Semantic Indexing via Knowledge Organization Systems: Applying the CIDOC-CRM to Archaeological Grey Literature

Andreas Vlachidis

A thesis submitted in partial fulfilment of the requirements
of the University of Glamorgan /
Prifysgol Morgannwg
for the degree of a Doctor of Philosophy.

July 2012

University of Glamorgan
Faculty of Advanced Technology

Certificate of Research

This is to certify that, except where specific reference is made, the work presented in this thesis is the result of the investigation undertaken by the candidate.

Candidate: .....

Director of Studies: .....

Declaration

This is to certify that neither this thesis nor any part of it has been presented or is being currently submitted in candidature for any other degree other than the degree of Doctor of Philosophy of the University of Glamorgan.

Candidate: .....

Abstract

The volume of archaeological reports being produced since the introduction of PG16[1] has significantly increased, as a result of the increased volume of archaeological investigations conducted by academic and commercial archaeology. It is highly desirable to be able to search effectively within and across such reports in order to find information that promotes quality research. A potential dissemination of information via semantic technologies offers the opportunity to improve archaeological practice, not only by enabling access to information but also by changing how information is structured and the way research is conducted.

This thesis presents a method for automatic semantic indexing of archaeological grey-literature reports using rule-based Information Extraction techniques in combination with domain-specific ontological and terminological resources. This semantic annotation of contextual abstractions from archaeological grey-literature is driven by Natural Language Processing (NLP) techniques which are used to identify "rich" meaningful pieces of text, thus overcoming barriers in document indexing and retrieval imposed by the use of natural language. The semantic annotation system (OPTIMA) performs the NLP 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 ISO Standard (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.

The results demonstrate that the techniques can deliver semantic annotations of archaeological grey literature documents with respect to the domain conceptual models. Such semantic annotations have proven capable of supporting semantic query, document study and cross-searching via web based applications. The research outcomes have provided semantic annotations for the Semantic Technologies for Archaeological Resources (STAR) project, which explored the potential of semantic technologies in the integration of archaeological digital resources. The thesis represents the first discussion on the employment of CIDOC CRM and CRM-EH in semantic annotation of grey-literature documents using rule-based Information Extraction techniques driven by a supplementary exploitation of domain-specific ontological and terminological resources. It is anticipated that the methods can be generalised in the future to the broader field of Digital Humanities.

[1] The Department of the Environment 1990 Planning Policy Guidance Note 16 (PPG16 (DoE 2010)

Acknowledgements

I would like to wholeheartedly thank my supervisor Professor Douglas Tudhope for his true kindness, invaluable support and inspiring guidance. Thanks are also due to the members of the Hypermedia Research Unit, Dr. Daniel Cunliffe, Ceri Binding and Dr. Renato Souza for their feedback and encouragement which helped me pursue and complete this study.

I would like also to thank

- Keith May (English Heritage) for his expert input that has greatly helped to understand issues relating to archaeology practice and for his quality feedback and work as the Super Annotator of the evaluation process.
- Phil Carlisle (English Heritage) for providing domain thesauri
- Paul Cripps and Tom Brughmans for their manual annotation input during the pilot evaluation phase
- The Archaeology Data Service for provision of the OASIS corpus. In particular I would like to thank Professor Julian Richards, Dr. Stuart Jeffrey and all the ADS staff members and University of York postgraduate archaeology students who provided manual annotations for the main evaluation phase.
- The external examiners of the thesis, Prof. Anders Ardö (Lund University, Sweden) and Dr.Antony Beck (University of Leeds, UK) for their kind, valuable and constructive feedback that has helped to improve parts of this work.

Finally, I would like to thank all my family members for their love and unwavering support not only during the period of this study but all my life so far but most specially I would like to thank my companion and partner in life Ms Paraskevi Vougessi for her unconditional love and patience that helped to make this work possible.

Γειά σου Μάρκο !

To the music of :

Markos Vamvakaris, 1905 – 1972

Manolis Rasoulis, 1945 – 2011

Nikos Papazoglou, 1948 – 2011

To my father, Spyros Vlachidis, 1946 – 1986

Table of Contents

List of Figures

List of Tables

Published Work

This research work of this thesis has resulted in a number of published research papers of which the most prominent papers are shown below.

- **Vlachidis, A. and Tudhope, D. (2012)** 'A Pilot Investigation of Information Extraction in the Semantic Annotation of Archaeological Reports', *International Journal of Metadata, Semantics and Ontologies* (IJMSO), 7 (3), 222-235.

- **Vlachidis. A., Binding., C., May, K. and Tudhope, D. (2011)** 'Automatic Metadata Generation in an Archaeological Digital Library: Semantic Annotation of Grey Literature', *In Proceedings CLA'11 Computational Linguistic Applications*, Jachranka, Poland, 17-19 October.

- **Vlachidis, A. and Tudhope, D. (2011)** 'Semantic Annotation for Indexing Archaeological Context: A Prototype Development and Evaluation. Metadata and Semantic Research, *In Proceedings MTSR '11*, Ismir, Turkey, 12-14 October, *Communications in Computer and Information Science*, 240, pp. 363-375

- **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*, 62 (4&5), 466 – 475.

Chapter 1

Introduction to Thesis

1.1 Prelude

The Oracle at Delphi was a paramount institution of ancient Greece. It had a strong religious and political influence and at the same time was a neutral place for storing the treasures of the ancient world. The key to its success was the elaborate use of natural language, known until today as oracular obscurity. The institution delivered oracles to pilgrims upon request, expressed in such language that could always be interpreted as true regardless of the result.

Some of the most well-known oracles that demonstrated this oracular obscurity are known as the "wooden walls" of Athens and the "great empire" of Croesus. When Athenians requested the Oracle to "advise" how to defend their city from the invading Persian army, the Oracle replied that the "wooden walls" will protect Athens. Although, Athens was protected by wood walls at the time, these did not stop the Persian army from capturing the city. It was though the battle of Salamis at sea where the Athenians defeated the Persian fleet and so the "wooden walls" were interpreted as being the Athenian fleet. Croesus, King of Lydia, before invading the Persian Empire asked the Oracle about the results of the war, who replied that upon the end of the war a "great empire", will be destroyed. Croesus encouraged by the oracle invaded Persia only to lose the war. His army was defeated and it was his empire that was destroyed.

In today's world, and in particular in modern web computing, being able to interpret the meaning of language is very desirable. Sir Tim Berners-Lee proposed the Semantic Web based on a view of a Web in which computers:

> *…become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A 'Semantic Web', which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The 'intelligent agents' people have touted for ages will finally materialize*

The Semantic Web is proposed to add logic to the Web for improving user experience and information seeking activities and so to use rules, to make inferences and to choose

courses of action that are defined by the meaning of information. It is said that the Semantic Web, when properly designed "can assist the evolution of human knowledge as a whole" (Berners-Lee et al. 2001)

1.2 Context and Motivation

Since 1990 there has been a significant increase in the number of archaeological projects being carried out in England and Wales which has directly affected the volume of archaeological reports being produced. This can be related directly to the introduction of the Department of the Environment 1990 Planning Policy Guidance Note 16 (PPG16) on *Archaeology and Planning*, followed by the 1994 (PPG15) on *Planning and the Historic Environment* (DoE 2010)[2]. As a result, a large number of archaeological investigations (28,000 between 1990-2000) have been undertaken delivering a similarly large volume of fieldwork reports (Falkingham 2005). Such fieldwork reports are often called "grey literature". The term, as agreed at the 3rd International Conference of Grey Literature 1997 in Luxembourg, refers to literature that "*is produced by all levels of government, academics, business and industry, in print and electronic formats, but which is not controlled by commercial publishers*" (Farace 1997). Being created and distributed to disseminate knowledge rather than to sell for profit, grey literature is not published in the conventional sense and so is not always widely available to the general public.

In archaeology, grey literature reports reflect different stages of a fieldwork project worth recording and disseminating information about, such as watching briefs, excavation, evaluation, survey reports and related artefact and ecofact analysis. From a research and scholarly point of view these reports have significant advantages over traditional types of publication. They are relatively cheap and very flexible (Weintraub 2000). Authors can elaborate and provide sufficient detail where necessary without being restricted by page limits. Grey literature reports can contain comprehensive explanations, diagrams, summaries and statistics that deliver in depth analysis and discussion usually not possible to be accommodated by traditional publication. The bulk of the UK archaeological grey literature comes from commercial archaeology units who may have been funded to describe the immediate investigation but not subsequent extended analysis

From a commercial archaeology point of view, such documents can be also very useful but their practice may lack some of the deeper analysis level found in academic work. On

2 Both PPG15 and PPG16 have been replaced in 2010 by the Planning Policy Statement 5 (PPS5) Planning for the Historic Environment (PPS5), which sets out the Government's planning policies on the conservation of the historic environment

the other hand, there are concerns regarding the accessibility of such reports. Grey literature is not always accessible to the general public and within the profession of archaeology access might be restricted within the boundaries of a commercial organisation or government institution. The current fragmented and often competitive situation in which archaeology fieldwork operates, where private contractors are often unaware of the work carried out at local or national level, does not support the full potential for collaboration and sharing of information. In response to this challenge, the All-Party Parliamentary Archaeology Group (APPAG 2003) provided a set of recommendations (174-177) that highlighted the need for managing and enabling access to "grey literature" and the role of the Archaeology Data Service (ADS) in archiving and making available excavation archives in digital form.

Dissemination of information via the WWW offers a huge potential for promoting collaboration and accessing information. In particular to archaeology, the WWW and Information Technology offer the opportunity to improve archaeological practice, not only by enabling access to information but also by changing how information is structured and the way research is conducted (Falkingham 2005). In response to APPAG recommendations, the ADS initiated the Online AccesS to the Index of archaeological investigationS (OASIS) project (Richards and Hardman 2008). The project is a joint effort of UK archaeology research groups, institutions, and organizations aiming to enable online dissemination and maintenance of a unified repository of archaeological grey literature reports. The repository stores and disseminates grey literature as word-processed documents or as PDF files. However, such file formats store information in a monolithic structure which has little or limited capacity to represent content in an interoperable and machine understandable way.

Additional efforts in the use of semantic technologies for the dissemination of archaeological information originate from the Semantic Technologies for Archaeological Resources (STAR) project (Tudhope, Binding and May 2008). The project aims to support 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 complex and semantically defined queries over archaeological digital resources.

A major hindrance to the swift development of archaeological research is the laborious and intensive process of finding new information in reports. 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" (Richards and Hardman 2008). Thus, it is highly desirable to be able to search effectively within and across archaeological reports in order to find information that promotes quality research. It was proposed that new and innovative ways of presenting content online based on XML technologies could break through interoperability barriers and enable long-term solutions to information discovery and maintenance (Falkingham 2005; Ross 2003).

The traditional model of Information Retrieval is based on the use of index terms (keywords). The process from full text (document) representation to the level of index terms is a multilevel abstraction traditionally achieved by the use of statistics. The statistical model of Information Retrieval based on keyword matching has been criticised as inefficient for overcoming language ambiguities which emerge from the use of natural language in query formulation and text authoring. Such language ambiguities concern use of polysemous terms, i.e. words that have more one meaning, as for example *bank* (a commercial or a river bank), and synonymous words where the same concept can be expressed by more than one word, as for example *car* and *automobile.* Other cases of ambiguity, such as *variability*, *conjunction* and *underspecification,* reflect the elaborate use of language where language expressions convey meaning that is open to interpretation.

It is suggested that adoption of Natural Language Processing techniques in document indexing and retrieval can support overcome such barriers imposed by the use of natural language (Smeaton 1997; Lewis and Jones 1996; Moens 2006). In addition, conventional Information Retrieval practices operate on the level of documents not on the level of information chunks. Usually users need to read through large passages of text before they find the piece of information that satisfies their information need and in the worst case scenario users might read through irrelevant pieces of information before they try another document in the results. Natural Language Processing techniques can be used to identify "rich" meaningful pieces of text (phrases), which can enhance document study and support retrieval of information closer to the users' needs.

A particular NLP technique which can be employed to address the above language ambiguity issues is Information Extraction (IE), defined as a text analysis task aimed at extracting targeted information from context (Cowie and 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. Such abstractions can connect 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). In particular the Conceptual Reference Model (CRM) of the International Committee of Documentation (CIDOC) aimed at "enabling information exchange and integration between heterogeneous sources of cultural heritage information" is believed to be capable of supporting NLP techniques to resolve free text information into a formal logical form (Crofts et al. 2009[3]).

The research question of the thesis is focused on the semantic indexing of archaeology grey literature. The semantic indexing result is targeted at supporting complex and semantically defined queries that facilitate information retrieval and cross searching over archaeological digital resources. The research effort contributes to the STAR project, which aims to develop semantic methods for linking digital archive databases, vocabularies and associated unpublished on-line documents (Tudhope, Binding and May 2008). The role of the CIDOC CRM ontology and its English Heritage extension CRM-EH for achieving semantic interoperability over diverse information resources is central to STAR and to the semantic indexing effort.

1.3 Thesis Layout

The thesis is organised into four main sections; 1) Background, 2) Preparation, 3) OPTIMA Pipeline, 4) Results and Conclusions. Each section contains two to three chapters which discuss the research phases from early development to final system evaluation.

The *Background* section contains chapter 1 (Introduction) and chapter 2 (Literature review) which present the main research question, relevant background information and literature review. Chapter 1 introduces the issue of interoperable semantic access to grey literature reports, the domain of the research and the motivations driving the system's development. Chapter 2 discusses the main literature review of the work focused on the topics of Natural Language Processing, Information Extraction, Ontologies, Semantics and relevant projects and tools. The review presents the main principles, theories, and technologies that support the research study. Individual chapters also contain additional elements of literature review that support the discussion of each chapter. Thus, Chapter 2 is an overview of the subject domains that contribute to the research study while literature review of individual chapters is more focused on the chapters' argumentation and topics.

[3] The ISO Standard (ISO 21127:2006) CIDOC-CRM is released under a regular version control. The thesis adopts version 5.0.1 (released March 2009) which was the current version during system development and also the version adopted by the STAR project.

The second section (*Preparation*) discusses the preparatory stages and work leading to the main system's development. The thesis is not organised according to a chronological order but discusses the development as a coherent whole with emphasis on the delivery of the final system. The *Preparation* section reveals early achievements and preparation of resources, evidence of the iterative process followed during development.

Chapter 3 discusses a prototype development aimed at exploring the role of ontological and terminological resources in the delivery of semantic indices via Natural Language Processing (NLP) techniques, in particular Information Extraction (IE). The prototype system investigated the method of semantic annotation, a form of metadata abstraction, using rule-based IE techniques. The chapter also discusses the evaluation method and results of the prototype development. The results of the evaluation were valuable and helped in drawing useful conclusions regarding the full potential of the method. The experience and knowledge gained during prototype development was directed towards improving and refining the full-scale system.

Chapter 4 introduces the topic of Named Entity Recognition (NER). The discussion provides an overview of NER with regards to origins, schools of thought and relating projects. The role of terminological resources in the NER task is also addressed. An analysis study of the contributing resources reveals their particular characteristics and arrangements. The chapter discusses in detail the integration of such resources in the system and their transformation and enhancement process to resources capable of supporting the NER with respect to a given ontology. The role of the pre-processing stage is also revealed for the delivery of generic annotation types, such as noun phrases, verb phrases and headings that are used by the succeeding stages of the pipeline.

The *OPTIMA Pipeline* section discusses the development stages of the main (full-scale) system aimed at the delivery of semantic indices of grey literature documents (archaeological reports) with respect to CIDOC CRM and CRM-EH ontologies. The term *pipeline* is used to describe OPTIMA due to the cascading order in which the system delivers results and outputs (an alternative would be to describe the system as *application).* The section contains three chapters each one discussing a different stage of the pipeline.

Chapter 5 addresses the process of Named Entity Recognition with respect to the CIDOC CRM ontology. The discussion reveals the various stages involved in the process of delivering textual abstractions (semantic annotations) with respect to the four CRM entities, Physical **O**bject, **Pl**ace, **Ti**me Appellation and **Ma**terial, from which the pipeline (OPTIMA) took its name. The chapter discusses the process of terminological resources

exploitation via a controlled semantic expansion technique that exploits synonym and hierarchical relationships of concepts. In addition, the individual stages that contribute in the NER process are also revealed, such as noun phrase validation, word sense disambiguation and adjectival expansion and conjunction. The chapter concludes with the negation detection phase, which reveals the process of adaptation and use of the NegEx algorithm (Chapman et al. 2001) in the domain of archaeological grey literature reports.

Chapter 6 is dedicated to the discussion of the task of Relation Extraction with respect to the CRM-EH ontology. The discussion reveals the role of the CRM-EH ontology in directing detection of textual instances (phrases) that relate in pairs, entities previously identified by the NER phase. The chapter commences with a literature review on the issue of Relation Extraction providing background information and relevant work. The role of the Zipfian distribution is revealed in the process of defining syntactical patterns capable of detecting 'rich' phrases of entity relation. The details of a corpus analysis process, which informed the definition of relation extraction patterns is revealed and rules and patterns are discussed via example cases.

