A Knowledge Based Approach to Information Extraction for Semantic Interoperability in the Archaeology Domain

Andreas Vlachidis
Faculty of Computing, Engineering and Science,
University of South Wales,
Pontypridd, CF37 1DL,
Wales, UK.
Email: andreas.vlachidis@southwales.ac.uk

Douglas Tudhope
Faculty of Computing, Engineering and Science,
University of South Wales,
Pontypridd, CF37 1DL,
Wales, UK.
Email: douglas.tudhope@southwales.ac.uk

Abstract

The paper presents a method for automatic semantic indexing of archaeological grey-literature reports using empirical (rule-based) Information Extraction techniques in combination with domain-specific knowledge organization systems. Performance is evaluated via the Gold Standard method. The semantic annotation system (OPTIMA) performs the tasks of Named Entity Recognition, Relation Extraction, Negation Detection and Word Sense disambiguation using hand-crafted rules and terminological resources for associating contextual abstractions with classes of the standard ontology (ISO 21127:2006) CIDOC Conceptual Reference Model (CRM) for cultural heritage and its archaeological extension, CRM-EH, together with concepts from English Heritage thesauri and glossaries.

Relation Extraction performance benefits from a syntactic based definition of relation extraction patterns derived from domain oriented corpus analysis. The evaluation also shows clear benefit in the use of assistive NLP modules relating to word-sense disambiguation, negation detection and noun phrase validation, together with controlled thesaurus expansion.

The semantic indexing results demonstrate the capacity of rule-based Information Extraction techniques to deliver interoperable semantic abstractions (semantic annotations) with respect to the CIDOC CRM and archaeological thesauri. Major contributions include recognition of relevant entities using shallow parsing NLP techniques driven by a complimentary use of ontological and terminological domain resources and empirical derivation of context-driven relation extraction rules for the recognition of semantic relationships from phrases of unstructured text. The semantic annotations have proven capable of supporting semantic query, document study and cross-searching via the ontology framework.

Introduction

Controlled vocabularies and semantically structured Knowledge Organization Systems (KOS) offer key resources for a variety of information science applications across a wide range of subject domains. They are seen as one route to overcoming the 'vocabulary problem' posed by differing terminology use by indexer and searcher (Bates, 1986). They support a range of use cases including (faceted) browsing, information indexing and retrieval, and document classification (see for example Golub et al. 2014, particularly with respect to KOS registries and application profile) and a variety of techniques for mapping between vocabularies (reviewed in Zeng and Chan 2004). This paper is concerned with the broad area of automated subject metadata generation.

In a review of different approaches to the automated subject classification of documents, Golub (2006) distinguishes three main approaches: *text categorization,* a machine learning approach based on predefined categories; *document clustering,* an unsupervised (information retrieval) approach which derives clusters of documents by statistical means; *document classification,* a (library science) approach based on intellectually created knowledge organization systems (such as classification schemes). The work discussed here has evolved out of the document classification approach. Library classification schemes, such as the Dewey Decimal Classification (DDC) have underpinned efforts in automatic subject metadata generation for many years and continue to be employed (eg Thompson, Shafer, Vizine-Goetz 1997; Golub, Lykke, Tudhope 2014). This paper reveals the role of information extraction for purposes of semantic indexing, as opposed to automated subject classification, via ontology based approaches rather than traditional library classifications.

Information Extraction (IE) is defined as a text analysis task aimed at extracting targeted information from context in the document (Cowie & Lehnert, 1996). Integrated with computational artefacts, such as information system ontologies that provide a common conceptual ground, IE can deliver a specialised form of document abstraction, known as semantic annotation. This connects natural language text with formal conceptual structures in order to enable new information access and to enhance existing information retrieval processes (Uren et al. 2006). The output of IE, in this case annotation for the purposes of semantic indexing, delivers abstractions of entities and relations enabling retrieval of facts and findings in context rather than the monolithic retrieval of documents as a whole.

Within the subject domain of archaeology, vocabulary standards have been envisaged as a potential solution to the data inaccessibility imposed by a fragmented fieldwork practice hindering the publication and dissemination of archaeological information (Richards & Hardman, 2008). Adoption of interoperable standards for encoding and dissemination of archaeological data and information would not only enable cross-searching and meta research studies of structured content but also offer an opportunity for a fundamental change in archaeological practice in terms of recording and disseminating fieldwork data and reports (Falkingham, 2005). There is a vast reservoir of archaeological reports largely untapped, for meta-research or crossing searching and comparison with published datasets.

The CIDOC CRM (ISO Standard 21127:2006) conceptual model for cultural heritage, in combination with standard domain thesauri and glossaries, offers the potential of a semantic-enabled architecture that supports information extraction and retrieval via a common layer of shared understanding of domain concepts and relationships (Crofts, Doerr, Gill, Stead, & Stiff, 2009). The architecture can enable cross-searching between disparate information resources. In the case of archaeology, being able to search across fieldwork reports and excavation data can significantly enhance the information seeking activities of research scholars and domain professionals (Tudhope, May, Binding, & Vlachidis, 2011).

The information extraction work described in this paper is based on a rule-based information extraction application driven by the CIDOC-CRM ontology supported by a range of domain vocabulary originating from English Heritage. Rule-based information extraction approaches do not rely on training data, as in the case of supervised text categorization (machine learning) approaches, for delivering results. Instead, a set of hand-crafted rules are supported by KOS resources, such as thesauri, glossaries, gazetteers and ontologies. The integration of different vocabulary sources is a distinctive aspect of the work, which employs a complementary use of ontology classes and the corresponding thesaurus concepts in information extraction, both expressed as URIs. Consuming applications may make use of both in combination, or one separately, depending on the use case, as discussed in the recent Thesaurus Standard (ISO25964-2 2013, ch 21).

Current state of the art portals offering access to collections of archaeological datasets and reports typically employ a faceted browsing interface combined with string search over metadata (Richards et al., 2011). The excavation report metadata consists of major objects (*finds)* sometimes with associated material, time period and the *context* (eg *'post-hole')* they were found in. The location and date of the excavation (or other archaeological intervention) are usually included, though that is not the focus of the information extraction described in this paper. The metadata can vary according to the practice of author/indexer but may sometimes include several terms for different temporal phases of the site, perhaps consisting of finds or evidence of settlement from different time periods. If best practice is followed the indexing refers to concepts from standard controlled vocabularies (such as the English Heritage thesauri). While the intellectual indexing can contain coordinated pairings of (say) object and period, typically this relationship is not recorded in the metadata and semantic search or faceted browsing of meaningfully associated terms is not possible (string phrase search gives limited help). Thus 'false drops' on pairings such as *'Roman urn'* can be quite common in searches or browsing sequences.

Particularly in the case of archaeological grey literature documents, aggregation of index terms on the basis of a simple co-existence does not guarantee their meaningful association at a contextual level. Archaeology site evaluation and excavation reports are typically long documents covering a wide range of excavation phases. Such reports may be indexed with a long list of terms covering various periods, materials, places, finds etc. and consequently are prone to false positive matching. Connecting faceted classification terms at a document retrieval level, for example What: *urn* and When: *Roman*, can result in matches albeit the index terms might not be coherently associated in text.

Taking full advantage of archaeological grey literature in research requires information retrieval systems capable of aiding semantically driven search scenarios where user query arguments reflect meaningful contextual associations (Tudhope et al., 2011). An example might be to retrieve across datasets and document collections physical objects of a certain type, which relate to a particular period or recovered from archaeological context, such as, *'arrowhead finds that relate to the Neolithic period'* or *'ditch containing pottery finds and flint flakes'*. To satisfy such search scenarios it is necessary to produce semantic metadata that can be utilised by an information retrieval system on the application layer.

This paper discusses a case study (with evaluation) of the automatic information extraction of meaningful entities and relationships from English language archaeological reports. This semantic indexing is targeted at supporting complex and semantically defined queries that facilitate information retrieval from archaeological grey literature reports and cross searching over reports and datasets. Since the aim is to make use of the results in semantic search applications, it is important to orient to the semantic vocabulary standards that search tools (or Linked Data) will employ. Thus the semantic indexing conforms to the CIDOC CRM core ontology, together with complimentary domain vocabulary from English Heritage (EH) thesauri and glossaries (the CIDOC CRM does not have a built-in terminology).

A variety of existing vocabularies were employed. An early pilot investigation (Vlachidis & Tudhope, 2012) employed a glossary for types of context and a single (Object) thesaurus but results indicated a vocabulary deficit particularly for places (combined with a highly contextual use of terminology within archaeology). The final system combines a variety of glossaries mainly taken from the EH (excavation) Recording Manual with various broader cultural heritage thesauri. Relying only on the recording system glossaries limits IE vocabulary coverage of the final excavation reports, which employ more elaborate terminology. On the other hand, employing the full thesauri can entail loss of precision. One strand of the evaluation reported here investigates the effect of different forms of semantic expansion from glossary to thesaurus and over different thesaurus relationships.

