediaeval” and c) Expansion only towards suffix i.e. “Prehistoric period”

```
({Lookup.minorType==Date_Prefix}{TimeAppellation}
{Lookup.minorType==Date_Post})|
({Lookup.minorType==Date_Prefix}{TimeAppellation}|
{TimeAppellation}{Lookup.minorType == Date_Post})
```

Additionally, JAPE patterns identify rich phrases of entity pairs, such as Time Appellation and Physical Object i.e. “Roman Pottery” or Time Appellation and Archaeological Context, i.e. “Mediaeval Deposit”. This last approach is elaborated further by the definition of JAPE patterns which match linguistic evidence of combinations between entities and verb phrases in the form of <Entity><verb><Entity>, for example “...coins dating to Roman period...”, and Time Appellation and Archaeological Context as for example “...pits are of prehistoric date...”. This is intended as a step towards investigating more elaborate contextual-dependency of annotations in further work.

The following grammar matches phrases that connect Lookup annotation via verb phrases i.e. “pits are of prehistoric date”

```
{Context}({Token.kind==word}|{Token.category==","})?
{VG}({Token.kind==word}|{Token.category==","})?
{PhysicalObject}
```

The annotations produced are pairs of entities mostly involving Time, such as ‘Context + Time’ and ‘Physical Object + Time’ expressed as bespoke annotations not (yet) connected with the ontology. The patterns employ verb phrases rather than simple offsets with the aim of favouring precision in the assertion of the relation extraction. This is discussed further in section 4.3.

JAPE grammars are also used by the pipeline for matching negation in phrases. The negation detection is based on matching an offset of ten words which are followed after the negation phrases “no evidence”, “without evidence” and “absence of”. The negation phrases were included in a specific Gazetteer list carrying the Major Type attribute “Negation”

3.4 The Andronikos Web-Portal

The annotations delivered by the prototype system were exported from the GATE environment as XML files using the Flexible exporter plug-in. The plug-in produces XML outputs that couple content and annotation tags together, allowing for interoperable handling of the annotations. The Andronikos web-portal (http://andronikos.kyklos.co.uk) utilised the exported XML annotations. The objective of Andronikos development is to utilise the resulting semantic annotation XML files for making the annotations available in HTML hypertext document format. Server side PHP technology is employed to handle the annotations from the XML files and to generate the relevant web pages. The resultant pages were organised under a web-portal structure for presenting annotation versions of grey literature documents, such as pre-processing and ontological annotations.

Andronikos (Figure 3) was developed to assist the evaluation of the extraction phase by making the annotations available in an easy to follow human readable

format and to demonstrate the capability of linking textual representations to their semantic annotations. The portal makes use of the DOM XML for processing the XML files and for revealing the annotations of documents, while employing a MySQL database server to store thesauri structures relevant to the annotations. In addition, for visual inspection and initial evaluation purposes, CSS files present the XML files and highlight annotations with colours to assist recognition of annotations within text.

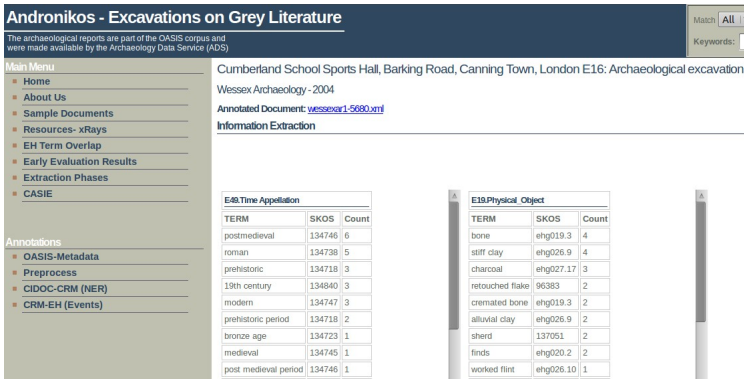

Fig 3. Andronikos Web portal, Semantic Annotations of Time Appellation and Physical object. Tables show the textual instance value, number of occurrences in document and the associated SKOS value (*postmedieval* and *post medieval period* share the same SKOS reference)