Chapter 7 discusses the delivery and usage of semantic indices of grey literature by information retrieval and document inspection applications. The transformation process of semantic annotation to semantic indices is discussed in detail along with the role of interoperable formats such as XML and RDF. The employment of semantic indices by two web applications is presented and real-world examples are discussed. The web applications enable document inspection, cross search and retrieval with respect to semantic attributes. The chapter also reveals examples of false positive results which are delivered by the semantic annotation process and passed into the definition of semantic indices.

The *Results and Conclusions* section contains chapter 8 (Evaluation) and chapter 9 (Conclusion and Future work). Chapter 8 discusses the evaluation methodology and results based on the system's performance which is benchmarked using established evaluation processes. The discussion reveals a set of system configurations which are used during evaluation in order assess the system's performance under different annotation types and conditions. Chapter 9 discusses the main conclusions regarding the achievement and contributions of the work. In addition, strengths and weaknesses of the system and methodology are highlighted and issues of generalisation of the work and future plans are discussed.

Chapter 2

Literature Review

2.1 Introduction

The current chapter provides an overview of the main domains that relate to the research aims of the thesis. The discussion presents basic and fundamental notions of the domains while defining the research environment within which the thesis contributes. Additional literature is also included in the individual chapters targeted at supporting the aims of each chapter.

In detail, the current chapter discusses the notion of Natural Language Processing (NLP) and in particular the role of Information Extraction (IE) in advancing Information Retrieval practises. The discussion also reveals the potential of semantic technologies for advancing information seeking activities and in particular the role of ontologies in encapsulating knowledge and describing semantic annotations that support rich metadata descriptions. A range of semantic technology projects, indicative of the contemporary approaches in supporting information needs of the Cultural Heritage domain are also discussed. The chapter concludes with a brief discussion on Language Engineering frameworks and tools leading to the adapted framework of the thesis.

2.2 Natural Language Processing

The field of NLP is not a new area of research and application. It has been in constant development since the early 1940's and it is still very active, continuing to thrive and to deliver applications and research outcomes. Jurafsky and Martin (2000) define speech and language processing as those "*computational techniques that process spoken and written human language as 'language'*". Although Jurafsky's definition describes the "holy grail" of speech and language processing, current NLP systems are not capable of comprehending and producing natural language at the same level as humans.

Dale et al. (2000) defines NLP as "*the design and implementation of computational machinery that communicates with humans using natural language*". Dale provides a broad definition for NLP, acknowledging that at its most ambitious, NLP research aims to design and develop artificially intelligent systems that are capable of *"using natural*

language as fluently and flexible as humans do". Other definitions of NLP include Fernandez and Garcia-Serrano (2000) who describe NLP as the Computing Science area that focuses on developing "*software systems that use language analysis functionalities to solve real problems*" and Liddy (2003) who defines NLP as a "*theoretically motivated range of computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achievement human-like language processing for a range of task or applications*".

While speech processing is not mentioned in the above definition, it would be misleading to define NLP as a discipline which is concerned only with the computer processing of the written form of natural languages. Gazdar (1996) argues that NLP and Speech Processing have been separate fields with no real overlap between personnel, journals and conferences. He highlights that Speech Processing, studies the problems that are specific to the processing of spoken forms of language while NLP, studies the problems that are common to the processing of both spoken and written forms of language.

The range of available definitions is indicative of the volume of research and development that has been offered in the field of NLP for a long period of time. The above definitions do not contradict but present different and complementary views on NLP. The thesis adopts Liddy's (2003) definition based on the merit that it is the most complete definition which includes both abstractness in defining NLP as the computational means for processing natural language as humans do, as well as specificity in defining NLP as the computational technique that is capable of analysing and representing natural language in one or more levels of linguistic analysis.

2.2.1 Levels of Linguistic Analysis

Liddy's definition of NLP emphasises the presence of linguistic analysis during the processing of natural language by computational techniques. Jurafksy (2000) provides an illustrative example for distinguishing data process from natural language process. In his example he highlights the fact that when *knowledge of language* is present then the process can be described as a language process. In any other case the process is most probably a data process. For example, counting the bytes of a text files is purely data processing. On the other hand, counting the number of words in a text file is natural language processing since knowledge about what constitutes a word must be present.

The study of language can be divided into *syntactic*, *semantic* and *pragmatic* (McGilvray 1999, Liddy 2003, Jurafsky and Martin 2000). *Syntax* deals with the lowest

level in the study of language, and is concerned with the way words are grouped and ordered when composition of sentences takes place. In other words, syntax is focused on analysing the grammatical and structural relationships between words. For example, syntax might be concerned with the structure of a specific verb phrase and how its various parts (verb, determiner, preposition) are constructed together to convey a meaning since the order and dependency between words are important to the composition of meaningful phrases.

Semantics deals with the meaning of words, connecting syntactic descriptions with the world to which the words refer. Referring to a particular item or thing is not always sufficient to provide the semantics of a word. For example "workstation" and "computer" both refer to the same device, but semantically the two terms are different and carry a different meaning. Thus, the concept of the sense of a word is equally important for defining the semantics of a word and for providing semantic disambiguation of polysemous words that have more than one meaning, as for example in the case of bank which can refer either to a river bank or to a commercial bank.

Pragmatics is the highest level in the study of language and is concerned with the study of how language is used to accomplish specific communication goals. To study a word pragmatically is to understand who is using the word in what context and to accomplish which goal. Without pragmatics it would have been impossible for metaphors and allegories to be used and understood.

Liddy (2003) and Jurafsky and Martin (2000) recognise *Phonology*, *Morphology* and *Discourse* as three more additional levels of linguistic analysis. *Phonology* is concerned with the study of linguistic sounds and the interpretation of sounds within words. *Morphology* deals with the study of morphemes which are the smallest components of meaning. Morphological analysis is capable of recognising the meaning conveyed by the use of specific morphemes, i.e. prefixes, roots and suffixes. *Discourse* analysis expands beyond the limits of a single utterance. This level of analysis is concerned with the properties of text that convey meaning by making connections between phrases and sentences.

2.2.2 NLP Schools of Thought

Since the early days of NLP, two very distinct schools of thoughts have dominated the field, the *Symbolic*, otherwise known as the *Rationalist* or *Logic* approach and the *Statistical*, otherwise known as the *Empirical* (Dale et al. 2000, Fernandez and Garcia-Serrano 2000, Jurafsky and Martin 2000, Liddy 2003).

The *Symbolic* approach is concerned with the construction of computational formalisms, heavily influenced by Chomsky's theory of generative linguistics. Based on representation of facts about language, symbolic systems develop human crafted rules, knowledge resources, and inference engines for the accomplishment of various language processing tasks. The approach delivers NLP systems that perform well under identifiable linguistic behaviour which is used to model the system's operation. On the other hand, symbolic systems are less flexible at coping with noisy and unexpected input and are hard to adapt dynamically to new domains.

The *Statistical* approach is a quantitative method dominated by the theory of statistics. It is based on the use of large text corpora input for the definition of mathematical models which, with the help of statistics can approximate linguistic phenomena. Statistical approaches have been used successfully in speech recognition and part-of-speech tagging and perform well in cases where linguistic phenomena are irregular and not easy to model. On the other hand, statistical methods rely heavily on the quality of the primary input. Insufficient input can harm the overall performance of the system and can make the system less flexible at coping with cases that have not been covered by the input resource.

Scholars do not see the two different approaches as being rival. Instead they understand them as being complimentary where each one has each own virtues. The *Hybrid* approach attempts to bring under a common ground both statistical and symbolic practises, motivated by the need for robustness and real-world application

2.2.3 NLP Potential in Information Retrieval

2.2.3.1 Information Retrieval

Information Retrieval (IR) addresses the task of finding relevant information resources from a collection of documents to satisfy specific user queries which, originate from generic information needs. The classical model of information retrieval is based on the idea that each document in a collection is represented by a set of terms, known as *index* terms (Baeza-Yates and Ribeiro-Neto 1999). The use of index terms as a "*bag of words*" to

represent the meaning of a document is well widespread in information retrieval, while a number of variations make use of the index terms in different ways as for instance assigning weighting to index term depending on the indexing approach followed.

Historically, index terms in a form of a "*bag of words*" have been used in IR processes to capture an abstract layer of textual document representations. The process from full text representation to the level of index terms is a multilevel abstraction from full document, to text structure, to word groups and to the final set of index terms "*bag of words*". The *tf*idf* weight (term frequency–inverse document frequency) is a widely used statistical figure that abstracts document keywords based on word frequency within document and across corpus while preventing from keywords of commonly occurring across-corpus words, such as articles and prepositions (Baeza-Yates and Ribeiro-Neto 1999)

Information Retrieval models are concerned with their effectiveness in responding to user queries and in retrieving results of value to the user. The evaluation of the effectiveness of information retrieval can be summarised under two distinct factors, the *Precision* and the *Recall* of the retrieved results. *Precision* (P) is defined as the fraction of documents retrieved which are relevant to a generic user need. *Recall* (R) is the fraction of the documents that are relevant to a query and have been successfully retrieved (Baeza-Yates and Ribeiro-Neto 1999).

$$Precision = \frac{(relevant\_documents \cap retrieved\_documents)}{retrieved\_documents}$$

$$Recall = \frac{(relevant\_documents \cap retrieved\_documents)}{relevant\_documents}$$

The above two metric factors are the focus of interest for evaluation exercises for retrieval effectiveness such as TREC (Text Retrieval Conference) which organises competitions on the effectiveness of information retrieval systems in performing retrieval tasks operating on large volumes of textual information.

Most operational information retrieval systems used today incorporate term indexing together with statistical model implementations to provide a framework for performing information retrieval tasks. Such IR systems have grown up and improved due to the considerable technological advances offered in computing the recent decades.

IR evaluation methods such as TREC have revealed that statistical methods of IR can operate well over a large information corpus. On the other hand, the tendency in such statistical models to rely on string matching, arguably limits the models' ability to overcome language ambiguities which emerge from the use of natural language in query

formulation and text authoring (Smeaton 1997; Lewis and Jones 1996; Moens 2006).

2.2.3.2 Language Ambiguities and NLP for Information Retrieval

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 systems.

Polysemous words that have multiple meanings and *synonyms* generate ambiguity which term matching statistical methods are ill-equipped to deal with (Smeaton 1997; Moens 2006). *Variability* in how concepts are expressed is a factor that also creates ambiguity. For instance the term "polished" could convey several meanings, such as lustrous and bright, or refined and updated, or even dressed and disguised.

Underspecification describes a situation where a term carries a partial and underspecified meaning, which is open to interpretation. The term "big" for example in the phrase "I saw him talking to the big guy" does not clearly specify whether the term big refers to the size or the status of the person. In addition, *Conjunction* is a form of language ambiguity frequently occurring during document authoring when conjunction between the head of noun phrases appears to make the language more concise (Lewis and Jones 1996). For instance in the phrase "Examine the enclosed memo and photo", it is not clear whether the photo is enclosed together with the memo for examination.

The ideal goal of any information retrieval system is to achieve high Precision and high Recall (Baeza-Yates and Ribeiro-Neto 1999). However attempts to increase the Precision of operational information retrieval systems often cause Recall to drop. It is highly desirable for information retrieval systems to overcome the language ambiguities described above and so to increase Precision. Researchers and scholars envisage different ways where the performance of information retrieval systems could be improved by the use of NLP tools, techniques and resources. The approaches of amalgamating NLP with information retrieval vary and proposals reveal different techniques for bringing NLP and IR together under a common application framework (Allan et al. 2003; Cunningham 2005; Lewis and Jones 1996; Moens 2006; Smeaton 1997; Wilks 2009).

Lewis and Jones (1996) recognised the potential of a hybrid approach for employing NLP techniques in statistical retrieval methods to improve retrieval performance. Smeaton (1997) also recognised the potential of Information Extraction (IE) in advancing IR performance and suggested that *Named Entity Recognition* could assist the indexing process by identifying index terms for document representation. Cunningham (2005) also supports the potential of IE in indexing while Allan et al. (2003) emphasise the potential of

IE not only in named entity extraction but also in relation (between entities) extraction, putting IE at the heart of the anticipated progress in NLP. In addition, Moens (2006) suggests that IE could be integrated with statistical vector based and probabilistic models of IR systems. She argues that the idea of using semantic information for indexing was initially expressed by Zellig Haris back in 1959, but it is only today that technology has matured to allow the computational overhead of using IE in IR.

On the other hand, sceptics of the potential of NLP argue that linguistically-motivated indexing (LMI) is not needed for effective retrieval and that experiments have shown only marginal benefits (Jones 1999). Voorhees (1999) argues that statistical indexing captures important aspects of natural language by implicit processing. She argues that unless done carefully, linguistic processing may harm the overall retrieval performance.

Wilks (2009) has a different point of view he argues that after "*40 years, IR ought to have improved more than it has*". While he acknowledges Jones' point of view on the marginal benefit of LMI in IR, he highlights that IR evaluation regimes have been in connection with statistical methods, often resistant to linguistic approaches. He then concludes that IE in combination with conceptual models and knowledge representations will have a major role in pattern-matching and template finding retrieval, something that remained untested by conventional IR.

2.2.3.3 Indexing and Classification with Terminological Resources

As discussed, NLP methods carry the potential to improve indexing techniques and to enhance information retrieval practices that deal with language ambiguity. However, indexing approaches can be further enhanced beyond use of NLP techniques. The following paragraphs discuss the role of terminological resources such as controlled vocabulary and thesauri in automatic indexing and classification.

The automatic keyphrase indexing system KEA (Medelyan and Witten 2006) assigns indexing terms to documents using a controlled vocabulary. The indexing algorithm (KEA++, currently updated to the Maui algorithm) that succeeded an earlier KEA algorithm, identifies thesaurus terms that relate to document content. A machine learning model then filters the most significant keyphrases based on a range of properties and features, such as position of keyphrase in document, length of keyphrase and frequency in terms of *TF*IDF*. KEA has been evaluated on 200 full-text documents originating from the UN Food and Agriculture Organization (FAO), using the Agrovoc domain specific thesaurus. The evaluation results demonstrated the ability of the system to eliminate meaningless and incorrect indexing phrases.

The use of controlled vocabularies is also evident in the field of term extraction. YaTeA (Aubin and Hamon 2006) is a tuneable term extractor that exploits linguistic-based rules and terminological resources for the extraction of noun phrases. The role of terminologies is to support disambiguation during chunking, parsing and extraction steps, delivering candidate maximal noun phrases that cannot be further split or deleted. The system has been evaluated on a biomedical corpus of 16,000 sentences describing genomic interaction, using three distinct controlled vocabularies; the Gene Ontology resource (GO), the Medical Subject Heading thesaurus (MeSH) and the Term Acquired in Corpus (TAC) resource. Results showed that use of terminological resources on a biomedical corpus supports identification and extraction of maximal noun phrases (Aubin and Hamon 2006).

The role of controlled vocabulary has also been explored in the field of automated classification aimed at supporting information retrieval. Automated (subject) classification "*denotes machine-based organization of related information objects into topically related groups*" (Golub 2006). It can be distinguished into 3 main approaches; *Text Categorisation,* which is a machine learning supervised approach where classification is learnt from a training corpus of manually assigned classes, *Document Clustering,* which is an unsupervised machine learning approach where classification classes and relationships between them derived automatically and *Document Classification,* which originates from library science and supports classification via intellectually created classification schemes (Golub 2006). *Document Classification* does not require a training set for providing classification while it can provide hierarchical browsing interfaces for accessing document collections, which is not well supported from unsupervised document clustering methods.

Golub, Hamon and Ardö (2007) devised a string matching algorithm for automated document classification for the purposes of information retrieval based on controlled vocabulary. The algorithm was applied to classification of documents in the field of engineering using the Engineering Information (Ei) thesaurus for supporting term identification and for providing hierarchical classification of engineering topics. Ei terms were assembled into a parameterised term list that assigned to terms class and weight (indicating how appropriate a terms is for the assigned class). The algorithm classified documents based on a string matching mechanism that exploited the term list and assigned weighted classes to documents, with final selection based on a heuristically defined cut-off. Results were reported to be comparable with supervised machine-learning algorithms.

2.3 Information Extraction

Information Extraction (IE) is a specific NLP technique defined as a text analysis task aimed at extracting targeted information from context (Cowie and Lehnert 1996; Gaizauskas and Wilks 1998; Moens 2006). It is a process where a textual input is analysed to form a textual output able for further manipulation. Such data manipulation may be then aimed for automatic database population, machine translation tasks, term indexing analysis, text summary algorithms and other.

Hobbs (1993) describes the generic information extraction system as "*a cascade of transducers or modules that at each step add structure and often lose information, hopefully irrelevant, by applying rules that are acquired manually and/or automatically*". He recognises that each information extraction system is dictated by its own set of modules however, he highlights a set of 10 individual modules that contribute to the general architecture of every information extraction system. These are;

- *Text zone* analyser to divide input into segments,
- *Pre-processor* to convert segments into sentences based on part-of-speech recognition,
- *Filter* to discard irrelevant sentences generated in the previous process,
- *Pre-parser* to detect small scale structures as noun groups, verb group and modifiers,
- *Parser* to produce a set of parse tree fragments possibly complete that describe the structure of a sentence,
- *Fragment combiner* to complete the parsing of incomplete parse tree fragments into a logical form for the whole sentence,
- *Semantic Interpreter* to generate meaning representation structures from a parse tree or a parse tree fragment,
- *Lexical Disambiguation* to resolve any ambiguities of terms in the logical form,
- *Coreference Resolution* to connect different descriptions of the same entity in different parts of text and
- *Template generator* to generate the final representations of the extracted text.