The paper presents the semantic annotation application, together with the archaeology case study and an evaluation based on the Gold Standard method. Related work and literature is presented and novel contributions to Named Entity Recognition (NER) and Relation Extraction (RE) techniques are discussed. These include use of shallow parsing NLP techniques over unstructured text for the recognition of CIDOC-CRM domain entities, use of contextual driven rules for the recognition of binary semantic relationships of interest and complimentary application of ontological and terminological definition for the purposes of semantic interoperability. The final part of the paper discusses results from an evaluation process that employed a set of manually annotated documents produced by archaeologists for the case study, in order to measure the system's performance in terms of Recall, Precision and F-Measure rates. The paper concludes with overall results of the semantic indexing effort and suggestions for future work.

Related Work

Since 1990 a significant increase in archaeological investigations in England and Wales has resulted in a large number of fieldwork reports, often called 'grey literature', which may reflect different stages of a fieldwork project (Falkingham, 2005). Grey literature is not published in the conventional sense and not always easily accessible to the general public or archaeology researchers. The laborious process of finding information in such reports is a major hindrance to the development of archaeological research. Researchers are required to read through large pieces of text, if not the whole document, in order to find new information about a particular period or a find type. University teaching cannot keep up to date with the latest discoveries and 'archaeologists of tomorrow are being taught the archaeology of yesterday' (Hardman & Richards, 2003). Thus, it is highly desirable to be able to search effectively within and across archaeological reports.

The work described in this paper contributed semantic indexing of grey literature in the Semantic Technologies for Archaeological Resources (STAR) project which, in collaboration with English Heritage (EH), addressed the above issues. STAR integrated data from diverse archaeological datasets and grey literature and demonstrated semantic cross search (Tudhope et al., 2011). The CIDOC CRM (ISO 21127:2006) formed the underlying conceptual framework, including classes, such as Physical Things, Places, Temporal Entities, with Events as the core of the model (Crofts et al., 2009). Types allow further detailed classification of any class instance. For its work in the archaeology domain, the STAR project adopted the English Heritage extension of the CIDOC Conceptual Reference Model (CRM-EH) with its archaeological subclasses.

Semantic Annotation produces metadata with respect to a given ontology, allowing users to search across textual resources for entities and relations instead of words (Bontcheva, Duke, Glover, & Kings, 2006; Uren et al., 2006). Information Extraction techniques automatically recognise, extract and associate information snippets with semantic elements. Named Entity Recognition (NER) is a subtask of Information Extraction aimed at the recognition and classification of units of information to predefined categories (Grishman & Sundheim, 1996; Nadeau & Sekine, 2007). In the cultural heritage domain, NER is evident in a range of projects. Grover, Givon, Tobin, & Ball (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. Byrne (2007) focused on NER from historical archive texts, originating from the Royal Commission on the Ancient and Historical Monuments of Scotland (RCAHMS) via a machine learning (ML) approach based on a maximum entropy classifier.

The Archaeotools project employed NLP techniques for enabling access to site metadata and archaeological reports via a faceted classification (Richards et al., 2011). For the NER task, the Archaeotools project adopted a hybrid approach, incorporating both machine learning and rule-based methods, with machine learning techniques being prominent. A rule-based approach was adopted for the identification of regular context, such as bibliographical information. On the other hand, machine learning was followed for the identification of entities, such as place names, temporal information, event dates and subjects (although training set definition proved challenging and time consuming Jeffrey et al., 2009; Zhang, Chapman, Ciravegna, 2010).

The employment of rule-based IE and complementary use of ontological and domain vocabulary resources distinguishes the work presented here from supervised machine learning work, which relies on the existence and quality of training data. The absence of a training corpus coupled with the availability of a significant volume of high quality domain-specific knowledge organization resources, such as a conceptual model, thesauri and glossaries were contributing factors to the adoption of rule-based techniques in the research.

Crucially, the semantic annotation is not only targeted at producing indexing abstractions to support semantic retrieval on a document level but also at the retrieval of significant phrases of information that can be modelled with ontological relationships. This is achieved via the IE technique known as Relation Extraction, sometimes referred to as Relation Detection and Recognition task. (US-NIST, 2003). The Automatic Content Extraction (ACE) programme defined relation extraction as an inference task addressing explicit relations between two entities that occur within a common syntactic construction (US-NIST 2003).

Rule-based systems, such as GENIES (Friedman, Kra, Yu, Krauthammer, & Rzhetsky, 2001),GenIE (Cimiano, Reyle, & Saric, 2005), and Genescene (Leroy & Chen, 2005) demonstrate the capacity of ontology based systems to tackle domain specific RE tasks in the biomedical domain with some success. GENIES uses a full parsing strategy combined with sub-language grammar constraints and domain dictionaries to extract information about cellular pathways. GenIE is an ontology-driven system that uses linguistic analysis and semantic representation formalisms to extract information on biochemical pathways and functions of proteins, while Genescene extracts complementary biomedical relations between noun phrases from MEDLINE abstracts via sentence structures expressed as relation templates. However, the capability of rule-based, ontology guided systems to tackle the task of RE in the archaeology domain has not yet been explored. Extracting semantic information conforming to the standard CIDOC CRM ontology and archaeological thesauri offers the potential for semantic interoperability with data resulting from other initiatives following the same standards.

The work presented here differs from previous cultural heritage work that has detected CIDOC CRM entities via intellectual methods (Ore & Eide, 2009). Some parallels can be drawn with Byrne and Ewan (2010) due to the comparable aims of NER and RE over archaeological text. However, the use of probabilistic ML techniques and absence of ontology, in particular the CRM or CRM-EH from their work, yields a significantly different method of IE than the one discussed in this paper. An early prototype used simple rules to identify basic coexistence of entities in phrases but this does not necessarily constitute detection of relations or events (Vlachidis & Tudhope, 2012). The full scale system discussed here employs a complex series of relation extraction rules, which, as discussed in the evaluation section, have significantly improved the precision rates of the system compared to the prototype.

The pipeline uses hand-crafted rules that employ syntactical structures for the detection of textual phrases potentially useful to archaeological research. The extraction method follows a shallow parsing strategy based on part of speech tag, entity types and domain dictionaries. Other projects have also found shallow parsing useful for tackling the task of relation extraction over binary representations (Zelenko, Aone, & Richardella, 2003), while deep parsing can be useful in the extraction of complex events, such as interactions between biological components (Ananiadou, Pysysalo, Tsujii, & Kell, 2010). The annotation technique is broadly informed by the ACE definition of Relation Detection and Recognition tasks (US-NIST, 2004).. A binary definition of relation is adopted, where each relation phrase consists of two arguments identified by a unique ID and a role.

The extracted relation phrases are modelled as CRM-EH events, which differ from the ACE definition of events. ACE events involve zero or more entities, values and time expressions, whereas the CRM-EH events targeted by the pipeline, connect CRM entities in a binary form. The aim is to detect phrases which can be modelled as CRM-EH events or properties, explicitly or implicitly mentioned in the text. Thus, neither the extent of the ACE task of Event Detection and Recognition nor the ACE event types, subtypes and attributes are appropriate to the pipeline.

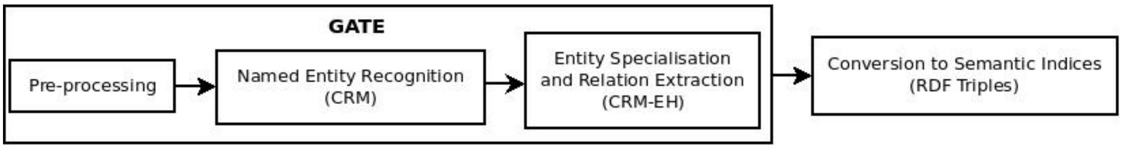
The pair of entities that participate in an event phrase are the arguments of the event. For example, the phrase *'hearth contains coin of the Roman period'* delivers two CRM-EH events (context find deposition event and context find production event). One event connects *'hearth'* and *'coin'* and another event connects the same *'coin'* with the *'Roman period',* with events having *'coin'* as a common argument. The events are implicitly defined in this example; there is no explicit mention of the event that deposited the coin in the hearth nor how the coin was originally produced. However, it can be assumed that since the coin has been found in the hearth, it must have been deposited in that place and since the coin is described as Roman, it has probably been produced during the Roman period (modelling the full complexity of spatio-temporal style periods is outside the scope of this paper). This modelling of events differs from the ML technique followed by Byrne and Ewan (2010), which detects events as mentions of verb phrases carrying a single event type (which might contain several arguments). The OPTIMA pipeline is driven by the CRM-EH ontological structure and generates event types defined by standard ontological definitions, which can be exploited by retrieval applications or used for information integration.