4. Evaluation of Semantic Annotations

The effectiveness of Information Extraction systems is measured by *Recall* and *Precision* rates. The measurement units originate from the IR domain but they have been redefined during the Machine Understanding Conference (MUC) to reflect matching and mismatching within the information extraction process (Grishman and Sundheim, 1996). According to the MUC definition, when the answer key is N_{key} and the system delivers $N_{correct}$ responses correctly and $N_{incorrect}$ incorrectly then

$$Recall = \frac{N_{correct}}{N_{key}} \quad \text{and} \quad Precision = \frac{N_{correct}}{N_{correct} + N_{incorrect}} \ .$$

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: $F_1 = 2\frac{Precision * Recall}{Precision + Recall}$. The score reaches its best value when 1 and its worst when 0. However, attempts to improve recall will usually cause precision to drop and vice versa. High scoring of F_1 is desirable since the unit can be used to test the overall accuracy of the system.

The evaluation task aimed at measuring the performance of the prototype information extraction mechanism with regards to Time Appellation, Physical Object

and Archaeological Context and the relations of Time Appellation with Physical Object and Time Appellation with Archaeological Context.

4.1 Evaluation process

The task had a largely investigative character, aiming not just to evaluate the performance of the prototype system but also to suggest development improvements to be taken on board by the full scale semantic annotation system. To evaluate system performance, a 'Gold Standard' (GS) test set of human annotated documents is typically employed for comparison with system produced automatic annotations.

Another aim of the pilot evaluation was to investigate the evaluation methodology and the difficulty of annotating archaeological reports with ontology entities. Thus the degree to which different annotators might agree or disagree and the influence of specialist domain knowledge was also of interest. Tools are provided within GATE to calculate an Inter-Annotator Agreement score (IAA) from separate annotations (Maynard, Peters and, Li. 2006). The creation of the GS is normally a collective effort of human annotators in order to achieve coverage of a wide sample range. Provision of a single and commonly agreed set of GS annotations is a subject of agreement between human annotator experts.

Within the constraints of the pilot investigation, four annotators provided manual annotation of 10 summary extracts originating from 5 archaeological 'excavation' and 5 'site evaluation' grey literature reports. One annotator was the system developer (AV), two annotators were STAR project members (CB, DT) and one was a senior archaeologist (KM). Each summary extract was annotated by all four annotators in order to get a pluralistic view of annotator agreement. The four manual annotation sets were collected and processed by the IAA GATE plug-in, delivering the results of Table 1.

The quadruple annotation was conducted in order to engage enough annotators for revealing pluralistic annotation results which could inform the development of the full scale system. For the pilot evaluation purposes, the annotations of the senior archaeology expert (KM) were treated as the GS for the evaluation, since the other annotators did not have the same level of specialist domain knowledge. The annotation input of the senior archaeology expert was selected as the most appropriate for delivering indicative performance results that correspond to the investigating focus of the study.

4.2 Evaluation Results

Examination of the IAA results (Table 1) of the four annotators reveal a low agreement score. This appears typical of manual annotation in an archaeological context. Zhang et al. (2010) agree with Byrne (2007) that manual document annotation in archaeology is a challenging task due to domain specific issues such as complexity of language, uncertainties, composite terms, acronyms and so on with overall IAA score ranging below 60%. The overall F_1 agreement score for all four

annotators is 51%, whereas the agreement score between different pairs varies from 35% to 65%.