Information Extraction and Information Retrieval operations are fundamentally different and as such cannot be seen as two competitive methods employed to resolve the same problem. They have been described as two complementary methods where their

combination promises the creation of new powerful tools in text processing (Allan et al. 2003; Cunningham 2005; Lewis and Jones 1996; Moens 2006; Smeaton 1997; Wilks 2009).

The two technologies have different and distinct historical backgrounds. Computational Linguistics and NLP have formed the environment within which IE developed, whereas Information Retrieval growth was based on Information Theory, Probability Theory and Statistics. For the average user, it would not be hard to imagine the operation of an IR system, since these kind of systems are widely used when searching the Web or a local library catalogue. On the other hand, IE systems arguably could not be described as applications available to the average user since such systems operation is usually closely bound to an application scenario or domain.

2.3.1 The Role of the Machine Understanding Conference (MUC)

The contribution of the Machine Understanding Conference (MUC) in a period of ten years from 1987 to 1997 has been significant and supported the growth of the IE field, providing the funds and a common ground, for evaluation, and sharing of knowledge and resources in Information Extraction. The conference adopted *precision* and *recall* measurements while redefining them to suite the information extraction task, including measurements for incorrect and partially correct results (Grishman and Sundheim 1996).

The fourth MUC marked the beginning of the conference inclusion in the TIPSTER programme. TIPSTER funded by DARPA and various other US Government agencies focused on three underlying technologies; *Document Detection*, *Information Extraction*, and *Summarisation.* Efforts involved the creation of a standard architecture for information retrieval and extraction systems, while improving the portability and re-usability of information extraction techniques. The programme has enjoyed three development phases from 1991 to 1998 and achieved its purposes under the directions of Ralph Grishman of NYU and the efforts of the TIPSTER Architecture Working Group.

The sixth MUC conference provided for the first time to participants the option to choose to perform one or more of four smaller evaluation tasks, described as *Named Entity Recognition* (NER), *Coreference Identification*, *Template Element Filling* and *Scenario Template*. The MUC programme concluded in 1999, an effort which occupied seven conferences and spanned for a decade. The extracts and conclusions that have been drawn from the MUC's have influenced the design and development of many information extraction systems since (Cunningham et al. 1996; Gauzaskas and Wilks 1998).

The Automatic Content Extraction (ACE) programme, successor of MUC in the evaluation of information extraction technology, directed the evaluation effort towards a finer inference analysis of human language. The programme described four evaluation tasks; *Recognition of Entities, Recognition of Relations, Event Extraction, Extraction from Speech and OCR input* (Doddington et al. 2004).

The evaluation tasks of the programme are challenging information extraction methods that operate on the semantic-entity level beyond the word-term limit. Recognition of entities in text involves *Coreference Resolution* for identifying all entity instances, an issue not addressed by NER. The tasks of *Recognition of Relations and Event Extraction* are also described and are targeted at detection and categorization of events and relations between entities. The latest April 2008 event of the ACE series, involved multilingual tasks focused on entity and relation recognition in Arabic and English within-document and cross-document tasks.

2.3.2 Types of Information Extraction Systems

Information extraction systems fall into two distinct categories; *Rule-Based* (hand-crafted) and *Machine Learning* systems (Feldman et al. 2002). During the seven MUCs, the involvement of rule-based information extraction systems has been influential. Systems such as TACITUS, FASTUS, PIE and LaSIE-II have used with success hand crafted rules to answer a range of information extraction scenarios set by the conference committee (Lin 1995; Hobbs et al. 1993; Humphreys et al. 1998).

The issue of information systems portability quickly gained attention. During MUC-4 the AutoSlog tool introduced a semi-automatic technique for defining information extraction patterns as a way of improving system's portability to new domains and scenarios. An updated and fully automated version of AutoSlog, named CRYSTAL, participated in MUC-5 introducing the involvement of machine learning information extraction systems in the conference. Although the performance of CRYSTAL did not match those of hand-crafted rules, it managed to deliver promising results that met 90% the performance of rule-based systems (Soderland et al. 1995; Soderland et al. 1997).

2.3.2.1 Rule-based Information Extraction Systems

Rule-based systems consist of cascaded finite state traducers that process input in successive stages. Dictated by a pattern matching mechanism, such systems are targeted at building abstractions that correspond to specific information extraction scenarios. Hand-crafted rules make use of domain knowledge and domain-independent linguistic syntax, in

order to negotiate semantics and pragmatics in context and to extract information for a defined problem. It is reported that rule-based systems can achieve high levels of precision between 80%-90% when identify general purpose entities from financial news documents such as Person, Location, Organisation etc. (Feldman et al. 2002; Lin 1995; Hobbs et al. 1993).

The definition of hand-crafted rules is a labour intensive task that requires domain knowledge and good understanding of the information extraction problem. For this reason rule-based systems have been criticised as costly and inflexible, having limited portability and adaptability to new information extraction scenarios. However, developers of rule-based systems claim that, depending on the information extraction task, the linguistic complexity can be bypassed and a small number of rules can be used to extract large sets of variant information.

2.3.2.2 Machine Learning Information Extraction Systems

The use of machine learning has been envisaged to be the element to break through the domain-dependencies of rule-based information extraction systems (Moens 2006, Ciravegna and Lavelli 2004). Machine Learning is a discipline that grew from the research of Artificial Intelligence, which is concerned with the design of algorithms that enable computers to "adapt" to external conditions. The term "learning" obviously does have the precise meaning that learn has in human intelligence context. Learning in the artificial intelligence context describes the condition where a computer programme is able to alter its "behaviour", that is to alter structure, data or algorithmic behaviour in response to an input or external information (Nilsson 2005).

Machine learning strategies can support *supervised* and *unsupervised* learning activities. When supervised the learning process is based upon the provision of a training data set which is used by the machine learning process in order to deliver generalisation of the extraction rules, able to perform a large scale exercise over a large corpus. The general idea of using supervised machine learning in Information Extraction systems is to use human experts to annotate a desired set of information fragments in an exercise involving a small corpus of training documents. The training set of documents is then utilised in a machine learning process for generalisation of the extraction rules, which are able to perform a large scale exercise on a large corpus. It is believed to be easier to annotate a small corpus of training documents than to create hand-crafted extraction rules, since the later requires programming expertise and domain knowledge (Moens 2006)

During *unsupervised* learning, human intervention is not present and the output of the training data set is not characterised by any desired label. Instead a probabilistic, clustering technique is employed to partition the training data set and to describe the output result, which generalisation of a larger collection would expand upon (Nilsson 2005). Unsupervised information extraction is very challenging and systems are not proven to be able to perform at an operational level (Uren et al. 2006; Wilks and Brewster 2009).

Supervised information extraction systems are more widely adopted and have managed to delivered successful results at an operational level. However, criticisms of the supervised learning methods highlight the dependence of the information extraction results on the quality of the training set, the impact of the type of learned data to the maintainability of the information extraction system and the difficulty in predicting which learning algorithm will produce the most optimum result (Wilks and Brewster 2009).

2.4 Ontology

Computer scientists today, more than ever before, express an appreciation of conceptual structures and their potential in mediating formal representations. Ontologies are becoming widely adopted in Artificial Intelligence, Computational Linguistic and Database systems while, many believe that the true potential of semantic computations resides in the potential of ontologies to aid understanding and standardisation (Guarino 1998).

The term Ontology was first used in a philosophical context by Aristotle in his work Metaphysics for defining the *very nature and structure* of "reality". Ontology focuses on the study of "things" and of their attributes *per se* without depending on a particular language or taking into account considerations about actual or physical existence of "things". Hence, it is perfectly valid to describe an ontology of mythological creatures and deities which does not depend on Greek or Latin naming practise i.e. Aphrodite versus Venus, Athena versus Minerva etc.

Information Systems (IS) on the other hand, adopts a more pragmatic approach for the definition of ontologies, defined as computational artefacts of specific vocabulary aimed at describing a certain "reality" (Guarino et al. 2009). In their simplest form, ontologies are hierarchical structures that describe hierarchical relationships between concepts. More advanced ontologies make use of sophisticated axioms that dictate the intended interpretation of certain relationships which expand beyond the definition of simple hierarchies.

Such ontologies can carry the role of an integration vehicle capable of connecting disparate datasets and information resources under a common semantic and schematic layer. Such layers can reconcile modelling and conceptual variations and enable domain interoperability between different schemas and resources that deal with common or alike data and information management issues. In the case of the STAR project (Tudhope et al. 2011) which is discussed in detail below (section 2.6.1), the archaeological extension of the CIDOC Conceptual Reference Model (CRM) (section 2.4.3) was adopted for enabling semantic interoperability over diverse archaeological datasets and information resources.

2.4.1 Conceptualization

The above section has intentionally made use of two forms of the word "ontology"; one written with capital "O" and another with lowercase "o". The first reading relates to a specific philosophic discipline, while the latter relates to a certain vision of reality. However, the latter sense also generates some dispute between the communities of philosophy and computer science. In philosophical terms an ontology does not depend on any language. Therefore, Aristotle's ontology regardless of language is always the same. On the other hand, computer science ontologies are constituted by vocabularies used to describe a given reality and assumptions about the indented meaning of vocabulary words hence, such structures are language dependent engineering artefacts.

Guarino (1998) addresses the above conflict between the two senses of the word "ontology" by adopting the term *conceptualization*, an abstract and simplified view of the world, for describing ontologies in their philosophical sense. Therefore, two ontologies can adopt two different vocabularies but they can share the same *conceptualization*. The level in which the intended meaning of a vocabulary is followed by an ontology is known as *ontological commitment* to a particular *conceptualization*.

An ontology can get closer to a particular *conceptualization* by adopting a rich set of axioms and domain relations to the level that can be "perfect", thus to exactly coincide with its target *conceptualization*. On the other hand, weak *ontological commitment* can bring ontology to a non-practical usage, diminishing all the benefits of a shared understanding model (Guarino et al. 2009).

2.4.2 Ontology Types

Guarino (1998) distinguishes three major types of ontologies. *Top level* (also known as Upper level) ontologies, that describe abstract concepts (e.g. time, place), and general axioms about concepts such as relationships and their intended use. Ontologies of this kind are not coupled to a specific problem or domain and therefore can act as unifying models of shared understanding. *Domain* ontologies that describe a generic domain (also referred as core ontologies), provide specialisations about related vocabulary and domain relationships. *Application* ontologies, which are specialisations of *Domain* ontologies, describe a specific domain coupled with a particular task.

Ontologies can be employed to support the creation and functionality of ontology-driven Information Systems. Guarino (1998) argues that ontologies can be used at development and at run time acting as an integral part of a database, a user interface or an application programme component. The level of integration distinguishes an ontology-aware from an ontology-driven Information System. In the first case, an Information System component is aware of an ontology which is invoked whenever is required by a specific function. In the second case the ontology is an integral component of the IS architecture.

Wilks (2003) follows a less "clean" and "pure" approach in his definition of ontologies and their purpose in IS. He argues that facts about word and about world are often mixed, as for example in the case of WordNet. Therefore, we cannot achieve to have pure logical representations that are detached from all language qualities that are used to describe them. He concludes that ontological and lexical resources "*do not differ in content only in principal"*. Hence, any model or structure is justified by its purpose which is evaluated against a desired outcome and such evaluation, according to his view, overrides any other consideration.

2.4.3 The Cultural Heritage Ontology CIDOC – CRM

Information System ontologies are computational artefacts aimed at providing common conceptual ground for information integration, logical inference and conceptualization at multiple levels. Such integration can be aimed at data analysis and understanding, use of descriptive vocabularies and automated mapping between data, metadata and ontological instances. Formal handling of information and integration in cultural heritage, poses significant challenges due to the inherited *diversity* and *incompleteness* of information

when recording cultural data (Doerr 2003). For example dating information about a museum collection and about an archaeological excavation differs significantly on the levels of complexity, inference and justification. Moreover, there is a natural difficulty of computer scientists to fully comprehend cultural concepts and equally it is difficult for cultural professionals to communicate such concepts to non-domain experts.

CIDOC CRM, the Conceptual Reference Model (CRM) of the International Committee of Documentation (CIDOC) is addressing the above issues by "*enabling information exchange and integration between heterogeneous sources of cultural heritage information*". Defined as an ISO Standard (ISO 21127:2006), CIDOC CRM is a comprehensive semantic framework that makes available *semantic definitions* and *clarifications* that promote shared understanding and enable transformation of disparate and localised cultural heritage information resources into a *coherent global resource* (Crofts et al. 2009).

The CRM ontology is a guide to good practice in conceptual modelling that aims to enable semantic interoperability of cultural heritage information. It aims to support domain experts and IT developers to address a range of software engineering tasks, such as system specification, data transformation, data migration, query formulation and retrieval from heterogeneous resources, as well as enabling natural language algorithms to resolve free text into formalistic structures.

To satisfy maximum interoperability and minimum ontological commitment the CRM ontology is based on the following modelling principles.

- *Monotonicity* to allow information integration in an "open world" where existing CRM constructs remain valid even when new constructs are added in the model,
- *Minimality* for constructing the model as economically as possible but without limiting the scope of the ontology,
- *Alternative Views* and *Shortcuts* for enabling modelling flexibility,
- *Coverage* and *Granularity* allowing for "underdeveloped" concepts to increase compatibility while restricting hidden concepts to allow for extensions and
- *Extensions* to allow linkage of compatible external constructs that specialise on the model.

IS ontologies contain *classes* (defined as in the Object Oriented paradigm) and *properties* that define relationships between classes. 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*

like *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 stable version of CIDOC CRM is consisted from 90 classes and 148 properties (Crofts et al. 2009).

2.4.3.1 The English Heritage Extension CRM-EH

English Heritage (EH) is an organisation that has a major role in the dissemination of standards in cultural heritage domain, both at a national and international level. EH attempted an initial modelling exercise of the EH archaeological domain to the existing CIDOC CRM ontology (Cripps et al. 2004). After consultation with CIDOC CRM-SIG the modelling exercise concluded that an extension of CRM ontology to archaeological domain entities was necessary.

Due to the state of current archaeological systems, described as an "*archipelago of diverse, specialised and rather isolated and independent information systems and databases*" (Cripps et al. 2004), the adoption of an ontological framework of shared meanings seemed highly relevant for assisting cross-domain searching by researchers within and beyond the archaeological sector. However, the complexity and specificity required in representing the broader archaeological processes has led EH to the construction of the supplementary ontology (CRM-EH). The extended model CRM-EH, comprises 125 extension sub-classes and 4 extension sub-properties.

The CRM-EH model is based on the "single context recording" methodology, which is widespread in the UK and elsewhere, with origins in recording systems from the Museum of London and English Heritage (Richards and Hardman 2008). An archaeological context can refer to a section of wall, a post-hole, a cut of a ditch or a skeleton. In CRM-EH archaeological context is modelled as *Place*, "*extents in space, in particular on the surface of the earth, in the pure sense of physics: independent from temporal phenomena and matter*".

The model also provides relationships between archaeological contexts and archaeological finds which allows storing information that links directly to deposition and production events (of finds). Besides the archaeological notion of context, the CRM-EH ontology describes entities and relationships that relate to a series of archaeological events such as stratigraphic relationships and phasing information (i.e. relations between layers corresponding to different time periods), finds recording and environmental sampling.

Phasing events allow for a phasing hierarchy, which can lead to the definition of groups of contexts and sub-groups enabling post-excavation site analysis. CRM-EH follows an event-based, object-oriented modelling technique for extending CIDOC CRM in the archaeology domain that represents a much closer abstraction of the real world than previously traditional data-modelling approaches (Cripps et al. 2004).

2.5 Semantics

In linguistics, *Semantics* reflect the meaning that is encoded in language and conveyed through syntactic structures (Liddy 2003, Jurafsky and Martin 2000). Today most popular search engines operate at a keyword level, enabling users to satisfy their various information needs by simply submitting words in a search box without specifying any particular semantics of keywords. A comparison study over nine search engine transaction logs revealed that the average number of terms that web users employ to express a single query is 2.2 terms (Jansen 2006). Considering that only 2% of the users are using operators to explicate their query and that eight out of ten users simply ignore those results that are displayed beyond the first page, it is not hard to realise the importance individual words have in query formulation.

It is not possible for today's popular search engines to cope with human language ambiguities, hence we witness sometimes dubious efforts in the optimisation of on-line content in favour of rankings (Berners-Lee 2007). Adoption of non-recommended spam-like Search Engine Optimisation (SEO) techniques, which are often described as black-hat techniques, has led major popular search engines to respond with protective measures as in the case of Google's Florida update (Hargrave 2007). This has given rise to a new debate on the credibility of search engine ranking systems and the lack of current IS to deal with information on the semantic level.