OPTIMA Pipeline

The OPTIMA pipeline is developed within the General Architecture for Text Engineering (GATE) environment, which is an NLP framework that provides the architecture and the development environment for developing and deploying natural language software components (Cunningham, Maynard, Bontcheva, & Tablan, 2002). Semantic approaches have been applied to GATE involving general concepts such as Persons, Organizations, Places and Date (Bontcheva, Tablan, Maynard, & Cunningham, 2004). The JAPE (Java Annotation Pattern Engine) language supports the definition of IE rules. JAPE grammar is a finite state transducer, which uses regular expressions for handling pattern-matching rules (Cunningham et al., 2000). The architecture allows the integration of new JAPE rules that extract information for specific IE goals. The main IE tasks of the OPTIMA pipeline employ such hand-crafted JAPE rules and terminology resources in the form of GATE gazetteers to deliver the semantic annotation result.

The semantic indexing process is conducted by the OPTIMA pipeline, which loosely takes its name from the four CRM entities, (Physical) Object, Place, Time Appellation and Material, targeted by the main NER phase. OPTIMA is described as a 'pipeline' due to the cascading processing order in which several NLP tasks and sub-tasks are invoked to deliver the semantic annotation results. The process of semantic indexing is divided into four main phases as seen in Figure 3. The first phase (Pre-processing) delivers a set of domain independent annotations, which are utilised further by the NER and RE phases (Vlachidis & Tudhope, 2012). The second and third phases of the pipeline focus on the task of NER and RE respectively and are discussed in detail in the following section. The last phase of the pipeline is the only phase conducted outside the GATE environment and delivers the RDF output of the semantic annotation.

FIG. 3. The main phases of the OPTIMA pipeline; the first three phases developed in GATE framework of natural language engineering, the last phase executed using bespoke PHP scripts.


GATE
Pre-processing
Named Entity Recognition (CRM)
Entity Specialisation and Relation Extraction (CRM-EH)
Conversion to Semantic Indices (RDF Triples)

Ontology and Vocabulary Resources

A complimentary use of ontological and vocabulary resources permitted semantic annotations to maintain a distinction between ontological and terminological references (both are needed for retrieval). The English Heritage (EH) vocabulary resources (English Heritage, 2014); Archaeological Object Thesaurus; Building Material Thesaurus; Monument Type Thesaurus; Period Thesaurus and a range of EH Recording manual glossaries were compiled as GATE gazetteer listings with individual entries carrying a SKOS terminological reference. These SKOS (Isaac & Summers, 2009) references are exposed to Information Extraction rules (JAPE grammars) enabling exploitation of broader – narrower thesaurus relationships. The semantics of the information retrieval thesaurus hierarchical relationship is looser than the formal ontological class-subclass relationship, being designed to support search and browsing use cases in retrieval (ISO25964-1:2011).

The task of NER is also supported by a small number of supplementary gazetteers. This specialised vocabulary is used in combination with Part of Speech (POS) input to support the tasks of Word-Sense disambiguation, Adjectival Conjunction and Negation Detection. Thesauri, glossaries and supplementary gazetteer listings were enhanced to include lexicon extensions of spelling variations and synonyms. Such extensions are recognised as helpful in improving the accuracy of domain-specific text mining tasks (Thelwall & Buckley, 2013).

The Named Entity Recognition Pipeline

The NER pipeline focuses on the recognition of the CRM entities; E19.Physical_Object, E53.Place, E49.Time_Appellation and E57.Material, using hand-crafted JAPE rules and a range of domain oriented terminological resources. Additional NLP modules aim at improving the accuracy of the NER pipeline, such as the word-sense disambiguation module which addresses the issue of vocabulary polysemy and the negation detection module which filters out any mentions of CRM entities which are negated. Figure 4 presents the cascading order of the pipeline starting from the involvement of gazetteer listings in the process of NER, followed by semantic expansion via thesaurus relationships and ending with the bespoke NLP modules of validation, disambiguation, expansion, and negation detection of semantic annotations.

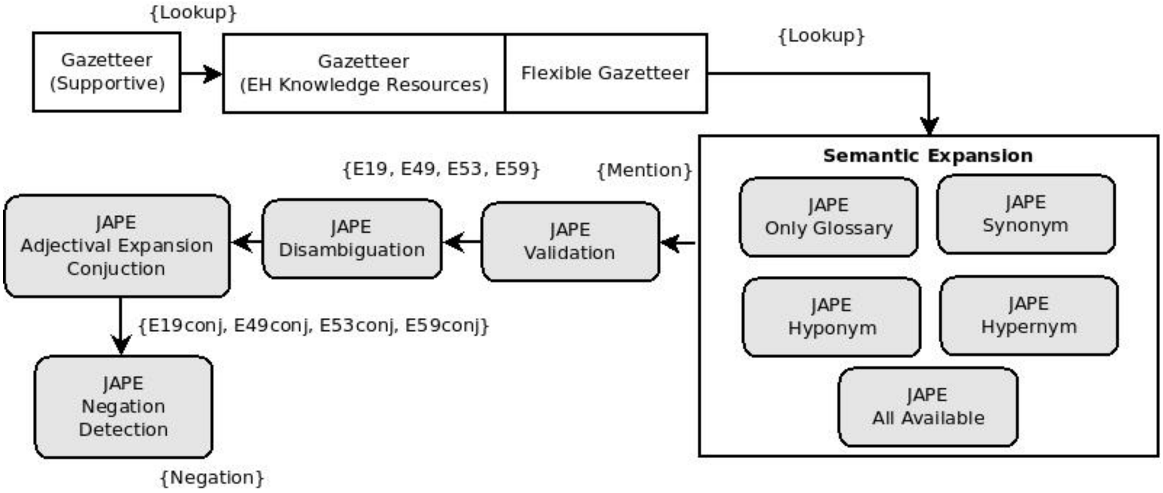


FIG. 4. The NER phase of the OPTIMA pipeline. Curly brackets show the annotation types produced at the different stages of the pipeline. White boxes are GATE (ANNIE) modules, grey are bespoke rules and modules.

Semantic Expansion for Information Extraction

The semantic expansion mechanism invoked by the NER pipeline is capable of a selective exploitation for IE purposes of the synonym, narrower and broader thesaurus relationships in the gazetteer listings (via the 'skosified' gazetteers). This flexibility is desirable due to the variable volume and specificity of the contributing terminological resources and important for controlling the volume of vocabulary terms contributing to the NER task. Some of the terminological resources, such as the glossaries, contain a limited number of highly relevant terms specialised to the archaeology domain. Other resources, such as the thesauri, contain a large number of general cultural heritage domain terms.

Relying only on glossary terms might benefit the Precision of the task but might harm its Recall. On the other hand, employing the vast range of available terms would improve Recall but potentially harm the system's Precision. Being able to control the volume of contributed terms via thesauri relationships can support tuning of the NER task towards Precision or Recall favouring configurations. The evaluation reveals the system's performance in a number of different configurations and discusses how Precision and Recall rates are affected by the level of semantic expansion involved in the NER task.

Expansion modes

There are three modes of semantic expansion, Synonym, Hyponym and Hypernym (ISO25964-1:2011) invoked by the NER phase of the pipeline. Two additional pipeline configurations exist that do not employ the semantic expansion mechanism; the Strict mode uses only the glossary terms and the All-available mode uses all available thesauri and glossary terms.

The Synonym expansion mode is a configuration that makes a modest use of the semantic expansion mechanism. The semantic expansion of the mode includes synonyms of the glossary terms, which are located in the thesauri structures. Based on the semantic alignment between ontological entities and terminology resources, the overlapping terms between glossary and thesaurus are assumed to have common word senses. For example the term *'grave'* (terminological reference: ehg003.35) originating from the glossary Simple Names for Deposits and Cuts shares the same sense as the term *'grave'* (terminological reference: 70080), originating from the Monument Type Thesaurus. Therefore, a glossary term can inherit the same semantic relationships of its equivalent (overlapped) thesaurus term

The Hyponym expansion mode exploits the narrower relationship in the contributing thesaurus structures. The mode builds on the overlapping terms between glossary and thesauri (from Synonym expansion) as entry points to the thesauri structures. Thus, the Hyponym mode exploits both synonym and narrower term relationships. For example, by expanding from the term *'grave',* the system will match its synonyms and narrower terms, such as 'pillow stone' and 'cremation grave', The Hypernym mode of semantic expansion enhances the Hyponym mode by extending matching to broader terms. For example *'funerary site'* is the broader term of *'grave'* in the Monument Types thesaurus. The Hypernym expansion matches all *'funerary site'* terms, such as *'animal burial pit, 'burial cairn', 'Tomb'* and their narrower terms.

NLP Modules for Improving NER

The work also investigated the effect of a range of modules that might improve the performance of NER in various contextual ambiguities: noun-validation, word-sense disambiguation, negation-detection and adjectival-expansion. Since the configuration of gazetteer matching is enabled at the level of word root (lemma), which allows matching of singular and plural forms but also verbs sharing the same word root, validation of matches using part-of-speech input is important for excluding verb matches from the NER task.