Table 1 Inter-Annotator agreement score of the different pairs

	Recall	Precision	F-Measure
All-Pairs	0.63	0.43	0.51
AV-CB	0.62	0.37	0.47
AV-DT	0.60	0.30	0.40
AV-KM	0.57	0.26	0.35
CB-DT	0.72	0.60	0.65
CB-KM	0.66	0.50	0.57
DT-KM	0.63	0.57	0.60

The lowest agreement score is between AV-KM where AV is the system developer and KM an archaeology expert while the highest score is between CB-DT where both are STAR project members. To some extent, the low agreement between annotators reflects the end-user focus. The evaluation was directed towards the (cross search retrieval) aims of the broader STAR project, being oriented to the audience of archaeology researchers and HE users. The instructions for evaluators were intended to be relevant to future cross search and hence neither the scope of the ontology elements, nor the associated vocabulary were specified exhaustively; annotators were expected to exercise judgment. The instructions directed annotators to identify textual instances of the targeted concepts including adjectival moderators and “rich” phrases containing two or more concepts. Information that could influence annotators, such as pattern matching clues and vocabulary coverage, was not made available. Hence, there was a significant difference between AV, the developer with a clear understanding of the system's functionality and vocabulary coverage, and KM, an archaeology expert with knowledge of the domain.

One major difference between the AV and KM was in the recognised vocabulary. As discussed in section 1.7, archaeology differs from other IE applications in that it employs many common words in a discipline specific manner. For example, AV followed precisely the ‘Simple Names for Deposits and Cuts’ Glossary, while KM exercised judgment and included words missing from the glossary, such as ‘road’, ‘occupation’ and ‘charcoal’ (interpreting Ecofacts as ‘objects’ along with Artefacts). Furthermore the scope of ontology elements is somewhat fuzzy at the boundaries – terms such as ‘villa’ and ‘settlement’ may be treated a little differently by different archaeologists according to context. KM did not annotate mentions of the ‘trenches’ dug as part of the excavation which were however annotated (incorrectly) by AV following a more literal approach. Additionally the issue of whether moderators and articles are included in an annotation and the scope of a rich phrase containing relations can affect results.

Table 2 System's performance for three ontological entities and for two relations (Context + Time and Physical Object + Time)

	Recall	Precision	F-Measure
Context	0.47	0.70	0.57
Physical Object	0.40	0.45	0.42
Time Appellation	0.70	0.96	0.81
Context + Time	0.38	0.75	0.50
Physical Object + Time	0.60	0.60	0.60
Overall	0.51	0.69	0.58

The prototype system performs well against Time Appellation entities delivering F_1 score 81% while it delivers reasonably good *Precision* for Context entities 70% and for Context plus Time relations 75% (Table 2). On the other hand *Recall* rates for Context and Physical Object entities are low (47% and 40%), which contributes to relatively low F1 scores. The system manages to extract relations with some limited success delivering F_1 score 55% (average) on relation extraction, although it only implements very basic matching rules.

4.3 Discussion

Although, results are not at an operational level, the evaluation suggests the potential of the method for identifying a set of ontological entities and relations. The overall F_1 score of the prototype system is 58% (Table 2) which is considered encouraging as a basis for further elaboration by the full-scale system, as discussed below. For comparison with other archaeological IE systems, semantic annotation systems targeted at archaeological entities have yielded F_1 score of 75% (Zhang et al. 2010), while systems targeted at historical text have yielded F_1 score of 73% (Grover et.al 2008).

The limited use of terminological resources in particular for the *Physical Objec*t entity has adversely affected Recall. The prototype delivered a low Recall rate (40%) mainly due to limited vocabulary coverage. Although, the MDA Object Thesaurus comprises approximately 4000 concepts, it does not contain concepts such as ‘finds’ and ‘samples’ that are relevant to excavation reports. Similarly, there proved to be a significant vocabulary deficit for archaeological contexts (places), as discussed above.