2.5.1 Semantic Web

The Web is designed for human communication not for computer programmes to manipulate information (Berners-Lee et al. 2001). Computer programmes such as web browsers parse content on a layout level and routine processing of hyperlinks but are unable to understand the information displayed. The availability of limited and author-based information in the form of metadata tags is not adequate to enable processing of the page on a semantic level and to support some form of reasoning over the information. Such inability allows phenomena like "Google Bombing" to occur for a number of different

cases. In the case of "miserable failure" for example, the term has been massively targeted (bombed) by user-links which, resulted in the White House website biography page of the former president of the USA George W. Bush to be ranked first in the particular search engine for this specific phrase (BBC 2003).

The Semantic Web is proposed to add logic to the Web for improving user experience and information seeking activities and so to use rules, to make inferences and to choose courses of action that are defined by the meaning of information. It is not proposed to be an alternative or separate web but instead the Semantic Web is envisaged as an extension of the current Web. Supporting access to diverse and distributed collections of information, the Semantic Web can enable sharing of information and to provide a coherent view to resources, where for example 'zip code' and 'postal code' are defined as resources carrying the same type of information. In addition, dealing with *polysemy*; same word carrying different meaning in different contexts and *synonymy*, different words having the same meaning, can significantly improve user information seeking activities. It is said that the Semantic Web, when properly designed "can assist the evolution of human knowledge as a whole" (Berners-Lee et al. 2001).

The architecture of the Semantic Web is realised by the Semantic Web Stack which arranges layers of languages and technologies in a hierarchy. Each layer of the hierarchy uses the capabilities of the layers below, whereas the architecture still evolves as its layers are materialised. The Universal Resource Identifier (URI) and Unicode language are found at the very bottom of the hierarchy for supporting the necessary unique identification of all web resources and for all natural languages. The next layer in the hierarchy consists of the eXtensive Markup Language (XML) for enabling the definition of structured data and the Resource Definition Framework (RDF) for representing web resources as graphs.

A set of technologies (Web Ontology Language OWL and RDF schema) enable reasoning and facilitate knowledge sharing and reuse of semantic web information. The layer of information retrieval is satisfied by SPARQL, a specialised language for querying semantic web structures. The top layers of the hierarchy describe a set of technologies that are still in development and have not been standardised yet, such as Proof and Trust of the derived resources and User Interface to enable users to use semantic web applications (Berners-Lee 2007). The Semantic Web Stack is still evolving and periodically revised to include additional layers that support the semantic web technology.

2.5.2 Semantic Annotation

The term Semantic Annotation refers to specific metadata which are usually generated with respect to a given ontology and are aimed to automate identification of concepts and their relationships in documents (Uren et al. 2006). It is proposed that a mechanism responsible for connecting natural language and formal conceptual structures (a mediator technology between concepts and their worded representations) could enable new information access methods and enhance existing ones. These annotations enrich documents with semantic information, while enabling access and presentation on the basis of a conceptual structure, providing smooth traversal between unstructured text and ontologies.

Semantic Annotation can aid information retrieval tasks to make inferences from heterogeneous data sources by exploiting a given ontology and allowing users to search across textual resources for entities and relations instead of words (Bontcheva et al. 2006a). Ideally the users can search for the term "Paris" and a semantic annotation mechanism can relate the term with the abstract concept of "city" and also provide a link to the term "France" which relates to the abstract concept "country". In another case, employing a different ontological schema the same term "Paris" can be related with the concept of "mythical hero" linked with the city of "Troy" from Homer's Iliad.

Semantic Annotations carry the critical task of formally annotating textual parts with respect to ontological entities and relations. Such annotations have the potential to describe indices of semantic attributes which are capable of supporting information retrieval tasks with respect to a given ontology (Bontcheva et al. 2006b).

2.5.2.1 Classification of Semantic Annotation Platforms

Semantic Annotation platforms are classified as *automatic* or *manual* depending on their mode of operation (Uren et al. 2006). Manual systems enable assignment of user-defined semantic annotation of HTML, XML and text files with W3C standards formats. On the other hand, *automatic* systems are classified as *pattern-based* that employ specific pattern techniques such as seed-patterns or hand crafted rules for capturing known facts about annotations or *Machine-Learning* which make use of probabilistic models for generating annotations. *Machine-Learning* systems are distinguished between *supervised*, requiring a set of training data from the user in order to "learn" from and to provide annotations relevant to the training set and *unsupervised*, where annotations are produced through a bootstrapping, and iterative process with little or no intervention from the user.

The level of use of ontologies is another potential aspect of the classification of

semantic annotation tools. Both rule-based and machine-learning tools can use ontologies to enhance their operation and to describe the conceptual arrangements of semantic annotations. Usually such systems are described as *ontology bas*ed or *ontology oriented* depending on the level of ontology engagement (Li and Bontcheva 2007).

2.5.2.2 Examples of Semantic Annotation Platforms

There has been a considerable amount of effort dedicated over the last few years in the design and development of Semantic Annotation Platforms and Knowledge Management Systems capable of supporting semantic interoperable access to information. The plethora of semantic annotation platforms that are available today describes an active research and development field aimed at enabling semantic and interoperable access to information. A detailed description of all available tools and platform expands beyond the scope of this thesis. The following presents in brief a range of the most well-known and successful examples of semantic annotation tools as described in literature (Bontcheva et al. 2006a; Reeve and Han; 2005; Uren et al. 2006).

Amilcare is a popular Information Extraction tool that has been used in many different applications. The system uses a supervised Machine-Learning algorithm which enables adaptation in new domains for adding XML annotations in documents. Amilcare is used by **MnM**, a tool for annotating web pages with semantic data and by **S-CREAM**, a trainable adaptive tool that makes use of the **Onto-O-Mat** manual annotation tool for the definition of the trainable set and the **CREAM** framework for the creation of relational metadata of documents.

An example of unsupervised machine-learning approach is **Armadillo**, which achieves learning from a handful of user selected example seeds that are used in a bootstrapping process controlled by the user. **KnowItAll** also employs unsupervised machine learning techniques but without requiring any initial set of seed examples or user intervention. Instead the application uses for its learning the Pointwise Mutual Information (PMI) measure for calculating the ratio between search engine hits obtained for a discriminator phrase (e.g. “Paris is a city”) and search engines hits obtained with an extracted fact (e.g. “Paris”). **PANKOW** also makes use of the Web for exploiting a range of syntactic patterns which enable the system to automatically annotate instances from text with respect to an ontology of 58 concepts of tourism.

AeroDAML is an example of ontology oriented information extraction system which employs a pattern-based approach for annotating proper nouns and common relationships with respect to the DARPA Agent Markup Language (DAML). Similarly **SemTag**, uses

the TAP ontology, consisted of 65,000 instances, for performing a large scale semantic annotation based on lookup definitions which are disambiguated by a vector space model named Taxonomy-based Disambiguation algorithm (TBD).

The Knowledge and Information Management (**KIM)** is an ontology-based information extraction system which uses KIM Ontology (KIMO), an upper-level ontology consisting of 200,000 instances. The platform uses the GATE framework and pattern-matching rules (JAPE) for creating named-entity annotations that are used as metadata for supporting information retrieval tasks. To enable information retrieval the platform employs the SESAME repository of RDF triples and a modified version of the Lucene search engine for keyword-based search.

On demand annotation tools such as **Magpie**, **Open Calais Gnosis** and **KIM plug-in**, operate as web-browser add-ons that enable "real-time" annotation of web documents by associating strings to ontological concepts. In the case of Gnosis and KIM plug-in, concepts are associated to embed upper-level ontologies. Magpie on the other hand, is capable of operating with different ontologies, depending on user choice.

2.6 Semantic Projects of the Cultural Heritage Domain

Four distinct projects are discussed below which related to the cultural heritage domain. The selection of the projects relates to the broad Cultural Heritage focus of the thesis research effort in the provision of semantic indices of archaeological grey literature documents. The discussion primarily reveals background information on the STAR project which employs the semantic annotation result of the research effort. The thesis is associated with the **STAR** project which originates from Hypermedia Research Unit (University of Glamorgan) and contributes semantic annotation metadata of archaeological excavation and evaluation reports.

Secondarily the discussion reveals three additional projects from the Cultural Heritage domain that also relate to some extent to the research focus. In detail, the **Archaeotools** project is discussed due to its similarities with the STAR project in the provision of semantic access to archaeological datasets and documents. Two well-known projects of the digital Cultural and Heritage domain are also discussed. The **Perseus** project, established in 1985, presents a pioneer effort in the creation of an online digital library with emphasis on interoperability and the **Europeana** project which is a recent integrated effort at the European level for semantically linking digital objects of culture and heritage domain.

2.6.1 STAR Project

As briefly mentioned in the introductory chapter, 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, often referred to as 'Grey Literature'. The project aims to support the considerable 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 complex and 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, the STAR project has adopted the English Heritage extension of the CIDOC Conceptual Reference Model (CRM-EH). The CRM-EH ontology is necessary for expressing the semantics and the complexities of the relationships between data and textual elements, which underline semantically defined user queries.

The project has completed a data extraction, mapping and conversion to RDF process, facilitated by an interactive custom mapping and extraction utility. Five datasets have been included in the conversion task; producing a triple store of about 3 million RDF statements. Unique identifiers have been assigned to the RDF providing a consistent convention mechanism for unique naming of entities.

The STAR project also aims at the integration of knowledge resources such as vocabularies and thesauri for assisting new methods in accessing information. Knowledge resources in the form of terminology services can provide term look up, browsing and semantic concept expansion of terms by using semantic relationships and synonyms, to assist users express queries at different levels of generalisation and semantic perspective. The project has converted the English Heritage National Monuments Thesaurus and the MDA Object Thesaurus to standard Simple Knowledge Organisation System (SKOS) RDF format. The resulted SKOS Thesauri have been connected to the CIDOC CRM ontology and SKOS concepts have been mapped to CRM entities to form the relationship.

The project developed a CRM-EH based search demonstrator which cross searches over disparate datasets (Raunds Roman, Raunds Prehistoric, Museum of London, Silchester Roman and Stanwick sampling) and a subset of archaeological reports of the OASIS grey literature corpus (Tudhope et al. 2011). The Demonstrator makes use of the rich metadata for some forms of semantic search, building on CRM and SKOS unique

identifiers. Also the project delivered a set of web services for accessing the SKOS terminological references and relationships of the domain thesauri and glossaries which are employed by the project.

2.6.2 Archaeotools

The Archaetools project, led by the ADS and the NLP Research Group at the University of Sheffield, aimed at creating an advanced infrastructure for archaeology research (Jeffrey et al. 2009). The project adopted faceted classification and IE extraction techniques for 'unlocking' access to datasets and grey literature previously hidden from archaeology scholars. Following an ontology based approach, the project adopted four hierarchical ontological structures to describe concepts relating to the four facets of the classification; What, Where, When and Media.

The development of Archaeotools was based on previous experience from the Armadillo project, aimed at the extraction of data from historical court records, and the work in faceted browsing delivered by the ADS Archaeobrowser service. Both adaptive supervised and rule-based IE techniques were employed to serve particular extraction objectives aimed at the following concepts; Subject (what), Location (where), Temporal (when), Grid reference (where), Report title, Event dates and Bibliographic references. The project has implemented a faceted classification browsing system in the context of aggregated archaeological records, anticipated to replace the existing ArchSearch II.

2.6.3 The Perseus Project

Significant contribution to the digital library domain has been made by the Perseus Project (Smith, Ryberg-Cox and Crane 2000). The project was established in 1985 and since then has encoded several thousand documents of early Greek and Latin text, creating the on-line Perseus Digital Library. The ambitious mission of the project is highlighted under the aim 'to make the full record of humanity - linguistic sources, physical artefacts, historical spaces - as intellectually accessible as possible to every human being, regardless of linguistic or cultural background'.

Perseus is a digital library project focused on the encoding of thousands of documents using structured mark-up techniques based on SGML and most recently XML. The vast majority of the encoded documents are tagged according to the guidelines established by the Text Encoding Initiative (TEI). The project has developed a generalisable toolset for managing XML documents of varying DTDs (Document Type Definitions), capable of

extracting structural and descriptive metadata that support retrieval of document fragments and enabling further analysis for linguistic and conceptual features of documents. Over the years, Perseus has delivered digital content on a variety of platforms, from standalone CDROMs, to custom client/server software, to the World Wide Web.

2.6.4 Europeana

Europeana is a leading example of the shift of digital libraries towards semantic contextualisation. Described as a digital library, the project links more than 6 million digital items from the cultural and heritage domain (Gradmann 2010). There is no repository to store the million digital objects and none of the objects is stored in Europeana's data space. Hence, Europeana can be understood as a common ground, an aggregation mechanism for linking digital objects of culture and heritage domain. In this respect Europeana has been perceived by the public as being a portal, but indeed Europeana is more than that.

The project delivers a significant semantic enrichment to its linked digital objects via an Application Programme Interface (API) on which portal services can be built. Aiming to enable "complex semantic operations" on the linked resources that would not be possible to deliver by traditional digital library environments, Europeana employs a synaptic data model that brings together qualities from a set of well-established conceptual, terminological, and metadata models.

This Europeana Data Model (EDM) uses the Resource Definition Framework (RDF) technology to provide rich and interoperable descriptions of digital objects (Doerr et al. 2010). Based on the Open Archives Initiative (OAI) Object Reuse and Exchange (OAI-ORE) specification as structural modelling framework, EDM integrates the Simple Knowledge Organisation System (SKOS), the Dublin Core (DC) and the Friend-of-a-Friend (FOAF) models to provide its interoperable characteristics. The integrative approach of EDM equips Europeana with a flexible interoperability mechanism, allowing various communities of the culture and heritage domain to provide data while enabling the project to become part of the emerging Semantic Web paradigm shift.

2.7 NLP Tools and Frameworks

There is a plethora of available NLP tools and frameworks written in range of different computing languages and platforms and distributed by a range of proprietary and general public licences. Java based tools like the Open NLP (http://opennlp.sourceforge.net) and

Stanford NLP tools (http://nlp.stanford.edu) are described as statistical NLP tools based on maximum entropy models for delivering a range of NLP components, such as sentence detector, tokenizer, part of speech tagger etc. Such tools can be deployed standalone or can be combined into larger NLP frameworks for contributing to larger scale NLP applications. The detailed discussion of such NLP tools is not within the scope of this thesis, however the following paragraphs briefly discuss popular NLP frameworks which could be used to support the IE and semantic annotation aims of the thesis.

2.7.1 GATE

General Architecture for Text Engineering (GATE) is an NLP framework that provides the architecture and the development environment for developing and deploying natural language software components (Cunningham et al. 2002). The architecture distinguishes two basic kinds of resources; *Language Resources* and *Processing Resources*. Language Resources can be text documents, including a wide range of different formats (HTML, XML, Plain text, MS word , Open Office, RTF and PDF) while ontologies of OWL-Lite format and Lexicons such as WordNet are also regarded as Language resources. Text documents can be loaded individually as GATE documents or as a collection of documents described as a GATE corpus.

Processing Resources are NLP components that are made available by the architecture, such as Tokenizer, Part-of-Speech tagger, Sentence Splitter, as well as Gazetteers, Export modules specialised taggers etc. The architecture is equipped with a repository of Processing resources which contains a large variety of available resources known as Collection of Reusable Objects for Language Engineering (CREOLE) plug-ins. The architecture is flexible due to its open source orientation to integrate with a range of JAVA based Processing Resources which are made available via the CREOLE repository. The GATE community also delivers new plug-ins which support a wide range of NLP needs.

A collection of processing resources organised in a cascading processing order is known as the GATE pipeline or GATE Application. The architecture enables users to name and save applications which can be quickly reloaded into GATE with the associated Language and Processing resources. Offering a rich graphical user interface, the architecture also provides easy access to language, processing, and visual resources that help scientists and developers produce GATE applications.

The architecture supports a Lucene based searchable data-store and a Serial data-store for storing Language resources. In addition it includes ANNIE (A Nearly-NEW

Information Extraction System), a ready-to-run Information Extraction system. ANNIE consists of processing resources such as Tokenizer, Sentence Splitter, Part-of-Speech Tagger, Gazetteer and ANNIE Named Entity transducer for providing a fundamental and adaptable framework for Information Extraction. The ANNIE transducer utilises a set of rules in combination with available gazetteer listings in order to deliver the named entity result.

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

JAPE grammars are constituted from two parts; the LHS (Left Hand Side) which handles the regular expressions and the RHS (Right Hand Side) which manipulates the results of the matching conditions and defines the semantic annotation output. The architecture allows the integration of user-defined JAPE rules which are customised to extract information snippets to satisfy user specific IE goals.

2.7.2 UIMA

Unstructured Information Management Architecture (UIMA) (Ferrucci and Lally 2004) is a language processing framework aimed at analysing large amounts of text and other forms of unstructured information. The framework concentrates on performance and scalability with emphasis on standards. UIMA originates from IBM but it has now moved to be an open source project incubated by the Apache Software Foundation, while its technical specifications are developed by the Organisation for the Advancement of Structured Information Standards.