Word Sense Disambiguation (WSD) refers to the computational ability to identify the different meanings of a word that has multiple meanings (Navigli, 2009). The terminology overlap analysis revealed that glossaries aligned to the ontological classes Physical Object and Material contain a large number of overlapping terms. The WSD module attempts to resolve the ontological polysemy of such words and assign an appropriate ontological classification and terminological classification whenever possible. Contextual collocation rules, conjunction and other phrasal patterns were empirically selected for resolving ambiguity between Physical Object and Material entities. The disambiguation module resolves the appropriate terminological (SKOS) reference to ambiguous terms using the appropriate thesaurus. Whenever the ambiguity of terms cannot be resolved, annotation is assigned to both senses. This favours Recall rather than Precision (in light of the cross search use case), resulting in a half-correct annotation of ambiguous terms since only one of the two applied senses can be correct. On the other hand, it ensures that annotations are not discarded due to their ambiguity but are still revealed by the NER process.

The adjectival Expansion module is targeted at all entity types (Physical Object, Place, Material and Time Appellations). The stage utilises part of speech (POS) input and gazetteer listings aimed at expanding Lookup annotations over phrasal moderators which add meaning, for example *'burned'* pottery, *'broken'* pottery etc. The stage invokes rules that make use of part of speech input and simple patterns which deliver the enhanced and expanded annotation spans. However, no semantic vocabularies were applied for moderators. Moderators proved problematic and the issue is discussed further in the evaluation.

The pipeline also implements a negation detection mechanism adapting the technique of the NegEx algorithm (Chapman, Bridewell, Hanbury, Cooper, & Buchanan, 2001), which uses specialised vocabulary in combination with phrasal offset rules for the medical domain. The aim of the Negation Detection module is to filter out any annotations that are detected within negation phrases. The algorithm is modified to enable negation detection in the archaeology domain. The adaptation strategy considered issues relating to the size of negation window, applicability and enhancement of negation moderator glossaries, and characteristics of pseudo negation moderators affecting the scope of negation phrases (for details see Vlachidis & Tudhope, 2013).

Relation Extraction

The Relation Extraction (RE) phase of the OPTIMA pipeline is targeted at detecting phrases that connect (previously identified by the NER phase) CRM entity types in a meaningful way. The aim of this part of the investigation was to evaluate whether it was possible to annotate such phrases as CRM-EH event or property entities that could enable retrieval of meaningful entity relationships at an application level. The RE phase also specialises the CRM entities previously extracted by the NER phase to the corresponding more archaeologically specific CRM-EH subclasses.

Relation Extraction Rules

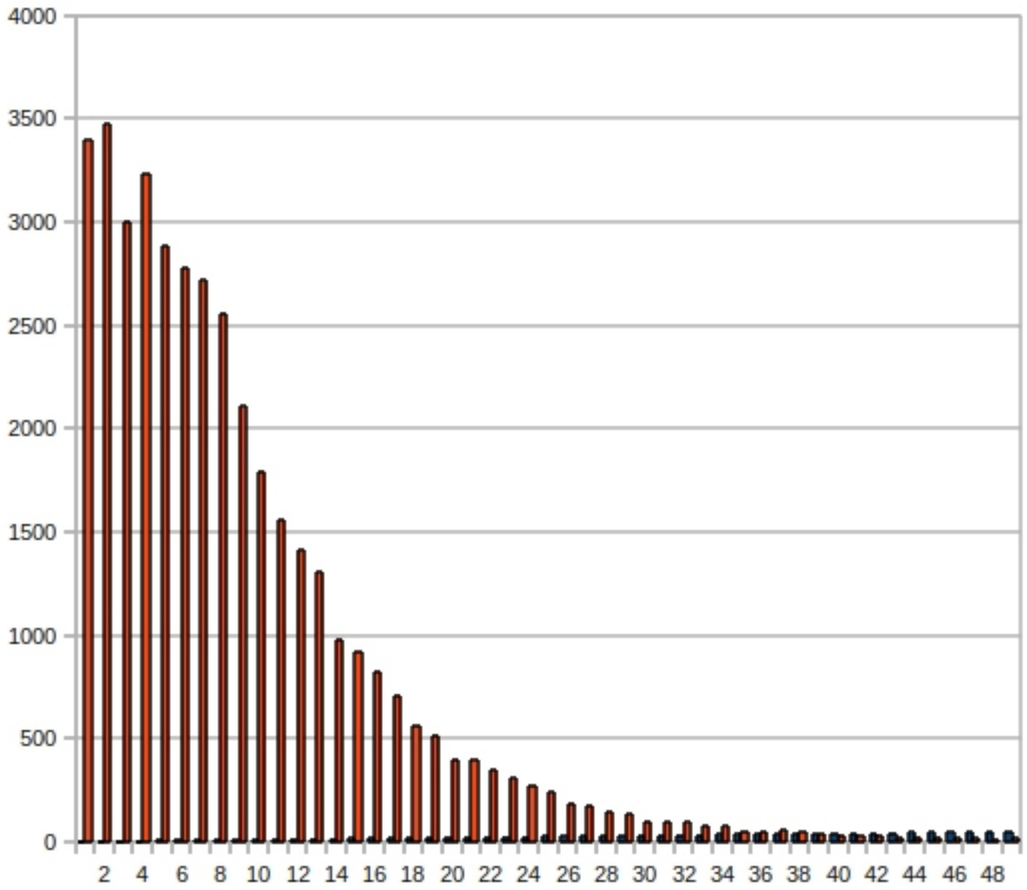
Definition of the hand-crafted relation extraction rules was informed by corpus analysis. This attempted to identify contextual patterns and other linguistic evidence which could be utilised by the RE rules. The first stage of the corpus analysis task developed a bespoke IE pipeline which extracted 146,008 text spans involving CRM entities. Each span was recorded in a CSV file containing the span string, its part of speech pattern and the number of tokens involved. For example, the span *'RB coin associated with hearth spot'* has the pattern NNP NN VBN IN NN NNP containing 6 Tokens. Similarly the span '*flint were recovered from the south end of the grave'* is reflected by the part of speech pattern 'NN VBD VBN IN DT NN NN IN DT NN' containing 10 tokens.

A subsequent intellectual analysis phase processed the statistics of the span patterns, in order to reveal commonly occurring pattern behaviours, which could be abstracted as JAPE grammars. The results of the occurrences were graphed to show the distribution of patterns with regards to the frequency of span length. The data, as discussed below, revealed a Zipf like distribution (Footnote 1), which formed the basis for determining the patterns selected and abstracted as JAPE grammars.

A projection of the frequency of extracted spans revealed the skewed graph (figure 5), where the vast majority of occurrences are observed between spans having 2 to 10 tokens. Given 20% of spans are responsible for 80% of occurrences, based on the Pareto principle, it was decided that a subsequent analysis of patterns should be focused on spans of maximum 10 tokens.

[1] Zipf's law (1935) states that, given a natural language corpus of sufficient volume, the frequency of any single word is inversely proportional to that word's rank associated with the frequencies.

FIG. 5. Frequency Distribution of Place - Time Appellation spans. Horizontal axis: span size in tokens. Vertical axis: count of spans. Spans of size up to ten tokens deliver the majority of matches.



The next stage analysed the most commonly occurring patterns containing up to 10 tokens. The number of occurrences of each unique pattern also approximates a Zipfian distribution similar to the trend of occurrences of span lengths described above. Thus, in the list of two tokens length, the most frequent pattern (NNP NN) occurs 342 times, delivering phrases such as, *'Roman settlement'* and *'Saxon cemetery'.* The second most frequent pattern (JJ NN) occurs 220 times delivering phrases such as *'prehistoric ditch'* and *'post-medieval deposit'.* The third most frequent occurs 182 times, the fourth 81, the fifth 55, the tenth 8 and so on. However, as we increase in span size, the step by which the number of occurrences declines is smaller, as well as the number of occurrences itself. For example the most frequent pattern of 10 token long phrases (i.e. NNS MD VB RB VBN TO DT JJ NN NN) occurs 27 times delivering phrases such as *'deposits can be soundly dated to the early medieval period'* and *'deposits can be directly related to the medieval field boundaries'.*

Again the Pareto principle was employed as a heuristic to yield a superset of patterns for intellectual analysis since many low-frequency examples were arbitrary spans containing two CRM entities which did not denote a CRM-EH event or a constitute valid phrase; 20% of the most frequent patterns of each span size up to ten tokens were analysed and used in the definition of JAPE rules. Frequency alone could not guarantee the validity of CRM-EH events and a subsequent intellectual analysis isolated a subset of patterns for JAPE grammar implementation.. The analysis grouped patterns under common structural characteristics that were then abstracted as JAPE rules of regular expressions supported by recursive Kleen and Logical operators. Overall, 43 JAPE rules were defined for the four distinct cases of relation extraction targeted by the pipeline

Translation from the selected linguistic patterns to hand-crafted IE rules was assisted by JAPE operators allowing complex expressions that matched a range of different linguist patterns. This technique allowed grouping of phrases that shared common pattern characteristics, which were then abstracted into JAPE grammars, for example the phrases *'Coin associated with hearth', 'The pit containing a group of flint', 'The ditch containing a single sherd of pot'* can all be matched by a single JAPE (Regular Expression) grammar as seen below