Lessons learned include the need to employ archaeologist-annotators in future evaluation for our project aims and to consider carefully the instructions for annotators. Future full-scale development will seek to improve the current prototype in order to deliver operational results. The current system can be improved by including additional specialised vocabulary resources in order to increase Recall. This includes further vocabulary for both finds (objects) and archaeological ‘contexts’ in the excavation. For the former, it is possible to draw on further EH glossaries for small finds and possibly materials sometimes treated as finds. For the latter, the EH

Monuments Type Thesaurus offers further vocabulary resources beyond the Simple Names glossary. Since there is no one integrated vocabulary resource, more sophisticated methods for combining thesauri with glossaries (word lists) will be investigated, For example, a core set of glossary terms might be expanded via the thesaurus to enable a selective use of the thesaurus vocabulary, without harming Precision by using too much irrelevant vocabulary. It is possible that employing additional vocabulary resources may result in a trade off between Recall and Precision rather than a simple gain in Recall. Therefore it will be necessary to closely consider the context of the vocabulary, as discussed in Section 1.7. NLP techniques, such as Word Sense disambiguation, Negation Detection and Part of Speech validation may be useful here.

The terminological resources should be enhanced to include spelling variations such as hyphenation, for example *post hole* and *post-hole.* The system should also be capable of exploiting the available vocabulary independently of plural or singular forms. The volume of false positive matches should be reduced by the use of Part of Speech input, which can be used for validating matches in order to distinguish verb from noun forms eg *Building*. Additional validation techniques such as word pair disambiguation can be invoked to improve precision, while negation detection can be further refined.

The prototype has managed to extract rich phrases revealing relations between CRM entities, using the simple JAPE grammars described in section 3.3.2. Although, current results are fairly low at 55%, we believe the methods have potential to target phrases carrying rich contextual evidence. More elaborate relation extraction methods will be used to deliver the specialised archaeological relations expressed by the CRM-EH model. Currently the system produces custom annotations – the ontology needs to be analysed to identify the appropriate relations between ontology elements and deliver results in ontological terms. The CRM (and CRM-EH) ontologies are event-based – the precise implications for IE techniques and patterns need to be explored. Neither the current verb phrase pattern methods nor simple offset based methods of combining named entities appear likely to yield results with sufficient precision. Instead we intend to investigate methods of relation extraction that use more sophisticated pattern-matching grammars based on likely syntactical constructs, in order to improve the performance of relation extraction.

5 Conclusions

This paper reports results from a prototype development and evaluation which is part of an ongoing larger scale GATE development effort and evaluation. The results reported show that information extraction techniques can be applied to archaeological grey literature reports in order to produce annotations in terms of the CIDOC CRM ontology. The development shows that it is possible to employ a complementary use of ontology and thesauri (plus glossaries) and extract both SKOS terminological elements for subsequent use in retrieval and CRM ontological elements for purposes of data integration and possible logical inferencing.

The initial evaluation results are not at an operational level. However they suggest the methods have potential when improved further by the steps outlined in section 4. These include further use of use of Part of Speech input, expanding the vocabulary resources for both objects and archaeological contexts and further refining the relation extraction techniques. The evaluation also highlights methodological issues arising from the nature of the archaeology domain and the cross search aims of the STAR project, which aims to integrate different archaeological datasets and grey literature via the CRM ontology. Further evaluation will seek to involve representative archaeological end-users.

The development has achieved its aims for the implementation of a prototype semantic annotation system capable of extracting concepts from archaeological grey literature with respect to both domain ontology and terminological resources. Further consideration is needed as to how to reflect the provenance of the IE contributions in the semantic search system; generally NLP elements from text reports are less reliable as 'facts' than extracted elements from excavation datasets.. The prototype techniques have also demonstrated the capability of semantic annotation to carry ontological and terminological references that can be used to support information retrieval with respect to semantics. The GATE framework has been negotiable in modification of its resources while JAPE grammars have proved flexible and robust for expressing grammars targeted at the extraction of CRM and CRM-EH entities and relations.

Acknowledgements

The STAR project was supported by the Arts and Humanities Research Council [grant number AH/D001528/1]. Thanks are due to the Archaeology Data Service for provision of the OASIS corpus and for many helpful discussions, to Phil Carlisle (English Heritage) for providing domain thesauri and for helpful input from Renato Souza (Visiting Fellow at Glamorgan) and Ceri Binding (Hypermedia Research Unit, University of Glamorgan). Special thanks are due to Keith May (English Heritage) whose comments and archaeology expertise have helped to improve this paper.