The architecture enables document processing applications (*Analysis Engines)* which encapsulate components (*annotators).* The UIMA components can be written in different programming languages, currently JAVA and C++, while the architecture allows installation of components from repositories. A standard data structure, the *Common Analysis System* (CAS) is operated by the Analysis Engines. CAS includes both text and annotation and supports interoperability by using the XML Metadata Interchange (XMI) standard.

An important architectural principle of UIMA is the use of strongly typed features for annotations and annotation features. Each Analysis Engine must declare what types of annotations are supported and must specify what feature each annotation type supports and what is the *type* feature each value may take, e.g. primitive, array, reference to another annotation type. The use of strongly typed features enables the architecture to control and check that output from one component has the right annotations types for input to the next component.

2.7.3 SProUT

Shallow Processing with Unification and Typed feature structures is a platform for the development of multilingual text processing systems (Drozdzynski et al. 2004). The platform is not as popular as GATE and UIMA but it has been adopted as the core IE component in several EU-funded and industrial projects, mainly originating from Germany and Poland. SProUT is developed by the German Research Centre for Artificial Intelligence (DFKI) and while not open source it can be used for research purposes free of charge. The motivations supporting the SProUT development relate to the trade-off between processing efficiency and expressiveness of grammar rules.

The platform utilises unification-based grammar formalisms which are designed to capture fine-grain syntactic and semantic details. Such formalisms use as their informational domain a system based on features and values. The main characteristic of the platform is that allows use of rich descriptive rules over linguistic structures which enable information sharing in the form of features among rule elements.

2.7.4 The Adopted Framework

All three frameworks that are discussed above have their merits but also their weak points. SProUT enables the definition of fine-grained rules but it is not a popular platform with limited availability of documentation and community support. UIMA on the other hand, might be a robust and scalable framework but the strongly typed features approach does not lend easily to prototype, exploratory or rapid developments. In addition, it has a steep learning curve since it relies on the Eclipse integrated development environment (IDE) for GUI support and on third party NLP tools for delivering language processing tasks. GATE might support rapid development via the ready-to-run ANNIE system and a unified GUI environment which controls all aspects of the development (language, processing and data-store resources), however performance and scalability are not its strongest points.

Considering the merits and weak points of the above framework, this research study adopts GATE as the core IE platform of the project. In detail, GATE supports rapid prototype and exploratory developments allowing use of loosely typed annotation types and features while making available a fast range of NLP plug-ins, including ontology and terminology (gazetteers) components. Thus, it fits well to the exploratory nature of the project and the requirement to deliver semantic annotation with respect to ontologies using terminological resources. In addition, the PhD work, being a research and not a commercial project, does not present any significant performance requirements. Thus GATE is suitable to negotiate the volume of grey literature documents since processing time is not top priority.

Furthermore, the platform has been in development for more than 10 years and has matured while used in range of projects. It is also supported by a strong community and available online documentation (tutorials, user forums, mailing lists etc). Regarding training, the GATE team organises annual summer schools which support developers to obtain new skills and discuss issues of their applications. The author has participated in two GATE summer schools, 2009 and 2010, which have significantly helped to improve skills and to develop the final Semantic Annotation application.

2.8 Corpus and Terminological Resources

2.8.1 OASIS Grey Literature

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.

Dissemination of grey literature in archaeology is a well-recognised problem (Falkingham 2005). The developer-funded archaeology in England has delivered a large volume of unpublished reports, which despite high quality and potential interest enjoy a limited distribution. 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 contain 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 thesis and in particular 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 archaeology research groups, institutions, and organizations, coordinated by the Archaeology Data Service (ADS), University of York, aiming to provide a unified online index to archaeological grey literature and a means by which the index can be maintained (Richards and Hardman 2008).

2.8.2 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, hence it is built upon 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 semantics 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 (hypernym) concepts. In addition, mapping relationships such as skos:exactMatch and skos:closeMatch, enable linking between concepts of different concept schemes according to varying degrees of match.

2.8.3 Terminological Resources

English Heritage 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 (English Heritage 2007) and EH National Monuments thesauri (English Heritage 2006) were previously converted from their original format (recording manuals and relational databases) to controlled terminology Simple Knowledge Organization System (SKOS) resources (Binding, Tudhope and May 2008).

The terminology resources adopted by the prototype are; the Simple Names for Deposits and Cuts glossary, which provides a controlled vocabulary for recording archaeological context; 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 contains both basic archaeological contexts (e.g. cut) and broader semantic groupings of basic contexts (e.g. ditch).

2.9 Summary

The current chapter has discussed a range of topics relating to the aims of the PhD research effort. The discussion revealed the origins of NLP and its potential to advance Information Retrieval practices in dealing with language ambiguities. The use of IE for achieving the NLP potential is addressed together with the role of Semantic Annotation in delivering concept aware metadata. The contribution of ontologies was also explained in fulfilling the aims of semantic aware applications, while a range of semantic efforts of the Digital Heritage domain was also discussed. The discussion concluded with the adopted Language Engineering framework employed to support the IE and Semantic Annotation efforts of the PhD project. The following chapters discuss the development effort of PhD project commencing with the Pilot System Development and Evaluation.

Chapter 3

Prototype Development and Evaluation

3.1 Introduction

This chapter discusses a prototype development and evaluation of an early Information Extraction system aimed at delivering semantic annotation metadata. The prototype stage is part of a larger project, investigating the use of NLP techniques in combination with Knowledge Organization Systems (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 (OASIS grey-literature), which can be exploited further by retrieval applications, such as STAR. The KOS employed by the prototype, are the CIDOC CRM ontology (Crofts et al. 2009) and the CRM-EH extension for archaeology (Cripps et al, 2004), together with the terminological resources English Heritage Recording Manual and English Heritage National Monuments Thesauri (English Heritage 2006; English Heritage 2007).

The prototype also investigates the capacity of 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. In detail, the prototype development explores the flexibility of GATE for modification and adaptation to a chosen semantic annotation task. The adaptation concerns the ability of JAPE rules to target concepts on ontological and terminological levels, using gazetteer listings containing entries that have unique terminological references. In addition, this chapter discusses an evaluation of the overall performance of the prototype information extraction system aimed to inform a full scale evaluation exercise discussed in chapter 8, following established evaluation measurements for assessing the performance of semantic annotation systems

The prototype is not targeted at delivering semantic indices of grey literature documents in the form of RDF triples; this is something that is addressed at later stages by the full scale semantic annotation system. However, the prototype delivers semantic annotations in XML interoperable format capable of further use and manipulation by external applications. The first part of the discussion introduces background information

and literature, while the development phases are discussed in the main body of the chapter. The last part discusses the evaluation phase and the accommodation of semantic annotation by the Andronikos web portal.

3.2 Background to Prototype Development

The semantic annotation output is targeted at supporting the STAR project for interoperability and semantic discovery of archaeological information, in this case grey literature. In order to achieve its aims, the prototype adopts the CIDOC CRM and its extension CRM-EH ontology while utilising a range of English Heritage terminological resources (glossaries and thesauri). The prototype is developed in the language engineering framework GATE and processes a corpus of archaeological excavation and evaluation reports originating from the OASIS grey literature library at the Archaeology Data Service.

3.2.1 Ontology-Based versus Thesauri-Based Semantic Annotation

As discussed in Chapter 2 (section 2.5.2), ontologies can be employed by semantic annotation systems to provide specialised vocabulary and relationships that are exploited during IE. Examples of such systems include h-TechSight (Maynard et al. 2005), which delivers semantic annotation via an ontology that consists of 9 main concepts (Location, Organization, Sector, Job Title, Salary, Expertise, Person and Skill); populated with a vocabulary of 29000 instances. The ontology is used to enable the task of Named Entity Recognition (NER) over job advertisements. Similarly, KIM (Kiryakov et al. 2004) uses KIMO, an ontology populated with a vocabulary of general purpose classes, such as Location, City, Currency, Date, Job Title etc. KIM is a comprehensive semantic application that utilises KIMO beyond NER in order to support document retrieval on semantic level.

Ontology-based IE projects such KIM and h-TechSight make use of ontologies that explicitly define classes and their properties. Classes and sub-classes form hierarchies which combine terminological and ontological specialisation. Bontcheva et al. (2004) describes a technique of ontology engagement in IE using the GATE OWLIM-Lite processing resource. The tool associates ontological classes with one or more vocabulary listings (gazetteers). Lists contain entries which populate ontological classes with instances, for example the class *Location* can be associated with the list *Cities* containing the entries *London, Paris, Athens* etc.

However, according to Tsujii and Ananiadou (2005) the tendency of an ontology-based approach to make explicit semantic associations between vocabulary and individual context is problematic. They argue that contextual dependencies strongly influence the IE process: "*... relationships among concepts as well as the concepts themselves remain implicit in text, waiting to be discovered*". Thus, inherited language ambiguity and diversity, as well as domain-dependent inferences and knowledge cannot be comprehensively encoded in ontological structures. Instead, they argue that terminological thesauri, as language oriented structures, can support implicit definition of semantics.

In particular, they highlight the case of the Biomedicine domain, which 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, where inevitably language is ambiguous.

> *"Most of the widely used ontologies have been built on a top-down manner. They are limited in their conceptual coverage and they are mainly oriented for human (expert) use. The difficulties and limitations lie with the definition of concepts (classes, sets of instances) since one is expected to identify all instances of a concept. This task demands evidence from text."* (Tsujii and Ananiadou 2005).

In some applications, the matching of instances with ontology classes may be less problematic, where language use is constrained or perhaps highly specialised.

The archaeology domain, however, shares some of the context-dependency discussed above. As in the Biomedicine domain, context-independent relationships as explicitly defined in logical ontologies are not the norm. Contextual factors dictate if, for example, a particular place is an archaeological "context", or if a physical object constitutes an archaeological "find". Such forms of entity specialisation cannot be inferred solely by a specialised vocabulary but are derived by contextual evidence. Therefore, complementary use of terminological and ontological resources may prove a promising avenue of investigation.

3.2.2 Development Pathway Criteria

The prototype development has an experimental focus aimed at obtaining practical experience and results to inform the large scale semantic annotation effort of this thesis as discussed in Chapters 4, 5, and 6. The prototype aims to explore an innovative semantic annotation process which does not rely on the use of a single ontology, as typical Ontology Based IE, but instead makes a complementary usage of ontological and terminological resources (Figure 3.1).

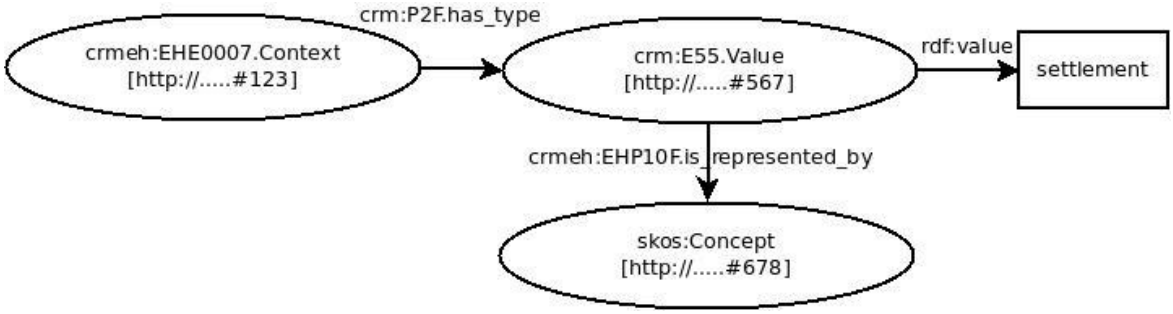


Figure 3.1: Semantic Annotation with terminological and ontological reference. The textual value of the annotation is "settlement" which is defined as an instance of the CRM-EH class EHE0007.Context. The value is linked (is_represented_by) a SKOS definition.

The reason for following this particular development pathway is based on the following criteria;

Semantic annotation (via archaeologically specific CRM-EH entities) cannot be reached by using only specialised vocabulary, as discussed in section 3.2.3. The archaeology domain vocabulary does not contain heavily specialised scientific terms and CRM-EH specialisation is subject to contextual dependencies.

- The CRM and CRM-EH ontologies have no directly associated vocabularies but define a range of entities and properties which provide semantic definitions and clarifications for the cultural heritage (CRM) and archaeology domains (CRM-EH)
- The semantic annotation effort is targeted at delivering semantic indices which will support information retrieval at the level of concepts. Thus, the prototype system is not concerned with the annotation of unique instances i.e. *post-hole A*, *post-hole B*, but with the annotation of concepts, i.e. the concept of *post-hole* not individual post-hole occurrences. However a concept may have term variants (e.g. *post hole).*
- Using both ontological and terminological resources empowers semantic annotations with a dual conceptual reference system that enables information retrieval on the ontological level, on the terminological level and the combination of both.

3.3 Information Extraction Pipelines of the Prototype

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. 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 examined and 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 3.1).

3.3.1 Pre-processing Phase

The pre-processing phase (Figure 3.2) 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 pipeline.

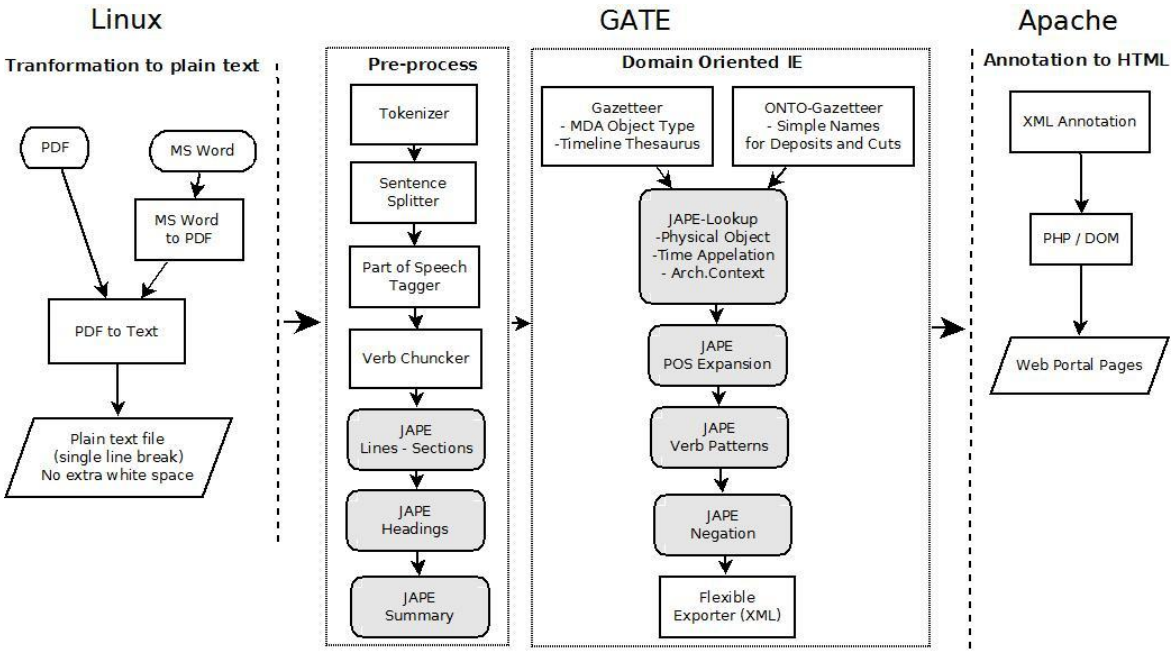


Figure 3.2: Phases of the Semantic Annotation process from File Transformation to GATE pipeline and to final conversion of annotations. Pilot study Grammars in shaded boxes.

As already discussed in section 3.2.3 contextual-dependency influences the validity of semantic annotation in archaeological text. Hence, the inclusion of document sections such as heading and table of contents (TOC) in the semantic annotation effort can lead to the delivery of annotations that enjoy a limited contextual-dependency. Headings and TOC might make use of EH vocabulary, however such sections do not use terms in a rich discussion setting, i.e. within argumentation, but instead use terms in isolation as in titles.

Additionally, detection of headings also supports extraction of document sections, such as summary sections, which contain rich discussion worth revealing.

Documents also contain tabular data and figures which make use of domain related terminology. However, such sections usually make use of domain vocabulary either as field labels or as descriptions of diagrams and figures. Extracting semantic information from such structures requires adoption of an alternative information extraction strategy since tabular areas normally do not contain free text but data. Thus, information extraction from tabular areas is unlike free text extraction and governed by its own specific semantic context. The priority of this study is the extraction of semantic information from discussion areas (free text) not from tabular sections. For this reason tabular sections are identified by the pre-processing pipeline as "Noise" and are excluded from the ontology oriented semantic annotation processes.