{E53}({!E53, Token.string != '.'})[0,4]{Lookup.majorType == E9_Verb}({!E53, Token.string != '.'})[0,4]{E19} (Footnote 2)

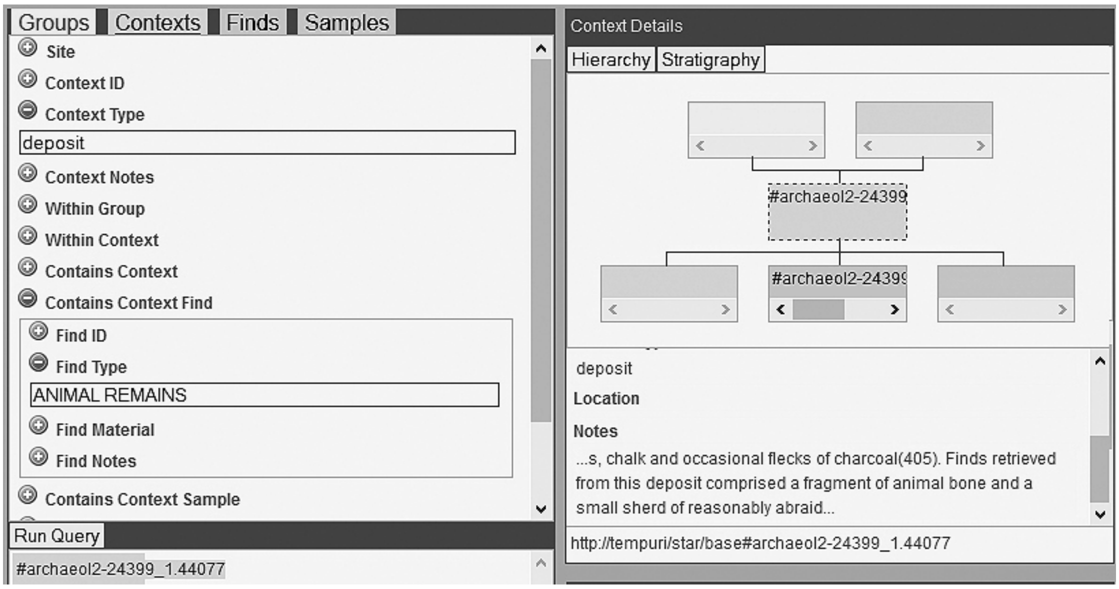
Similarly the grammars {E49}{E19} and {E19}({VG}|{Token.category == IN}){E49} match phrases, such as *'Roman coin', 'Medieval arrowhead'* and *'finds of Roman period',* respectively. The JAPE grammars that resulted from the formulation stage, were incorporated into the CRM-EH Relation Extraction pipeline.

Semantic Output

The output was expressed in standard interoperable format (RDF) and used in the STAR Demonstrator, enabling cross searching between grey literature documents and datasets via semantically defined user queries. The Demonstrator (Figure 6) enables users to build semantic searches from a subset of CRM-EH relationships relating to archaeological contexts and finds e.g. Archaeological Context of type *Deposit* containing finds of type *Animal Remain.*

[2] E53.Place (used for archaeological context), E19.Physical_Object (used for archaeological finds), E9_Verb domain oriented verb list (result of corpus analysis)

FIG. 6. The STAR semantic cross-search demonstrator. CRM-EH faceting and controlled vocabulary query and results for the relationship Context Type ‘Deposit’ containing Find of type ‘*Animal Remains*’ http://hypermedia.research.southwales.ac.uk/resources/star-demonstrator/



Site
Context ID
Context Type
deposit
Context Notes
Within Group
Within Context
Contains Context
Contains Context Find
Find ID
Find Type
ANIMAL REMAINS
Find Material
Find Notes
Contains Context Sample

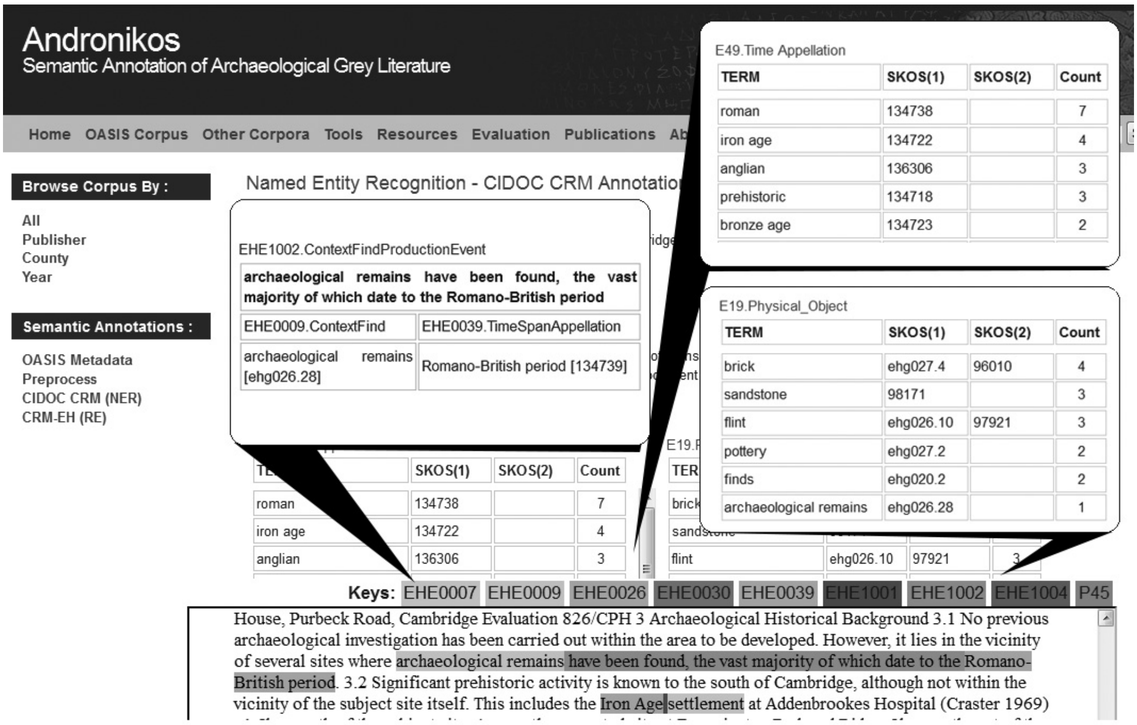
#archaeol2-24399_1.44077

deposit
Location
Notes
...s, chalk and occasional flecks of charcoal(405). Finds retrieved from this deposit comprised a fragment of animal bone and a small sherd of reasonably abraid...

http://tempuri/star/base#archaeol2-24399_1.44077

The indices are also made available in a web portal (Andronikos, 2012) supporting document inspection and browsing of both NER and RE results with respect to CRM semantics. Figure 7 presents the semantic inspection views of Andronikos. The portal offers tabular abstractions and contextual views of semantic annotations from a corpus of 2460 OASIS reports. The NER tabular abstractions hold SKOS references and frequencies of identified concepts - dual SKOS reference is assigned to overlapping glossary and thesauri terms. The RE abstractions hold the type of CRM-EH relationship and SKOS references of the relationship parts. The contextual view enables inspection of the extracted phrases and concepts in the context of the report with respective parts of relationships highlighted in key colours.

FIG. 7. The Andronikos web portal: Tabular and contextual view of semantic annotations of the document "Archaeological Evaluation: Purbeck House, Purbeck Road, Cambridge" http://www.andronikos.co.uk/Anno_CRM-EH.php?id=2395



Andronikos
Semantic Annotation of Archaeological Grey Literature

Named Entity Recognition - CIDOC CRM Annotatio

EHE1002.ContextFindProductionEvent

archaeological remains have been found, the vast majority of which date to the Romano-British period	
EHE0009.ContextFind	EHE0039.TimeSpanAppellation
archaeological remains [ehg026.28]	Romano-British period [134739]

E49.Time Appellation

TERM	SKOS(1)	SKOS(2)	Count
roman	134738		7
iron age	134722		4
anglian	136306		3
prehistoric	134718		3
bronze age	134723		2

E19.Physical_Object

TERM	SKOS(1)	SKOS(2)	Count
brick	ehg027.4	96010	4
sandstone	98171		3
flint	ehg026.10	97921	3
pottery	ehg027.2		2
finds	ehg020.2		2
archaeological remains	ehg026.28		1

TE	SKOS(1)	SKOS(2)	Count
roman	134738		7
iron age	134722		4
anglian	136306		3

Keys: EHE0007 EHE0009 EHE0026 EHE0030 EHE0039 EHE1001 EHE1002 EHE1004 P45

House, Purbeck Road, Cambridge Evaluation 826/CPH 3 Archaeological Historical Background 3.1 No previous archaeological investigation has been carried out within the area to be developed. However, it lies in the vicinity of several sites where archaeological remains have been found, the vast majority of which date to the Romano-British period. 3.2 Significant prehistoric activity is known to the south of Cambridge, although not within the vicinity of the subject site itself. This includes the Iron Age settlement at Addenbrookes Hospital (Craster 1969)

Evaluation

The evaluation phase assessed the NER and RE performance of the pipeline (OPTIMA) with respect to ontology (CIDOC CRM and CRM EH) driven semantic annotation. In addition, a set of dedicated evaluation tasks assessed the contribution of the separate NLP modules relating to the semantic expansion, syntactic pattern relation extraction, negation detection, word sense disambiguation and noun phrase validation.