References

Ananiadou S. Friedman C, Tsujii J. (2004) Introduction: named entity recognition in biomedicine', Guest Editorial / Journal of Biomedical Informatics 37 (2004), pp 393–395

Ananiadou S, Chruszcz J, Keane J, McNaught J, Watry P. (2005) The National Centre for Text Mining: Aims and Objectives. Ariadne 42, http://www.ariadne.ac.uk/issue42/ananiadou

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) 'The Semantic Web', Scientific American, Vol.284 No.1, pp.28–37

Binding, C., Tudhope, D., May, K. (2008) 'Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM', Proceedings (ECDL 2008) 12th European Conference on Research and Advanced Technology for Digital Libraries, Springer-Verlag, Berlin, Germany pp.280-290

Bontcheva K., Tablan V., Maynard D., Cunningham H. (2004) 'Evolving GATE to Meet New Challenges in Language Engineering', Natural Language Engineering Vol.10 No.3/4, pp. 349-373

Bontcheva, K., Cunningham, H., Kiryakov, A., Tablan, V.(2006) Semantic Annotation and Human Language Technology, in Davies, J. et al (Eds.), Semantic Web Semantic Web Technology: Trends and Research in Ontology Based Systems, John Wiley and Sons. Sussex

Bontcheva, K., Li, Y., Cunningham, H.(2007) 'Hierarchical, Perceptron-like Learning for Ontology Based Information Extraction', Proceedings of the 16th International World Wide Web Conference, ACM, New York, USA, pp.777–786

Byrne K. (2007) 'Nested named entity recognition in historical archive text'. In Proceedings of the International Conference on Semantic Computing (ICSC 2007), Irvine, California, pp. 589 596.

Cowie, J. and Lehnert, W. (1996) 'Information extraction', Communications ACM, Vol. 39 No. 1, pp.80-91

Cripps, P., Greenhalgh, A., Fellows, D., May, K., and Robinson D. E. (2004) Ontological Modelling of the work of the Centre for Archaeology. CRM – EH model diagram. English Heritage, Portsmouth, England

Crofts, N., Doerr, M., Gill, T., Stead, S., and Stiff, M. (2011) Definition of the CIDOC Conceptual Reference Model. Available at http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf (Accessed 6 March 2012)

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan V. (2002) 'GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications', Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics ACL'02, Philadelphia, USA. Available at http://gate.ac.uk/sale/acl02/acl-main.pdf (Accessed 18 January 2012)

Cunningham, H., Maynard, D., and Tablan, V. (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(Accessed 18 January 2012)

Debachere M.,C. (1995) 'Problems in Obtaining Grey Literature', Journal of the International Federation of Library Associations and Institutions, Vol. 21 No.2, pp 94-106

English Heritage Recording Manual (2006) English Heritage internal document

Francis, W.,N. and Kucera, H. (1964) Brown Corpus Manual. Brown University, Available at http://icame.uib.no/brown/bcm.html (Accessed 18 January 2012)

Gaizauskas, R., and Wilks, Y. (1998) 'Information extraction: beyond document retrieval', Journal of Documentation, Vol. 54 No. 1, pp.70-105

Gradmann S. (2010) 'Knowledge = Information in context: on the importance of semantic contextualisation in Europeana', Europeana White Paper, Available at http://version1.europeana.eu/web/europeana-project/whitepapers (Accessed 18 January 2012)

Grishman, R., Sundheim, B. (1996) 'Message Understanding Conference-6; a brief history', Association for Computational Linguistics, Vol. 1 No.1, pp.466–471

Grover C., Givon S., Tobin R., and Ball J. (2008) 'Named entity recognition for digitised historical texts'. In Proceedings of International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco

Guarino, N. (1998), 'Formal Ontology and Information Systems', in N. Guarino, (ed.), Formal Ontology in Information Systems, IOS Press, pp.3–15.