3.3.1.1 Transformation to Plain Text

The grey literature corpus consists of 2460 archaeological reports originating from the OASIS digital library. GATE allows processing of a range of different document formats such as Adobe PDF (.pdf), Microsoft Word (.doc), and plain text (.txt). GATE does not maintain the morphological aspects of imported documents, such as font size and type framework but only maintains tags of HTML and XML documents and stores them in a specialised annotation set called "Original mark-up". The OASIS corpus contains both Microsoft Word and Adobe PDF files. All Microsoft Word documents were initially converted to PDF files, enabling transformation of all the documents to plain text files by a single application. It was necessary to transform the corpus documents into plain text files outside GATE. Import of PDF or Microsoft Word files into GATE documents causes documents to lose their morphological arrangements. Tabular sections turn to single word lines and necessary line breaks are introduced between paragraphs, headings and table of contents.

Specific pre-processing JAPE rules exploit single carriage return definition for identifying document sections and headings. It is important to avoid generation of blank areas and multiple line breaks during the import process which can harm the performance of such rules. For the transformation of files to plain text the Linux shell application "*pdftotext*" is used. The application allows transformation of files to a raw dump format that suppresses page breaks and blank areas; while it enforced specific text encoding (Latin 1) and single carriage return. File transformation enables the construction of JAPE rules for detecting document sections by exploiting the single carriage return definition of text files.

3.3.1.2 Annotating Document Sections

The first step of the pre-processing pipeline identifies sentences containing verbs or verb phrases. The detection of verb and verbs phrases within sentences is considered to be an indication that the discussion uses some form of argumentation. Hence, sentences that contain verbs most possibly do not use terms in isolation, as for example headings and tabular sections, but instead make use of vocabulary within a richer discussion setting. In order to detect sentences that contain verbs, the pipeline uses the ANNIE sentence splitter with *single new line* configuration. The configuration delivers sentence annotations on splits (i.e. full-stops) and on new line definitions (carriage return); for example the line "*evidence, and stratigraphic evidence is either non-existent or unclear. However with the*" delivers two different sentence annotations.

JAPE grammars exploit the single carriage return definition of the documents to identify the beginning and end token of each line (JAPE Grammar 1). For document tokenization, the pipeline uses the default ANNIE tokenizer. The verb input annotation is delivered by the ANNIE Part of Speech (POS) tagger and the Verb chunker processing resource.

Sentences containing a verb are annotated as "*Rich_Sentence*" (JAPE Grammar 2), however non-verb sentences between "*Rich_Sentence*" annotations are also turned into "*Rich_Sentence*" (JAPE Grammar 3). This enables inclusion of both incomplete sentences introduced by the Sentence Splitter configuration and of sentences that do not contain a verb but consist of noun phrases that are part of a richer discussion.

All "*Rich_Sentence*" annotations are then joined together with a simple pattern rule to identify a *Section* annotation. *Section* annotations, due to the verb-based sentence selection, enable delivery of semantic annotation from discussion document areas. The semantic annotation pipeline is targeted at delivering semantic annotations only from document areas that are annotated as *Section*. The aim is to avoid annotation from tabular, partially-worded or heading areas which might make use of EH vocabulary in a context-independent manner.

JAPE Grammar 1

"*Match the beginning of each line*"

```
({SpaceToken.kind==control})+
({SpaceToken.kind==space})*
({Token}):Begin
```

The grammar annotates as *Begin* the first Token after one or more(+) line breaks followed by zero or more spaces (*).

"Match the end of each line"

```
({Token}):End
({SpaceToken.kind==space})*
({SpaceToken.kind==control})+
```

The grammar annotates as *End* the first Token before zero or more spaces (*) followed by one or more (+) line breaks.

JAPE Grammar 2

"Match Sentence containing at least one verb"

```
{Sentence contains VG}
```

The grammar matches a *Sentence* (annotated by ANNIE Sentence Splitter) that contains at least one verb (*VG*). *Sentence* is re-annotated to *Rich_Sentence (pseudo-line).*

(JAPE Grammar 3)

"Match Sentence between Rich_Sentence"

```
{Rich_Sentence}
({Sentence}):match
{Rich_Sentence}
```

The grammar matches a *Sentence* wrapped between two *Rich_Sentence* annotations.

The next stage of the pipeline identifies single worded line sections (Noise). The explicit identification of such sections prevents their annotation as headings. Empirically, headings do not contain words written only in lower case and normally the length of single worded heading descriptions is greater than three letters. Typically words of two or three letters are articles and determiners which cannot stand on their own as heading descriptions. Therefore, strong cases of "*Noise*" annotations target single worded lines, which contain words that have less than four letters length and single worded lines containing a lowercase word (JAPE Grammar 4). Other "*Noise*" cases are lines that contain only numbers and symbols with no words, usually originating from tabular areas (JAPE Grammar 5).

(JAPE Grammar 4)

"Single-worded Line has length less than 4 letters"

```
{Token.length < 4, BL contains EL}
```

"Single-worded Line is in lowercase"

```
{Token.orth == "lowercase", BL contains EL}
```

The BL (Beginning of Line) contains EL (End of Line) is used for matching only single-worded Lines. The Token annotations results from the ANNIE Tokenizer.

(JAPE Grammar 5)

"Lines containing numbers and symbols and no words"

```
{BL, Token.kind == number}
({Token.kind != word})[0,10]
{EL, Token.kind != word}
```

The grammar matches Lines stretching up to 12 Tokens that do not contain any Tokens of the kind *word*.

The identification of heading spans is based on a collection of eight different pattern-matching rules. Two rules annotate heading areas 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 for example "*3.1 Prehistoric phase*". Such heading cases enable definition of simple and precise rules, capable of annotating headings having numerical prefix conventions (JAPE Grammar 6).

Another set of rules is targeted at annotating single worded headings having upper initial or capitalised case that do not commence with a numerical prefix, for example "*Introduction*". In such cases, rules require that a *Section* annotation or another *Heading* annotation follows the single worded phrase. Similarly multi-worded phrases in upper initial or capitalised case must be followed by a *Section* annotation or a *Heading* annotation, in order to qualify as *Heading* annotations. A specific set of rules targets heading cases that are followed by a sequence of dots. Such cases are frequently found in table of contents (TOC). The identification of TOC is based on a simple pattern that joins four or more previously identified *Heading* annotations together (JAPE Grammar 7).

Annotation of *Summary* sections is based on a JAPE rule which annotates a 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 while the second *Heading* annotation is simply the next available *Heading* annotation of the document (JAPE Grammar 8).

(JAPE Grammar 6)

"Annotate Headings that commence with a numerical prefix"

```
({BL, Token.kind == number, Token.length <= 2}
(({Token.string == "."})?
({SpaceToken.kind == space})?
({Token.kind == number, Token.length <= 2})?)*
({SpaceToken.kind == space})+
({Token.orth !="lowercase", Token.kind == word})
({Token.kind == word}|{Token.kind ==number}|
{Token.kind == punctuation}|
{SpaceToken.kind == space}|{Dots})*
{EL}):match
```

The grammar matches headings of numerical commencement. The rule matches phrases which commence with numbers like 1, 1. , 1.1, 1.1.1. etc. followed by a non-lowercase word Token, which is then followed from any number of Tokens including sequence of Dots (previously identified) until the end of line EL token.

(JAPE Grammar 7)

"Annotate TOC by joining previously identified Headings"

```
{Heading}
{Heading}
{Heading}
({Heading})+
```

The grammar annotates TOC by matching four or more Headings in a row which are required in order to avoid annotation of succeeding Headings within document (empirically two to three) which are not TOC.

(JAPE Grammar 8)

"Re-Annotate Headings as Summary"

```
{Heading contains Lookup.type =="Summary"}
```

Heading containing any Lookup of the type Summary (this kind of Lookup originates from a gazetteer which contains the terms summary, abstract and overview).

"Annotate Summary Sections"

```
{Heading.type=="Summary"}
{Heading}
```

GATE enables annotation of large chunks by configuring the input of annotation types that are processed by a JAPE rule. Introducing to the rule only the necessary annotation types, we control which types are transparent and processable by the rule. Thus, we manage to annotate large chunks of text using simple rules that by-pass annotation types which are found within the text chunk but are not visible by the rule. The rule matches a section between a *Heading* of type *Summary* and the next available *Heading* annotation.

3.3.2 Domain Oriented Information Extraction Phase

The domain-oriented pipeline (Figure 3.2) 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 use case scenarios and project discussions with EH, specifically with the project collaborator Keith May. The use case scenarios describe simple and complex information seeking activities targeted at range of CRM-EH entities. Simple scenarios require the retrieval of single entities, such as archaeological contexts, finds, samples and activities, whereas more complex scenarios are targeted at retrieving entities, their attributes and their relationships with other entities, such as archaeological contexts containing finds or samples, stratigraphic information and grouping definitions. A detailed list of all available use case scenarios can be found in [Appendix D1]. After discussion and consideration of available use case scenarios it was decided that the prototype system should focus on the extraction of the following concepts:

- E19 Physical Object defined as *"items having physical boundaries that separate them completely in an objective way from other objects",* such as arrowhead, bone, pot, etc.
- E49 Time Appellation defined *as "appellation 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.
- E53 Place with emphasis on EHE0007.Context defined 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.

The thesis adopts a composite definition of Context, to include broader interpretive groupings. In the CRM-EH, *EHE0005.Group* is defined as "*The process of grouping together the various places that represent contexts into interpretive spatial entities*". After consultation with archaeologists it was decided that the IE definition of Context should include both entities (EHE0005 and EHE0007). Distinction between individual contexts and groups cannot be addressed by specialised vocabulary while grey literature reports tend to reflect a higher level of generality in reporting excavations which does not distinguishes the individual contexts from groups. Thus, the thesis adopts the CRM-EH *EHE0007.Context* entity to model both individual contexts and groups. The adoption concerns the delivery of semantic annotation definitions and thesis reporting terminology.

3.3.2.1 Domain Specific Knowledge Based Resources

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 glossary Major and Minor (SKOS Type) Features respectively associated gazetteer entries with both an ontological (CRM) class definition and with a *skos:concept* terminological reference (one of the 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 GATE resource OWLIM to represent the CRM-EH ontology. In practice, the immediate outcome was equivalent. However, since the aim was to annotate at the concept (rather than individual) level, it was decided that the first approach was more appropriate for our purposes than making an explicit instance connection between an ontology class and archaeological terminology, based on the concerns with context dependency discussed earlier. The incorporation of thesauri into GATE gazetteers was a practical solution without requiring the development of a new GATE CREOLE module. It would have been inappropriate to represent thesauri as a formal OWL ontology within GATE as 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. In the future, if an

appropriate GATE language resource were available, that might be another option.

3.3.2.2 JAPE Grammars of the Prototype Pipeline

The prototype pipeline implemented fifteen different JAPE grammars for identifying the three main ontological concepts (Physical Object, Context, Time Appellation). JAPE grammars are employed for matching gazetteer terms in text (JAPE Grammar 9). 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 uses the CRM-EH class property, made available via the OWLIM plug-in (JAPE Grammar 10). Additional rules are used for extending the initial Lookup annotations to include meaningful moderators which are identified by the Part of Speech module (JAPE Grammar 11). For example, initial Lookups are expanded to include adjectives, adverbs and passive voice verbs that precede them. Especially in the case of Time Appellation Lookup, two gazetteer listings are used for expanding over prefix terms (Earlier, Later, etc.) and suffix terms (Period, Century, etc.). These gazetteers listings originated from the ANNIE system and were modified based on the use of dictionaries to accommodate additional time related terms relevant to the IE task. A full list of the part of speech Hepple tagger categories that are used by the rules can be found in Appendix F1.

(JAPE Grammar 9)

"Annotate Lookup via MajorType"

```
{Lookup.majorType == "Physical Object"}
```

The grammar matches Lookup of majorType "Physical Object".

(JAPE Grammar 10)

"Annotate Lookup via Ontological Reference"

```
{Lookup.class == "EHE0007.Context"}
```

The grammar matches Lookup of ontological class "Context".

(JAPE Grammar 11)

"Extend initial Lookup to include moderators"

```
({Token.category== VBN}{Context})|
({Token.category== JJ}{Context})|
({Token.category== CD}{Context})
```

The grammar extends initial Context Lookup towards passive voice verb (VBN), adjectives (JJ) and cardinal numbers (CD).

(JAPE Grammar 12)

"Extend initial Time Appellation Lookup to include prefix and suffix terms"

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

The 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".

Moreover, 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" (JAPE Grammar 13). 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> (JAPE Grammar 14). Such patterns are aimed at matching relations between Time Appellation and Physical Object, as for example "...*coins dating to Roman period*...", and Time Appellation and Archaeological Context as for example "...*pits are of prehistoric date*...". The above pattern-matching approach is aimed at supporting the required contextual-dependency of annotations as discussed previously in section 3.2.3.

Simple JAPE grammars are also used by the pipeline for matching negation in phrases (JAPE Grammar 15). 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".

(JAPE Grammar 13)

"Annotate Lookup pairs"

```
{TimeAppellation}{PhysicalObject}
```

The grammar matches pairs of Lookup annotation i.e. "Roman Coin".

(JAPE Grammar 14)

"Annotate Lookup connected in phrases with verbs"

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

The grammar matches phrases that connect Lookup annotation via verbs phrases i.e. "pits containing pottery".

(JAPE Grammar 15)

"Annotate Simple Negation Phrases"

```
{Lookup.majorType== "Negation"}
({Token.kind == word})[0,10]
({PhysicalObject}|{TimeAppellation}|{Context})
```

The grammar matches negation phrases followed by an offset of maximum 10 word tokens and Lookup annotations i.e. "No evidence of pottery".

3.4 Evaluation of the Pilot System

The evaluation task aimed at measuring the performance of the prototype information extraction mechanism with regards to the concepts of Time Appellation, Physical Object and Archaeological Context and the relations of Time Appellation with Physical Object and Time Appellation with Archaeological Context. As discussed in Chapter 2 (section 2.3.1) and Chapter 8 (section 8.3), the effectiveness of Information Extraction systems can be measured by Recall, Precision and F-measure rates. The performance of the prototype extraction mechanism was evaluated against the above measurements. For the purposes of the evaluation, a manually annotated version of the intended IE results was created and made available to the GATE Corpus Benchmarking Utility.

The task had a largely investigative character, aiming not just to evaluate the performance of the prototype system but also to suggest the necessary development improvements that have 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 3.2. The results and experience gained from the prototype evaluation study informed a full scale evaluation of the final system which is fully discussed in Chapter 8 (Evaluation).

Instructions to annotators targeted the task of manual annotation at the concepts of Physical Object, Place (Archaeological Context), and Time Appellation advising for a flexible and rich annotation approach [Appendix D2]. The purpose of the guidelines was to direct annotators to identify phrases carrying a rich meaning with regards to the targeted concepts, hiding any algorithmic and pattern matching clues that could influence and bias their performance. The manual annotation task followed an end-user oriented approach allowing flexibility in annotation and inclusion of modifiers and rich phrases.

For the pilot evaluation purposes, the annotations of the senior archaeologist (KM) were treated as the GS for the evaluation, since the other annotators did not have the same level of specialist domain knowledge. However, table 3.1presents the system's performance against all four manual annotation sets. The aim of the evaluation was to reveal the potential of the prototype system rather than concluding to a definite benchmarking evaluation. 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. Thus, the annotation input of the senior archaeologist was selected as the most appropriate for delivering indicative performance results that correspond to the investigating focus of the study

3.4.1 Evaluation Results

The annotators used MS Word highlight and underline tools to identify (mark) the annotations in text. The manual annotations were then exactly replicated in the GATE environment using the available Ontology Annotation Tool (OAT). The four versions of manual annotations were processed by the GATE Corpus benchmark utility and produced the following overall scores (Table 3.1). AV is the developer and was involved in the construction of pattern matching rule, having a clear understanding about coverage of gazetteers and system functionality. Thus his annotation is closer to system's capacity, delivering a high overall score. On the other hand, KM is the archaeology expert and his input challenges the system's performance to a desirable user-centred result. CB and DT are project members who have a limited knowledge on gazetteer coverage and pattern matching rules. Their annotations are not as challenging as KM annotations but are valuable from an average user, non-archaeology expert point view.

	AV	**CB**	**DT**	**KM**
Precision	0.85	0.68	0.73	0.51
Recall	0.85	0.69	0.61	0.69
F-measure	0.85	0.68	0.66	0.59

Table 3.1: Prototype system performance

A close examination of the IAA results (Table 3.2) of the four annotators reveals a low agreement score. This appears typical of manual annotation in an archaeological context. Zhang, Chapman and Ciravegna (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-measure agreement score for all four annotators is 51%, whereas the agreement score between different pairs varies from 35% to 65%.

	Precision	Recall	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

Table 3.2: Inter-Annotator agreement score of the different pairs

The lowest agreement score is between AV-KM where AV is the system developer and KM an archaeologist while the highest score is between CB-DT where both are STAR project members. To some extent, the low agreement between annotators reflects an 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 precise vocabulary employed was specified exactly. 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 were 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. 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.

The prototype system performs well against Time Appellation entities delivering F_1 score 81% while it delivers good *Precision* for Context entities 70% and for Context plus Time relations 75% (Table 3.3). 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.