The system performance was benchmarked via a 'Gold Standard' set of manual annotations, defined by archaeologists via an iterative process. For the intended cross-search use case, the Gold Standard aimed to represent the desirable result of semantic annotation of archaeological documents with respect to end-users of such documents (see below for more detail). Results are reported on the measurement of Precision and Recall and their weighted average F-measure, established as standard measurement units for measuring the performance of IE by the second Machine Understanding Conference, MUC 2 (Hobbs, 1993).

The above metrics examine a system's response in terms of correct or incorrect matches. This binary approach does not always provide enough flexibility to address partially correct answers, which are frequently delivered by semantic annotations. Thus partial matches are also incorporated. In this case, the Precision and Recall formulas can be defined as

$$Recall = \frac{Ncorrect + \frac{1}{2}Partial\ Matches}{Nkey}, \qquad Precision = \frac{Ncorrect + \frac{1}{2}Partial\ Matches}{Ncorrect + Nincorrect + Partial\ Matches}$$

The value of the weight can reflect the importance attached to partial matches. When partial matches are treated as correct matches the assigned weight is set to 1 and the approach is described as *Lenient*. *Strict* is the case when partial matches are not taken into account (weight is 0), while *Average* is the case where partial matches weight is set to 0.5 as above.

Evaluation Method

The evaluation method was based on an iterative process of Gold Standard definition. The first phase of the definition involved a pilot evaluation study, testing and improving the clarity of the semantic annotation instructions. The pilot evaluation used a small corpus of 10 summary extracts, of archaeological excavation and evaluation reports manually annotated by three volunteer archaeologists. Typically, annotation agreement scores of archaeological text is moderate-low as a result of the many embedded language ambiguities of the domain (Byrne, 2007; Zhang et al., 2010). Ensuring the clarity and unambiguity of annotation instructions can benefit the validity and usefulness of the Gold Standard definition, which in turn directly influences the accuracy of evaluation results.

The pilot evaluation phase revealed several instruction issues relating to the inclusion of entity moderators, the scope of archaeological context annotation and the length of entity relations annotation. The revised instructions defined both smaller (eg *'cut')* and larger (eg *'post-hole structure')* archaeological context groupings as within scope. They restricted the annotation spans of relations to be within the boundaries of a sentence between entities. They suggested that only the most immediate moderator of an annotation be included (eg '*burnt* flint'). The amendments resulted in a significant improvement of overall manual annotation correctness, increasing Precision from 62% to 84% while the Recall score remained unchanged (71%).

The main manual annotation task was conducted at the Archaeology Data Service (ADS, York University), with the voluntary participation of 12 archaeologists, including ADS staff and post-graduate students. Annotators were instructed to annotate at the level of archaeological concepts rather than attempting to identify abstract ontological entities in context. The instructions in effect directed annotators to adopt the principles of orthography, topicality, phrasal annotation and negation detection, following the practice of the SEKT project (Peters, Aswani, Bontcheva, & Cunningham, 2005). In detail, the instructions directed the task of manual annotation at the concepts of archaeological place, archaeological find, the material of archaeological finds and time appellation. The annotators were directed to annotate textual instances relevant to the STAR project's archaeological (cross search) research questions and to identify phrases containing two or more of the targeted concepts in meaningful relations. Such phrases were used in the evaluation of the CRM EH relation extraction phase.

The evaluation methodology followed a user-oriented perspective, where annotators were expected to exercise judgement as competent users. With the ecological validity of the results in mind, the instructions for evaluators were intended to be relevant to future cross search and hence neither the scope of the ontology elements nor the precise vocabulary were specified exactly. This approach differs from some more specific forms of evaluation deriving from the ML tradition, where the annotation criteria are spelled out in detail or the vocabulary is provided .

Each annotator was assigned a document containing 2500-3000 words. Overall, six composite documents containing 55 new summary passages not previously used in the pilot stage were annotated by six groups, where each group consisted of two annotators. The evaluation summaries originate from archaeological evaluation and post-excavation reports of the ADS (OASIS) repository of archaeological grey literature (Hardman & Richards, 2003). The reports chosen were representative of the OASIS corpus with respect to the contributing archaeological units. Summaries were selected for the Gold Standard since they reflect a report's main findings and support the end-user focus of the evaluation due to their density and richness. In addition, manual annotation of summary sections is significantly less labour intensive than manual annotation of the complete reports, which can be very extensive. Summaries are relatively straightforward to isolate and are common to all reports. It was considered important that the Gold Standard take account of different types of reports originating from different archaeological units rather than focusing only on a few complete reports (footnote 3). The overall length of the 55 summaries used by the main evaluation phases is 11306 words.

The final stage in the definition of the Gold Standard delivered an explicit and unambiguous Gold Standard via a reconciliation process, which resolved any disagreement between the different manual annotation sets of each document. This was done by assigning the role of Super Annotator (Savary, Waszczuk, & Przepiórkowski, 2010) to a senior archaeologist research collaborator, who acted as a referee between individual annotation sets and reviewed the cases of annotation disagreement. The Gold Standard contained overall 1215 annotations: 120 Physical Object, 309 Time Appellations, 511 Place, 61 Material, and 214 Relation Phrases (128 Context Event, 42 Production Event, 21 Deposition Event and 23 Consists of Material).

3 In addition to the evaluation, results are available on the complete reports from the Andronikos web portal http://www.andronikos.co.uk

Evaluation Results

Discussion of the evaluation results is divided into two main sections (NER and RE). The discussion concludes with the summative evaluation results and the findings of the semantic annotation work.

The first phase benchmarked the NER outcome of the pipeline against five different system configurations, corresponding to the five modes of semantic expansion (Table 5.1). *Average* mode results are about 6% lower than the *Lenient* mode. The decision to include entity moderators in the Gold Standard definition affected the overall precision of the system due to the subjective and inconsistent use of moderators by manual annotators, which resulted in discrepancies at annotation boundaries. Annotating moderators proved too difficult even with revised instructions and yielded little semantic benefit, since the focus was on CRM entity recognition, rather than moderators which were not semantically modelled and cannot be fully utilised on an application level, apart from a flat textual enhancement of entity matches. Thus, the Lenient mode (which treats partially matches as full matches) provides a fairer view of the system's performance with regards to CRM entity matches.

Based on *Lenient* mode and F-measure score, the best performing semantic expansion mode is *Hypernym* (82%). However, Table 1 shows that the *Hypernym* mode does not provide the best Precision score, which is delivered by the *Hyponym* and the *Synonym* modes (both 80%). *Hypernym* delivers best Recall (87%), very close to the *All-Available* mode's Recall (86%). On the other hand, *All-Available* delivers the lowest Precision (71%), which is expected since the mode uses all the available terms from thesauri and glossaries, including those which do not relate strongly with the targeted entities.

TABLE 1. Precision, Recall and F-measure results for the 5 Semantic expansion modes

	Recall		**Precision**		**F-measure**	
	Average	*Lenient*	*Average*	*Lenient*	*Average*	*Lenient*
Only-Glossary	0.60	0.65	0.71	0.78	0.64	0.70
Synonym	0.66	0.72	0.73	0.80	0.70	0.76
Hyponym	0.70	0.76	0.73	0.80	0.71	0.78
Hypernym	0.80	0.87	0.71	0.78	0.75	0.82
All-Available	0.79	0.86	0.65	0.71	0.70	0.78

The five different modes of semantic expansion represent different but to some degree overlapping exploitations of glossaries and thesauri. From the smallest (Only-Glossaries) to the largest volume (All-Available), the contribution includes more and more terms from synonyms to narrower and to broader concepts. The results clearly validate expanding the immediate glossaries with broader domain thesaurus resources. Results also clearly support the selective use of thesaurus expansion compared to All-Available terms where Precision is concerned. Differences between the modes are fine grained but suggest that use of semantic expansion in IE may generally follow the search literature on query expansion via thesaurus relationships (reviewed in Tudhope Binding, Blocks, and Cunliffe 2006). Based on these results, the Hypernym mode delivers the best F-measure. Although, it does not provide the best Precision, it could be regarded as a good choice for supporting an Information Retrieval task focused on Recall rather than Precision. On the other hand, the Hyponym mode delivers better Precision and could be employed where Precision is more important than Recall.