Isaac A., Summers E. (2009) 'SKOS Simple Knowledge Organization System Primer', Available at http://www.w3.org/TR/skos-primer/ (Accessed 18 January 2012)

ISO 25964. Information and documentation – Thesauri and interoperability with other vocabularies. http://www.niso.org/schemas/iso25964/

Jeffrey S., Richards J., Ciravegna F., Waller S., Chapman S., Zhang Z. (2009) ‘The Archaeotools project: faceted classification and natural language processing in an archaeological context’ in P. Coveney (Eds.), Special Themed Issue of the Philosophical Transactions of the Royal Society A, Crossing Boundaries: Computational Science, E-Science and Global E-Infrastructures, Vol.367, pp.2507-19

Kiryakov, A., Popov, B., Terziev, I., Manov, D. and Ognyanoff, D. (2004) 'Semantic Annotation, Indexing, and Retrieval', Web Semantics: Science, Services and Agents on the World Wide Web, Vol.2 No.1, pp. 49–79

Lewis, D. and Jones, K. (1996) 'Natural Language Processing for Information Retrieval', Communications ACM, Vol. 39 No. 1, pp. 92–101

Mascardi V, Cordì Valentina , Rosso. 2006. A Comparison of Upper Ontologies. Technical Report DISI-TR-06-21. Dipartimento di Informatica e Scienze dell’Informazione (DISI), Università degli Studi di Genova.

Maynard D., Yankova M., Kourakis A., and Kokossis A. (2005) 'Ontology-based information extraction for market monitoring and technology watch', In ESWC Workshop “End User Apects of the Semantic Web”, Heraklion, Crete

Maynard, D., Peters, W. and Li, Y. (2006) 'Metrics for Evaluation of Ontology-based Information Extraction' in Proceedings of WWW 2006 Workshop on "Evaluation of Ontologies for the Web". 22 May 2006. Edinburgh, Scotland

Moens, M.F. (2006) Information Extraction Algorithms and Prospects in a Retrieval Context. Springer, New York

National Monuments Record Thesauri (online), Available at http://thesaurus.english-heritage.org.uk/ (Accessed 18 January 2012)

Ohta,T., Tateisi,Y. and Kim,J. (2002) 'The GENIA Corpus: an annotated research abstract corpus in molecular biology domain.' In Proceedings of the second international conference on Human Language Technology Research, pp 82-86.

Online AccesS to the Index of archaeological investigations OASIS (online), Available at http://oasis.ac.uk/ (Accessed 18 January 2012)

STAR project (online) Available at http://hypermedia.research.glam.ac.uk/kos/star/ (Accessed 18 January 2012)

Tsujii J., Ananiadou S. (2005). ‘Thesaurus or logical ontology, which one do we need for text mining?’ Language Resources and Evaluation, Vol.39, No.1, pp.77-90.

Tudhope D., Koch T., and Heery R. (2006) 'Terminology Services and Technology', JISC state of the art review

Tudhope D., May K., Binding C., Vlachidis A. (2011) 'Connecting archaeological data and grey literature via semantic cross search', Internet Archaeology, Vol.30, Available at http://intarch.ac.uk/journal/issue30/tudhope_index.html/ (Accessed 18 January 2012)

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E. and Ciravegna F. (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, Vol. 4 No. 1, pp 14-28

Vlachidis, A., Binding, C., May K. and Tudhope, D. (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 journal, Vol. 62, No.4-5, pp.466-475

Wilks, Y. (2008) 'The Semantic Web as the apotheosis of annotation, but what are its semantics?' Intelligent Systems, Vol. 23 No. 3, pp.41–49

W3C Library Linked Data Incubator Group Report. (2011). Available at http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/ (Accessed 6 March 2012)

Zhang Z., Chapman S., Ciravegna F. (2010) 'A Methodology towards Effective and Efficient Manual Document Annotation: Addressing Annotator Discrepancy and Annotation Quality'. Lecture Notes in Computer Science, 6317 pp301–315