	Precision	**Recall**	**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

Table 3.3: System's performance for three ontological entities (Context, Physical Object and Time Appellation) and for two relations (Context + Time and Physical Object + Time)

3.4.2 Discussion on Evaluation Results

Although, results are not yet 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 3.3) which is considered encouraging as a basis for further elaboration by the full-scale system, as discussed below. For example, full scale semantic annotation systems targeted at archaeological context have yielded F_1 score of 75% (Zhang, Chapman and Ciravegna 2010) while full scale systems targeted at historical text have delivered 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. Additionally, 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 e.g. *Building*. Additional validation techniques such as word pair disambiguation can be invoked to improve precision, while negation detection might 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 sophisticated pattern-matching grammars based on likely syntactical constructs, in order to improve the performance of relation extraction.

3.5 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 objective of Andronikos web-portal (http://andronikos.kyklos.co.uk) 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 3.3) 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. The open source search engine indexing algorithm FDSE is also deployed in the portal to index the web-pages of the semantic annotations and the full text version pages.

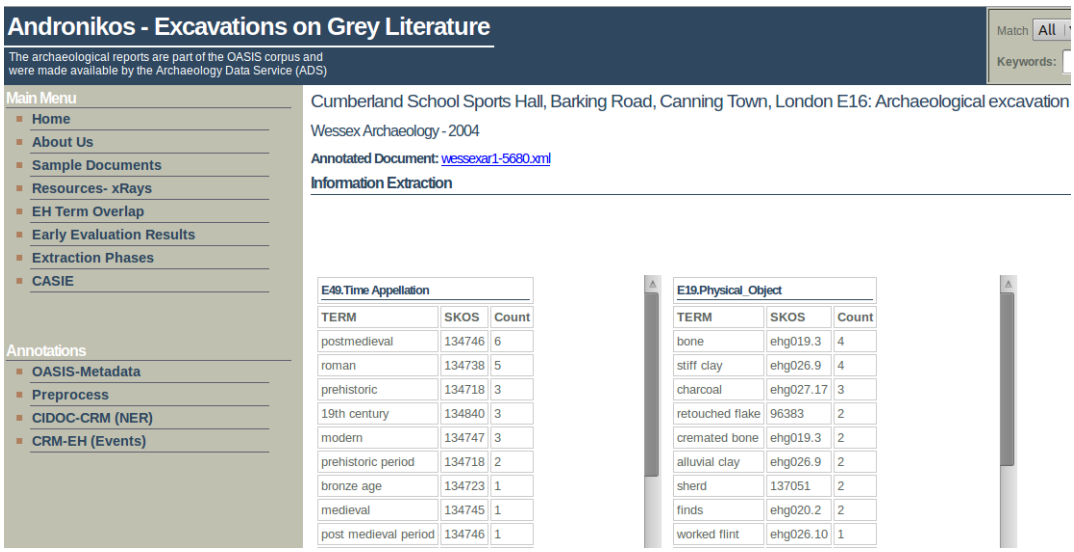


Figure 3.3: Andronikos Web portal. Tables show the textual instance value, number of occurrences in document and the associated SKOS value (postmedieval and post medieval period share the same SKOS reference)

3.6 Summary

The results reported in this chapter 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 prototype 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 inference.

The initial evaluation results are not at an operational level. However they suggest the methods have potential when improved further by 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 prototype development has reached its aims for the implementation of a prototype semantic annotation system capable of extracting concepts from archaeological grey literature with respect to domain ontology and terminological resources. The prototype explored an innovative method for the complementary use of such resources, capable of delivering semantic annotations which enjoy both an ontological and SKOS terminological reference. In addition, GATE proved to be adequate to undertake the task of semantic annotation for a large set of grey literature documents. The framework has been negotiable in modification of its resources while JAPE grammars proved to be flexible and robust for expressing grammars targeted at the extraction of CRM and CRM-EH entities and relations.

Overall, the prototype study suggests that further elaboration of the method by a full-scale system is feasible. The following chapters discuss the further improvement of the method aimed at exploiting the IE potential in the provision of rich semantic indices with respect to CIDOC CRM and CRM-EH ontologies.

Chapter 4

System Preparation and Adaptations

4.1 Introduction

The chapter discusses the preparation phase of the full-scale IE system aimed at delivering semantic annotation of archaeological (grey literature) documents. The full-scale system aims to explore further the NLP techniques of the prototype development, in order to deliver semantic indices capable of supporting document retrieval at the semantic level. The process of semantic indexing is dived into two main phases (Figure 4.1), the semantic annotation phase and the document indices production phase. The first phase is conducted in GATE and is divided into three sub-phases; Pre-processing, Named Entity Recognition (NER) with respect to CIDOC CRM, and CRM-EH specialisation phase. The second phase transforms the semantic annotation output to Resource Description Framework (RDF) triples, capable of supporting document retrieval with respect to ontological and terminological semantics.

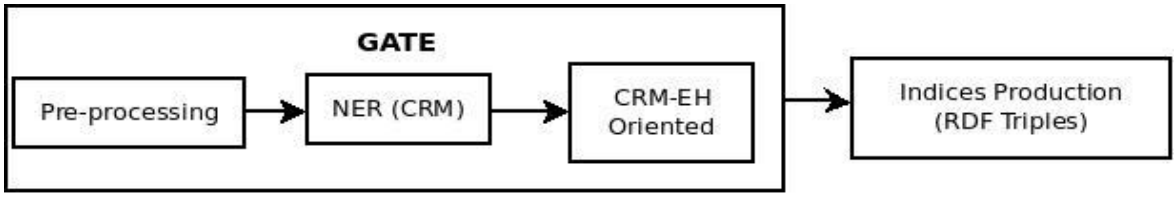


Figure 4.1: Birdseye view of the process of semantic indexing

The chapter provides background information with regards to the task of NER. The discussion reveals the role of NER in the process of semantic annotation with respect to ontological entities while discussing relevant projects that dealt with NER in the cultural heritage domain. The central role of terminological resource in the process of NER is also discussed. The chapter reveals the terminological resources that are employed by the full-scale system and discusses their preparation and transformation to GATE gazetteers. A detailed analysis of the terminological resources is given followed by the enhancement process of gazetteers. The chapter concludes with a discussion on the modifications of the prototype Pre-processing pipeline, which is employed by the first phase of the full-scale semantic annotation system.

4.2 Named Entity Recognition (NER)

The term Named Entity Recognition (NER), also sometimes referred to as Named Entity Recognition and Classification (NERC), is a particular subtask of Information Extraction aimed at the recognition and classification of units of information to predefined categories, such as names of person, location, organisation, expressions of time, money, percentage etc. (Nadeau & Sekine 2007). The term was first coined during the sixth Machine Understanding Conference (MUC 6), which focused on the extraction of structured information of company and defence related activities from unstructured text (Grishman & Sundheim 1996).

Nadeau and Sekine (2007) presented a study that examined the progress of NER systems within a period of fifteen years, from 1991 to 2006. They revealed a large number of projects targeted at the task of NER covering a range of language, textual genre and domain factors. According to the study, a great deal of research has been devoted to the study of NER for the English language, while progress has also been made in a wide range of other languages ranging from major European languages to Arabic, Hindi and Korean.

Nadeau's study also revealed projects beyond the business and defence domain, including systems that addressed the task of NER in domains, such as sport, gardening and humanities. In addition, different systems were targeted at different textual genres; including newspaper articles, scientific journals, emails, informal documents and religious texts. With few exemptions, most systems were targeted at a single domain, extracting a specific set of entities. The study concluded that almost any domain and text genre can be supported but that the portability of NER systems to different domains and textual inputs always presents a major challenge.

4.2.1 Named Entity Recognition Schools of Thought

There are two main development roots in the design of NER systems, namely Rule-based and Machine Learning, which correlate to the schools of thought in Information Extraction, as discussed in section (2.3.2).

4.2.1.1 Rule Based NER

The Rule-based approach of NER is based on the definition of hand-crafted rules that exploit a range of lexical and syntactical attributes for the identification of the extraction results. Lexical attributes describe word level features, such as word case (lower-upper), morphological features (prefix, suffix, stems) and digits that are used in the identification

of numerical entities. Hand-crafted rules can also exploit part of speech attributes that are assigned to words by NLP Part of Speech (POS) modules. Part of speech attributes enhance the definition of rich syntactical patterns, which are employed by the NER process.

Rules can invoke input from gazetteers, lexicons, dictionaries and thesauri to support the purposes of NER. Such word classification systems contain specific terms of predefined groups, such as person names, organisation names, week days, months etc., which can be made available to the hand-crafted rules. Mikheev et al. (1999) describe a Rule-based system, where NER does not entirely rely on the use of very large gazetteers but instead exploits what he calls, *internal* (phrasal) and *external* (contextual) evidence. The system without any use of gazetteers scored around 80% for names of organisation and people, although it did not perform that well for locations. The same system using gazetteers (4900 place names, 30000 companies and 10000 first names) scored around 93%, which according to Mikheev is indicative of the contribution of word lists in the NER task.

4.2.1.2 Machine Learning NER

The second school of thought employs Machine Learning methods for the identification of extraction entities. Such methods can support supervised, semi-supervised or unsupervised learning. The basis of supervised learning is a training set, which provides positive and negative examples of named entities over a collection of annotated document. A machine learning technique (Hidden Markov Model, Maximum Entropy, or Support Vector Machine) exploits the training set to automatically induce rules that are responsible for the identification of named entities. The quality of the final result heavily relies upon the quality of the training set, and since the construction of a large and good quality training set is a laborious process, the use of hand-crafted rules remains the preferred technique when a training set is not available (Nadeau & Sekine 2007).

Semi-supervised or unsupervised techniques aim to compensate for the lack of a training set in a machine learning environment. However, the performance of such techniques is not always comparable with the results of supervised learning or rule-based techniques (Uren et al. 2006; Wilks and Brewster 2009). A semi-supervised approach does not require the annotation of a large training set. Instead it uses a handful of "seeds" (examples) to learn from. The system then tries to identify common contextual clues from the given examples and uses such clues to identify new examples and new contextual clues, iterating through a bootstrapping process. Unsupervised learning techniques do not

require the provision of training examples, instead they use input from statistics or lexical resources, such as WordNet, as the basis for their learning.

4.2.2 Related NER Projects

Since 1996, the year MUC6 was held, the task of NER has gained increased interest (Nadeau and Sekine 2007). However, it is beyond the purposes of the current thesis to provide a comprehensive overview of all such NER developments. The following paragraphs discuss a number of related NER projects, selected according to criteria concerning the domain of application or the method of development. Hence, the selected projects either use a Rule-based NER technique, or address the problem of NER in the culture and heritage domain. The objective of the following discussion is to reveal successful design choices and techniques, which can be employed from the current study to address the issue of NER with respect to the thesis aims and objectives.

4.2.2.1 GATE Related NER

The MUSE system used GATE in order to identify the three main MUC6 entity types (ENAMEX name entities, NUMEX numerical entities, TIMEX time entities) and two other project specific types (Address and Identifier). The project examined three different textual genres; religious monologues, scientific books and medical emails (Maynard et al 2001). The system delivered Precision and Recall results ranging from 63% to 93%. However, a minimal adaptation of the system was necessary in order to deliver robust results for all three different types of textual input. The system used a range of gazetteer inputs and hand-crafted rules to address the task of NER.

The knowledge management platform h-TechSight (Maynard et al 2005) is another project that used GATE and rule-based techniques to answer the NER task. The project used ontologies instead of flat gazetteers to support the recognition task. h-TechSight used an ontology consisting of 9 main concepts (Location, Organisation, Sector, Job Title, Salary, Expertise, Person and Skill) and populated with 29000 instances. The system processed 38 documents of job advertisements delivering Precision 97% and Recall 91.5%.

Similarly to h-TechSight, the KIM system uses GATE and Rule-based techniques. However, KIM is a comprehensive semantic application that expands beyond the limits of NER, supporting document retrieval on a semantic level. KIM is equipped with KIMO an upper level ontology containing about 250 entity classes, 100 attributes and relations and about 200,000 instances including 50,000 locations, 282 countries, 4,700 cities which makes KIM a good basis for location-based services (Popov et al. 2004) .

4.2.2.2 NER in the Culture and Heritage Domain

NER in the cultural heritage domain is evident by a range of different projects. 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. The system used in-house XML tools (LT-XML2 and LT-TTT2) to perform the NER task, delivering Precision and Recall scores ranging from 65.19% to 81.81%.

Byrne (2007) focused on NER from historical archive texts, originating from the Royal Commission on the Ancient and Historical Monuments of Scotland (RCAHMS). Employing a corpus of 1,546 text documents, the NER task employed 11 classes; Organisation, Person-name, Role, Site-type, Artefact, Place, Sitename, Address, Period, Date, and Event. The project adopted a machine learning approach based on the CandC maximum entropy classifier, using the MMAX2 annotation tool to provide annotations for the training of the machine learning algorithm.

The NER task dealt with the identification of nested entities in phrases such as "Aberdeen School of Architecture" assigning "Aberdeen" a *Place* classification. The system delivered Precision and Recall results ranging from 66% to 87%. It also made limited use of gazetteers. According to the study, use of gazetteers produced no improvement, due to the machine learning approach adopted, which tends to deliver good results without gazetteers if the training set is sufficiently representative.

The Archaeotools project, led by the ADS and the NLP Research Group at the University of Sheffield, aimed at creating an infrastructure for archaeology research to enable to archaeologists to "*discover, share, and analyse datasets and legacy publications*" (Jeffrey et al. 2009). The project aims are comparable to some extent with the aims of the STAR project since both projects concentrate on opening access to archaeological information on a semantic level. Hence, the methods and techniques employed by Archaeotools for addressing the task NER have a particular interest for the thesis.

Archaeotools adopted a faceted classification and NLP techniques for providing access to datasets and grey literature previously hidden from archaeology scholars. The project adopted 4 hierarchical ontological structures to describe concepts relating to the 4 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).

Each facet, apart from *Media*, was populated from existing Thesauri, i.e. *What* - Monument Types Thesaurus, *Where* – UK government list of administrative areas, *When* – Monument Inventory Data Standard (MIDAS) period list.

For the NER task, the project adopted a hybrid approach, incorporating both machine learning and rule-based, 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.

Machine learning techniques proved to be laborious and expensive to build in the project due to the time consuming constrains of the training set definition (Jeffrey et al. 2009). The NER output was exploited by the faceted classification browser in order to support discovery in an information retrieval context. However, the use of extracted entities as metadata supporting document indexing was not straightforward. Documents usually refer to a number of different periods, object and places depending on the progress and features of an archaeological fieldwork. Distinguishing the significant (for the document) entities from those incidentally identified proved to be a challenging issue. To address this issue the project applied a frequency threshold, taking the top 10% of named entities from each document (Bateman and Jeffrey 2009).

4.3 NER Terminology Resources and Ontological Entities

The prototype development for the PhD research revealed the capability of hand-crafted rules and gazetteers to deliver named entity output associated with CRM ontological entities using GATE. However, evaluation results of the prototype system suggested that a full scale NER system should not make blind use of all available knowledge resources but should instead aim to make optimum use of gazetteer resources and contextual evidence, in order to improve both Precision and Recall rates. Thus, a full-scale system should be able improve NER Recall rates without significantly harming Precision.

4.3.1 Selecting Ontological Entities for NER

It is a common practice of faceted classification and NER systems aiming to enable access to cultural heritage information to focus on entities, which relate to Places, Periods, Persons and Physical Objects (Byrne 2007, Grover et al 2008, Jeffrey et al 2009). Archaeotools employed a faceted classification, which clustered individual entities under 4 main groups; What, Where, When and Media. It seems that this form of classification is

useful for enabling access to archaeological information. It is a classification that was followed, except the Media facet, by the prototype system that was targeted at the NER of Time Appellation (When), Place (Where) and Physical Objects (What).

The focus of the prototype-system and subsequently of the main, full-scale system is complementary to the focus of Archaeotools project. The full NER system is focused on the generation of detailed, lower level, "rich indexing", in relation to the domain ontology CRM and CRM-EH, whereas the Archaeotools NER did not consider this particular level of detail of the domain ontologies and neither did it utilise the ontological definition of CIDOC CRM and CRM-EH.

The task of more archaeologically specific CRM-EH specialisation is not addressed by the NER phase since implementation of the CRM-EH specialisation possibly requires use of contextual evidence in addition to the specialised vocabulary. Thus, the NER (CRM) phase of the IE system is targeted at identifying named entities with respect to the general CIDOC CRM ontology. A second phase (CRM-EH Oriented) of the system, discussed in a later chapter, identifies the required contextual evidence for extracting information with respect to the CRM-EH specialisation.

Co-reference is another subtask of Information Extraction introduced during MUC6 aimed at providing resolution between proper nouns and the pronouns and the aliases used to refer to them. Similarly to the prototype system, the full-scale NER task is not targeted at extracting entities of individual instances of Person, Place names etc. Thus Co-reference resolution is not within the scope of the development.

The full-scale NER task is focused on the extraction of the following entities.