Regarding the different ontology entities (Table 2), the system performs best (Precision 97%, Recall 99%, F-Measure 98%) for the Time Appellation entity type (E49). The good performance of the system is due to the completeness and non-ambiguity of the terms in the Timeline thesaurus, combined with enhancement of lexical variations (*Earlier, Early, Mid, Middle, Mid-late* etc.) during the 'skosifiation' processes.

Depending on the mode of Semantic expansion (best results for Hypernym), the system delivers F-measure scores: for Physical Object (E19) between 63% and 81%; for Place (E53) between 69% and 85%; for Material (E57) between 50% and 63%. The slightly better performance for Place compared to Physical Object is probably due to the clarity of the Place related glossary resources, which did not suffer the same level of overlap as the Object related glossary resources. On the other hand, NER of Material is supported by terminological resources that contain a large amount of overlapping concepts between glossaries and thesauri aligned both to Physical Object and Material entities. This overlap has influenced Material performance, as evident from the low Precision. The Material entity is influenced by ambiguities particular to the archaeology domain. For example the same concept *('iron', 'pottery',* etc.) can be treated by archaeologists as a Find (i.e. Physical Object) or as the Material of an object. In some cases, an Object is implicit (e.g. *a fragment of* pottery).

TABLE 2. Recall and Precision scores of four CRM entities (E19.Physical Object, E49.Time Appellation, E53.Place and E57.Material) for the five modes of semantic expansion.

	Physical Object		**Time Appellation**		**Place**		**Material**	
	Recall	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*
Only-Glossary	0.53	0.78	0.99	0.97	0.57	0.88	0.50	0.51
Synonym	0.69	0.84	0.99	0.97	0.68	0.87	0.52	0.53
Hyponym	0.72	0.83	0.99	0.97	0.79	0.86	0.55	0.53
Hypernym	0.80	0.81	0.99	0.97	0.91	0.80	0.77	0.54
All-Available	0.84	0.64	0.99	0.97	0.92	0.76	0.68	0.49

Contribution of Entity Validation Modules

A secondary aspect of the evaluation assessed the contribution of various NLP techniques to the NER phase in five system configurations, executed in *Hypernym* semantic expansion mode. The IE system was stripped of all NLP modules used by the NER pipeline (*Noun Phrase Validation*, *Negation Detection*, *Word Sense Disambiguation)*. Some additional concepts, added to the matching mechanism after the pilot evaluation, were also removed. A basic configuration (*Basic)* was used as an indicator of the system performance, without the use of accuracy techniques. The contribution of individual NLP module was then evaluated by adding each individual module to the *Basic* configuration and all three combined Table 3.

TABLE 3. Evaluation metrics of contribution to NER of the bespoke NLP modules; Negation Detection, Noun Phrase Validation and Word-Sense Disambiguation

	Recall	**Precision**	**F-Measure**
Basic	0.89	0.55	0.67
Negation Detection	0.89	0.57	0.68
NP Validation	0.88	0.62	0.72
WS Disambiguation	0.87	0.61	0.71
All modules	0.87	0.78	0.82

Overall, Precision improves slightly when adding each individual NLP validation module. For example, Negation Detection improves Precision by 2% without harming Recall, although the impact is probably affected by the limited number of negation phrases included in the evaluation corpus (Footnote 4). The key point is the significant improvement in Precision of all the NLP validation modules combined together (increasing from 55% basic configuration to 78%). This slightly reduces Recall from 89% to 87% but the overall F-measure score improves from 67% to 82%.

Relation Extraction Evaluation Results

The second phase of evaluation benchmarked the system with regards to Relation Extraction (RE), addressing system performance in the identification of phrases that relate entities to CRM-EH ontology event or property descriptions. The more complex Syntactic-based approach was compared with a basic Offset-based configuration to see if the additional effort of the hand-crafted relation extraction rules yielded benefits. The Offset-based system used basic rules in the form of *<entity><up to 5 tokens><Verb><up to 5 tokens><entity>* whereas the Syntactic-based system employed syntactical pattern rules. Table 4 compares the performance of the two different system configurations in terms of Recall, Precision and F-measure both in *Average* and *Lenient* mode of reporting.

4 From the 1099 NER entities delivered by the system, running on the Hypernym expansion mode, only 33 were negated phrases

TABLE 4. Precision, Recall and F-measure of relation extraction (CRM-EH event types) between the Offset-based and Bottom-up system configurations.

	Recall		**Precision**		**F-measure**	
	Average	*Lenient*	*Average*	*Lenient*	*Average*	*Lenient*
Offset-based	0.67	0.83	0.52	0.64	0.57	0.70
Syntactic-based	0.67	0.75	0.76	0.86	0.70	0.80

Results show that the Syntactic-based configuration delivers higher F-measure and Precision scores, while the Offset-based system delivers better Recall results on the *Lenient* mode of reporting. The Offset-based configuration delivers 8% higher Recall results, while the Syntactic-based configuration delivers 22% higher Precision results. Based on the F-measure score, the Syntactic-based configuration outperforms the Offset-based system by 10% on the *Lenient* mode and by 13% on the *Average* mode. The significant improvement in the Precision, in combination with the constrained drop in Recall, gives a considerable advantage to the Syntactic-based configuration.

At a basic comparative level, the overall NER F-measure results (82%) are competitive with full scale semantic annotation systems targeted at archaeological context that have yielded F-measure scores ranging from 68% to 83% (Zhang *et al*. 2010) and full scale systems targeted at historical text that have delivered F-measure scores of 73% (Grover *et al*. 2008). In terms of RE the overall F-Measure results (80%) compare favourably with ML approaches targeted at extracting relations from archaeological text (Byrne & Ewan, 2010) and rule-based, ontology guided systems targeted at biomedicine text, which deliver F-measure scores between 64% to 76% (Cimiano *et al*. 2005). As discussed above, there is a high level of ambiguity in archaeology domain use of Material entity terms and the distinction between Material and Object may not be of significance for many archaeological applications. If Material is excluded from the evaluation then *Lenient* mode results of NER increase to 90% Recall, 88% Precision and 89% F-measure.

However, such comparisons are broad brush; details of evaluation methodologies may differ or not be available. Additionally, the method, scope and purpose of Relation Extraction differs significantly between projects. Byrne & Ewan (2010), for example, focuses on the identification of verbs, which act as nodes for relating entities in terms of *hasLocation*, *hasPeriod*, *partOf* relations etc., rather than complete phrases which can be modelled as CRM-EH events for purposes of semantic cross search over reports and datasets. On the other hand, Cimiano et al. (2005) use deep parsing for identifying biochemical events such as control/regulation and biochemical interaction with emphasis on discourse analysis driven by classification of domain specific verbs and a taxonomy of biochemical events.

Conclusion

The semantic indexing results demonstrate the capacity of rule-based Information Extraction techniques to deliver interoperable semantic abstractions (semantic annotations) with respect to the domain ontologies. Major contributions of the semantic indexing effort, include recognition of CIDOC-CRM entities using shallow parsing NLP techniques driven by a complimentary use of ontological and terminological domain resources and employment of context-driven information extraction rules for the recognition of semantic relationships from phrases of unstructured text.

Results of the Gold Standard evaluation are at least competitive with related work, although as discussed direct comparisons of performance measures can be misleading due to application domain features and individual system characteristics. The evaluation shows clear benefit in the use of assistive NLP modules relating to word-sense disambiguation, negation detection and noun phrase validation, which improve the precision performance of the NER outcome. The results also demonstrate the capacity and agility of the controlled thesaurus expansion to enable a configurable NER behaviour in favour of either Precision or Recall. In addition, RE performance clearly benefits from a syntactic based definition of relation extraction patterns derived from domain oriented corpus analysis as opposed to a mechanistic definition of offset span patterns.

With regards to the task of RE, the development adopted a novel approach in the application of the Zipfian distribution principle for the selection and definition of a manageable set of relation extraction rules originating from a large volume of syntactical patterns delivered by corpus analysis techniques. This has permitted the extraction of meaningful combinations as patterns of CIDOC CRM ontology classes associated with SKOS concepts. This has the potential to significantly reduce the 'false drop' problem in digital archaeology portals discussed in the Introduction, where individual entities are combined inappropriately in retrieval response.

The RE semantic output has been directly applied in the STAR Demonstrator (Tudhope et al. 2011). The information extraction results have been expressed in the same integrating semantic format as the archaeological datasets extracted by different means. This has enabled precise semantic search across diverse archaeological datasets and also grey literature over patterns such as *'hearth containing a coin', 'deposit associated with animal remains',* etc. The evaluation focused upon report summaries, which can be considered representative of the key findings of an excavation. Other use cases include an NER overview of a complete document, where the frequency counts of the different entities provide a statistical view of the various contexts, finds and periods mentioned in the report.

Future work

Future steps towards improving the current results could include methods of advancing contextual analysis to further the development of entity validation modules. It would be interesting to investigate whether the negation detection and word-sense disambiguation modules could benefit from deep parsing techniques delivering richer contextual evidence of the extracted elements. A deeper contextual analysis might produce a finer negation detection result avoiding blanket exclusion from indexing of all entities involved in a negation phrase and also provide further evidence and input to word-sense disambiguation grammars. In addition, if it were possible to (manually) remove the false positive cases semantic annotation output of the RE pipeline then it might be possible to form a training set for a supervised ML pipeline, which could potentially improve the generalizability and performance of the system.

The methodology has been expanded beyond the immediate archaeological domain in a cultural heritage pilot study of information extraction from catalogue descriptions of classical vases originating from collection fascicules. The CASIE (Classical Art Semantics Information Extraction) a collaborative project between the Hypermedia Research Group (University of South Wales) and the Beazley Archive (Oxford University) automatically extracted information about cultural objects from classical art scholarly texts (high quality structured catalogues 'fascicules') and represented it in terms of the CIDOC-CRM. Generalisation of the method to a multilingual archaeological context is currently in progress, as part of the ARIADNE European archaeological e-infrastructure project.

References

Ananiadou, S., Pysysalo, S., Tsujii, J., & Kell, D. (2010). Event extraction for systems biology by text mining the literature. Trends in Biotechnology, 28(7), 381-90.

Andronikos Web Portal. (2012). Semantic Annotation of Archaeological Grey Literature. Retrieved September 1, 2014, from http://www.andronikos.co.uk

Bates M. (1986), Subject access in online catalogs: a design model, Journal of the American Society for Information Science, 37(6), pp. 357-376.

Bontcheva, K., Tablan, V., Maynard, D., & Cunningham, H. 2004. Evolving GATE to meet new challenges in language engineering. Natural Language Engineering, 10(3/4), 349-373.

Bontcheva, K., Duke, T., Glover, N., & Kings, I. 2006. Semantic Information Access. In J. Davies, R. Studer, & P. Warren, P. (Eds.), Semantic Web Semantic Web Technology: Trends and Research in Ontology Based Systems. Chichester: John Wiley and Sons Ltd.

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

Byrne, K., & Ewan, K. 2010. Automatic extraction of archaeological events from text. In B. Frischer, J. Crawford, & D. Koller, (Eds.), Making History Interactive: Computer Applications and Quantitative Methods in Archaeology. Oxford: Archaeopress.

Chapman,W.W., Bridewell, W., Hanbury, P., Cooper, G.F., & Buchanan, B.G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics, 34(5), 301–310.

Cimiano, P., Reyle, U., & Saric, J. (2005). Ontology-driven discourse analysis for information extraction. Data and Knowledge Engineering, 55 (1), 59–83.

Cowie, J., & Lehnert, W. (1996.) Information extraction. Communications ACM, 39(1), pp. 80–91

Crofts, N., Doerr, M., Gill, T., Stead, S., & Stiff, M. (2009). Definition of the CIDOC Conceptual Reference Model, *FORTH Greece*. Retrieved September 1, 2014, from http://cidoc.ics.forth.gr/docs/cidoc_crm_version_5.0.1_Mar09.pdf

Cunningham, H., Maynard, D., & Tablan, V. (2000). JAPE a Java Annotation Patterns Engine (Second Edition). Technical report CS--00--10, University of Sheffield, Department of Computer Science. Retrieved September 1, 2014, from http://www.dcs.shef.ac.uk/intranet/research/resmes/CS0010.pdf

Cunningham, H., Maynard, D., Bontcheva, K., & 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 (ACL2002). Stroudsburg, Philadelphia

English Heritage Linked Data Vocabularies for Cultural Heritage (2014). Retrieved September 1, 2014, from http://www.heritagedata.org/blog/vocabularies-provided/

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. Retrieved September 1, 2014, from http://intarch.ac.uk/journal/issue17/falkingham_index.html

Friedman, C., Kra, P., Yu, H., Krauthammer, M., & Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles, Bioinformatics 17, 74–82

Golub, K. (2006). Automatic subject classification of textual Web documents. Journal of Documentation, 62(3), 350-371.

Golub K, Lykke M, Tudhope D. (2014). Enhancing social tagging with automated keywords from the Dewey Decimal Classification. Journal of Documentation 70(5), 801-828. Emerald.

Golub K, Tudhope D, Zeng M, Žumer M. (2014). Terminology Registries for Knowledge Organization Systems – Functionality, Use, and Attributes. Journal of the Association for Information Science and Technology, 65(9), 1901-1916. Wiley.

Grishman, R., & Sundheim, B. (1996). Message Understanding Conference-6: a brief history. Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996). Copenhagen.

Grover, C., Givon, S., Tobin, R., & Ball J. (2008). Named entity recognition for digitised historical texts. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008). Marrakech.

Hardman, C., & Richards, J.D. (2003). OASIS: Dealing with the digital revolution. In M. Doerr, & A. Sarris, (Eds), The Digital Heritage of Archaeology: Computer Applications and Quantitative Methods in Archaeology (pp. 325–329). Heraklion: ICS Publications.

Hobbs, J.R. (1993). The Generic Information Extraction System. Proceedings of the *5th Message Understanding Conference* (MUC-5). Baltimore

Isaac, A., & Summers, E. (2009). SKOS Simple Knowledge Organization System Primer. Retrieved September 1, 2014, from: http://www.w3.org/TR/skos-primer

ISO 25964-1 2013. ISO 25964. Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval. Retrieved September 1, 2014, from http://www.niso.org/schemas/iso25964/

ISO 25964-2 2013. ISO 25964. Thesauri and interoperability with other vocabularies. Part 2: Interoperability with other vocabularies. Retrieved September 1, 2014, from 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 Special Theme Issues of the Philosophical Transactions of the Royal Society A, Crossing Boundaries: Computational Science, E-Science and Global E-Infrastructures, 2507–2519

Leroy, G., & Chen, H. (2005). Genescene: An ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical text. Journal of the American Society for Information Science and Technology, 56(5), 457-468

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3–26

Navigli, R. 2009. Word sense disambiguation: a survey. ACM Computing Surveys 41(2), 10–11

OPTIMA (2012) Project Resources, Retrieved September 1, 2014, from: http://sourceforge.net/projects/optimacidoc/

Ore, C-E., & Eide, Ø. (2009). TEI and cultural heritage ontologies: Exchange of information. Literary and Linguist Computing, 24 (2), 161-172.

Peters,W., Aswani, N., Bontcheva, K., & Cunningham, H. (2005). Quantitative evaluation tools and corpora v1. Technical report, SEKT project deliverable D2.5.1

Richards, J., & Hardman, C. (2008). Stepping back from the trench edge. In M. Greengrass, & L. Hughes, (Eds.) The Virtual Representation of the Past. Farnham England: Ashgate.

Richards, J., Jeffrey, S., Waller, S., Ciravegna, F., Chapman, S., & Zhang, Z. (2011). Archaeology data services and the Archaeotools project: Faceted classification and natural language processing. In S. Whitcher Kansa, E.C. Kansa & E.Watrall (Eds.) Archaeology 2.0. New Approaches to Communication & Collaboration (pp 31-56). Los Angeles: Cotsen Institute of Archaeology Press

Savary, A., Waszczuk, J., & Przepiórkowski, A. (2010). Towards the annotation of named entities in the national corpus of Polish. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'10). Valletta.

Thelwall, M., & Buckley, K. (2013). Topic-Based sentiment analysis for the Social Web: The role of mood and issue-related words. Journal of the American Society for Information Science and Technology, 64(8), 1608-1617

Tudhope, D., Binding, C., Blocks, D., & Cunliffe, D. (2006). Query expansion via conceptual distance in thesaurus indexed collections. Journal of Documentation, 62 (4), 509-533.

Tudhope, D., May, K., Binding, C., & Vlachidis A. (2011). Connecting archaeological data and grey literature via semantic cross search. Internet Archaeology, (30). Retrieved September 1, 2014, from http://intarch.ac.uk/journal/issue30/tudhope_index.html/

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., & Ciravegna, F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 4(1), 14–28.

US NIST (2003). The ACE 2003 evaluation plan. US National Institute for Standards and Technology (NIST). Retrieved September 1, 2014, from http://www.itl.nist.gov/iad/mig/tests/ace/2003/ (Accessed 12 June 2012).

Vlachidis, A. & Tudhope, D. (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.

Vlachidis A, & Tudhope D. (2013). The semantics of negation detection in archaeological grey literature. In Garoufallou E. & Greenberg J. (Eds), Metadata and Semantics Research. Communications in Computer and Information Science, 390, 188-200.

Zelenko, D., Aone, C., & Richardella, A. (2003). Kernel methods for relation extraction. Journal of Machine Learning Research, (3), 1083–1106.

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, 301–315

Zipf, G.K. (1935). The Psycho-biology of language: An introduction to dynamic biology, second edition (1965), Cambridge: MIT Press.

Zeng M, Chan, L. (2004), Trends and issues in establishing interoperability among knowledge organization systems. Journal American Society of Information Science, 55, 377–395.