1) **Physical Object** defined from CIDOC CRM documentation (Doer 2003) as : "*items of a material nature that are units for documentation and have physical boundaries that separate them completely in an objective way from other objects*"
2) **Place** defined as: "*space, in particular on the surface of the earth, in the pure sense of physics: independent from temporal phenomena and matter*"
3) **Time Appellation** defined as: "*all forms of names or codes, such as historical periods, and dates*"
4) **Material** as the concept of materials " to denote properties of matter before its use, during its use, and as incorporated in an object, such as ultramarine powder, tempera paste, reinforced concrete"

The first three entities had already been targeted by the prototype system. Material is the only entity not previously addressed. Based on the evaluation results of the prototype

and discussions with English Heritage collaborators, it was decided to include Material in the NER task due to the strong relation materials have with physical objects and the importance materials have in the science of archaeology. Following the first letter(s) of the above four entities the NER pipeline is given the acronym OPTIMA (**O**bject, **P**lace, **TI**me, **MA**terial), which is used throughout the thesis to refer to the large-scale NER pipeline.

4.3.2 Terminology Resources for NER

From the range of the available terminology resources which were made available to the STAR project by the English Heritage, a subset was selected to support the NER task. Following a closer examination of the available resources and liaising with English Heritage (Keith May and Phil Carlisle), 4 thesauri and 5 glossaries are identified as most appropriate for supporting the task of NER. The selection of the resources is based on the criteria of suitability and specialisation. Hence, only the resources that are suitable to support the classification task of Physical Object, Place, Time Appellation and Material entity types are selected.

Following the Archaeotools example, the selected resources are initially associated with the ontological entities on the basis of single specialisation, meaning that each resource is specialised to support classification of a single entity type. A main difference with the Archaeotools project is that the Monument Types Thesaurus is associated with the *Place* entity, thus being closer to the *Where* facet instead of the *What* facet, which is the choice of Archaeotools implementation. Monuments are addressed by the PhD study as *Places* not as *Physical Objects,* following the CRM-EH definition of archaeological context as a CRM Place entity. Examples of Physical Objects are the different types of archaeological finds.

The CRM-EH ontology defines spatial elements that constitute an individual archaeological unit of excavation as *EHE0007.Context*. Grouping of individual contexts is also permitted and defined as *EHE0005.Group*. Both group and context entities are a specialisation of the CRM *E53.Place* entity which is immobile and consumes space. The thesis adopts the above definition and treats monuments as *Places* not as *Physical Objects*. However, at later stages (CRM-EH oriented phase) the ontology distinction between primitive contexts and larger groupings of contexts is not adopted. As discussed in section 3.3.2, due to the higher level of generality of grey literature reports, the distinction between primitive contexts and larger groupings is not feasible. Following the prototype development choice, OPTIMA treats both primitive contexts, such as cuts and deposits,

and also larger groupings, such as well, floor and structures, as archaeological contexts assigned to the EHE0007.Context class. The final mapping between terminological resources and ontological classes is shown below in table 4.1.

Ontological Entities	**Terminology resources**
Physical Object	• Box Index Form: Find Type (Glossary) • Small Finds Form (Glossary) • MDA Archaeological Object Type (Thesaurus)
Place	• Simple Names for Deposits and Cuts (Glossary) • Monument Types (Thesaurus)
Time Appellation	• Timeline (Thesaurus)
Material	• Box Index Form: Material (Glossary) • Bulk Finds Material List (Glossary) • Main Building Materials (Thesaurus)

Table 4.1: Mapping between Ontological Entities and Terminology resources

4.3.2.1 Glossaries

The following Recording Manual glossaries (English Heritage 2007) relate to the recording practices of the English Heritage and are adopted to support the NER task:

Simple Names for Deposits and Cuts: The glossary originates from a recording manual aimed at providing a control vocabulary for recording archaeological context, which generally are considered types of primitive contexts and groups. The glossary consists of 96 terms of Cuts and Deposits including possible variations, such as Hearth pit, Hearth pit: fill, Hearth pit: debris.

Box Index Form (Material):

Boxing relates to the process of archaeological fieldwork and is vital to the management of finds for keeping track both of individual finds and boxes of finds. The glossary consists of 20 terms regarding the material(s) from which the objects are made.

Box Index Form (Find Type):

Similar to the above, this glossary contains terms that reflect the broad method by which finds have been recorded. The glossary consists of only 4 terms (Bulk Find, Small Find, Skeleton, and Sample)

Small Finds Form: The glossary is a controlled vocabulary for recording all three-dimensional small items on site, such as coins, most metals, worked bone, etc. It consists of 27 terms describing small finds.

Bulk Finds Material List: The glossary is a controlled vocabulary for recording bulk

finds, such as pottery, ceramic building materials, etc. The glossary consists of 18 terms.

4.3.2.2 Thesauri

Four EH thesauri (English Heritage 2007) were selected to support the NER task:

Monument Types: The resource is a thesaurus of monument type records arranged by function. It includes types of monuments relating to the built and buried heritage in England. It contains 6,567 terms classified under 18 hierarchies: agriculture and subsistence, civil, recreational, religious ritual and funerary, transport, etc.).

MDA Archaeological Object Type: The resource contains physical evidence that can be recovered from archaeological fieldwork, such as objects and environmental remains which are portable and cover all historical periods. The thesaurus contains 2,231 terms classified under 25 hierarchies, such as agriculture and subsistence, animal equipment, architecture, armour and weapons, etc.

Main Building Materials: The resource contains constructional material types of monuments relating to the built and buried heritage. It contains 626 terms classified under 9 hierarchies, such as animal, earth mix, manmade material, etc.

Timeline: An experimental thesaurus for dates and periods. It is made available by the EH to support the aims of the STAR project. It contains 582 terms classified under the following hierarchies: artistic period, cultural period, geological period, historic period, political period and religious period

4.3.3 Study of Terminology Resources Overlap

A one-to-one association between terminology resources and ontological entities is not always a straight forward option. In the case of the STAR project, the availability and range of terminology resources does not enable a simplified association between resources and entities. Apart from the straight forward case of the Timeline thesaurus, which can be associated with the Time Appellation entity, the rest of terminological resources require a thorough analysis before they can be associated with ontological classes.

The following sections present the result of a study focused on the investigation of terminological resources and their level of overlap. The outcome of the study informs the construction of GATE gazetteers that are used by the NER pipeline.

4.3.3.1 Term Overlap Tool

A bespoke tool was built to support the study of overlaps, aimed at revealing the overlapping terms between the various knowledge resources (Figure 4.2). Deploying the

tool as a web application served the purpose of making it easily available to the members of the STAR project. The tool is available from the Andronikos web-portal (access currently restricted to members of the STAR project, since the knowledge resources are owned by other organisations).

The web deployment runs on an Apache 2.0 web-server. The tool uses the selected terminological resources, which are accommodated in a MySQL database server. The user interface was developed using PHP and HTML technology. The tool is built to support the research aims of the thesis and is not aiming to support end-user tasks. The tool supports identification of overlaps between thesauri, glossaries and user defined terms. In detail, the tool operates in the following two modes.

The first mode of search supports identification of term overlap between the terminology resources. The tool enables checking for potential overlaps between thesauri, between glossaries and between glossaries and thesauri without restriction on the number of the examined resources. For example the user can check for overlaps between two, three or all available thesauri, depending on user selection. Similarly, a user can check for overlaps between any number of glossaries or between any number of glossaries and thesauri.

For example, find all overlapping terms between the thesauri Monument Types, MDA Object Type and the glossary Simple Names for Deposits and Cuts. The tool returns matches in the following form:

```
->MATCH:"WALL" 70426, 1 WITH: "wall" ehg003#93, ehg003
->MATCH:"WALL" 96129, 128 WITH: "wall" ehg003#93, ehg003
```

The first line of the above result translates as "WALL" with terminological reference "70426" of the Monument Type Thesaurus (1) overlaps with the terms "wall" with terminological reference "ehg003#93" of the glossary Simple Names for Deposits and Cuts (ehg003). The second line is a about a match of the same term ("wall") between the MDA Object Type Thesaurus and the glossary.

Notice that each terminological resource provides its own unique reference. At this point a manual examination of the scope of resources is required in order to identify if the terms originating from the different resources correspond to the same concept or if they are polysemous terms. If they correspond to the same concept, the *skosExactMatch* property is used to link the different references, as described in section 4.4.2

Also the tool supports *wild-card* matches searching for partial term matching. For example, the above search including the *wildcard* option returns 6 results (Figure 4.3), including "Wheel" and "wheel rut" not previously identified. Using the *wild-card* option allows a thorough examination of the resources but distorts the semantic equivalence of matches. Partial matching at the string level may encourage ambiguity, which requires more rigorous intellectual reviewing of overlapping terms than exact match results, before resuming to a conceptual link via the *skosExactMatch* property.

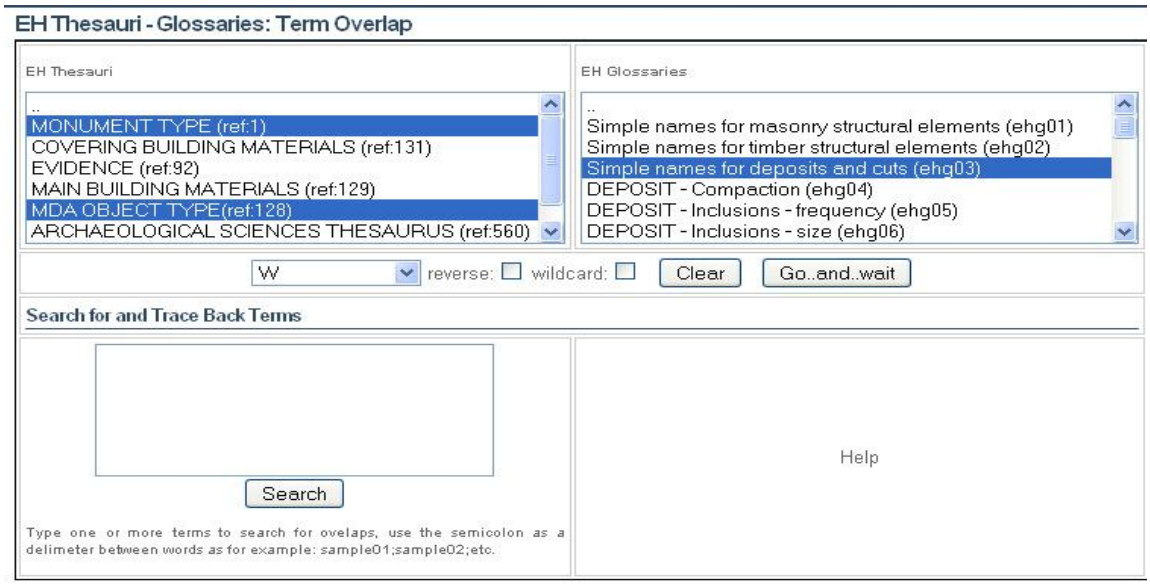


Figure 4.2: The search mode of the term-overlapping tool, displaying results from terminology resources for the letter "W"

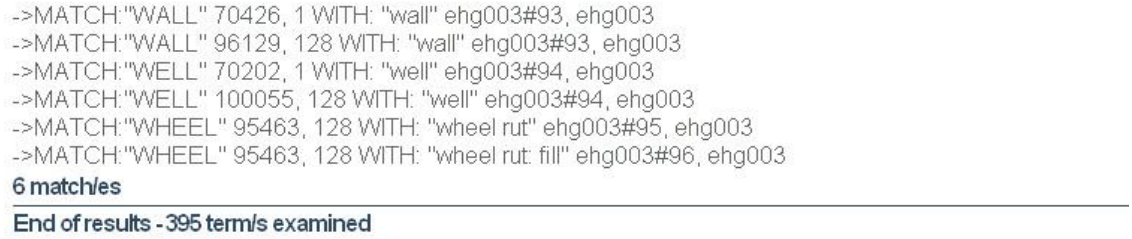
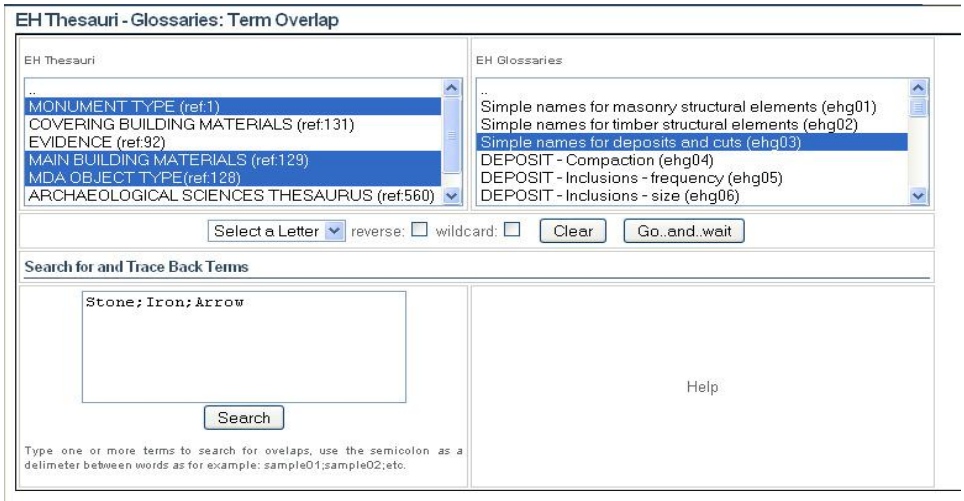


Figure 4.3: Results of overlapping terms between 3 different resources for the letter "W" using the wildcard option

The second mode of search supports overlap matching between user-defined terms and terminology resources. Users can check if a term or a list of terms separated by semicolon exists in any of the selected resources glossaries and thesauri. For example the user can check whether the terms "Stone", "Iron" and "Arrow" are found in any of the selected thesauri and glossaries (Figure 4.4). The tool returns results in a tabular format including references to their broader and top terms. This functionality is very useful for finding overlapping points between thesauri and glossaries, which are used as entry points for the semantic expansion techniques (section 5.3) employed by the NER task.

The results (Figure 4.4) reveal that the term "Stone" having terminological reference "70391" is found in the Monument Type thesaurus. The term has a broader term "Natural Feature" having terminological reference "70383" and top term "Monument" having terminological reference "102872". The same term ("Stone") is also found in the Main Building Materials thesaurus and in the glossary Box Index Form: Material (ehg019). Similarly the term "Iron" overlaps in the Main Building Materials thesaurus and in two glossaries, Box Index Form: Material and Small Find Form, while the term "Arrow" is only found in the MDA Object Type thesaurus.



Term	TermSKOS	Broader Term	BT SKOS	TOP Term	TT SKOS	Origin
STONE	70391	NATURAL FEATURE	70383	MONUMENT	102872	MONUMENT TYPE
STONE	98216			STONE	70391	MAIN BUILDING MATERIALS
Stone	ehg019#20	-N/A-	-N/A-	-N/A-	-N/A-	
IRON	97992	METAL	98053	METAL	98053	MAIN BUILDING MATERIALS
Iron	ehg019#11	-N/A-	-N/A-	-N/A-	-N/A-	
Iron	ehg026#15	-N/A-	-N/A-	-N/A-	-N/A-	
ARROW	95131	PROJECTILE	100011	ARMOUR AND WEAPONS	97083	MDA OBJECT TYPE

Figure 4.4: The search mode of the bespoke term overlap tool displaying tabular results of 3 overlapping terms in 4 different terminology resources.

4.3.3.2 Term Overlap Analysis

The bespoke tool was used to collect a range of data regarding term overlap between resources. Each terminological resource was examined for overlaps against the rest of the resources and the results were recorded in a 9x9 matrix (Table 4.2). Examining the matrix, some immediate readings can be made. For example, the Timeline thesaurus does not overlap with any other resource. Thus, the specific thesaurus can be treated as a "pure" Time Appellation resource capable of supporting NER of Time Appellation entities without conflicting with other entity types. On the other hand, the Timeline thesaurus is the only terminological resource that does not have any overlaps. All other terminological resources present overlaps, some to a larger degree than others.

	T1	**T128**	**T129**	**T486**	**G03**	**G19**	**G20**	**G26**	**G27**
T1	-	110	4	0	25	2	1	1	1
T128	110	-	18	0	11	3	1	3	6
T129	4	18	-	0	0	11	0	10	10
T486	0	0	0	-	0	0	0	0	0
G03	25	11	0	0		0	0	0	0
G19	2	3	11	0	0	-	0	12	13
G20	1	1	0	0	0	0	-	0	0
G26	1	3	10	0	0	12	0	-	5
G27	1	6	10	0	0	13	0		-

Table 4.2: Number of overlaps between knowledge resources. Key: T1(Monument Types), T128(MDA Object Types), T129(Main Building Materials), T486(Timeline Thesaurus), G03(Simple Names for Deposit and Cuts), G19(Box Index Form: Material) G20(Box Index Form: Find Type), G26(Small Finds Form), G27(Bulk Finds Material List)

To support the analysis of overlaps between the resources, a diagram is arranged (Figure 4.5) based on the associations between terminology resources and ontological entities. The diagram groups together the resources under the four ontological entities targeted by the NER pipeline. Dotted lines are drawn that are accompanied by numbers to indicate the volume of overlapping terms between the resources. The plethora of lines is indicative of the number of overlaps, while the diagram depicts a situation where apart from the Timeline thesaurus all other resources have overlaps.

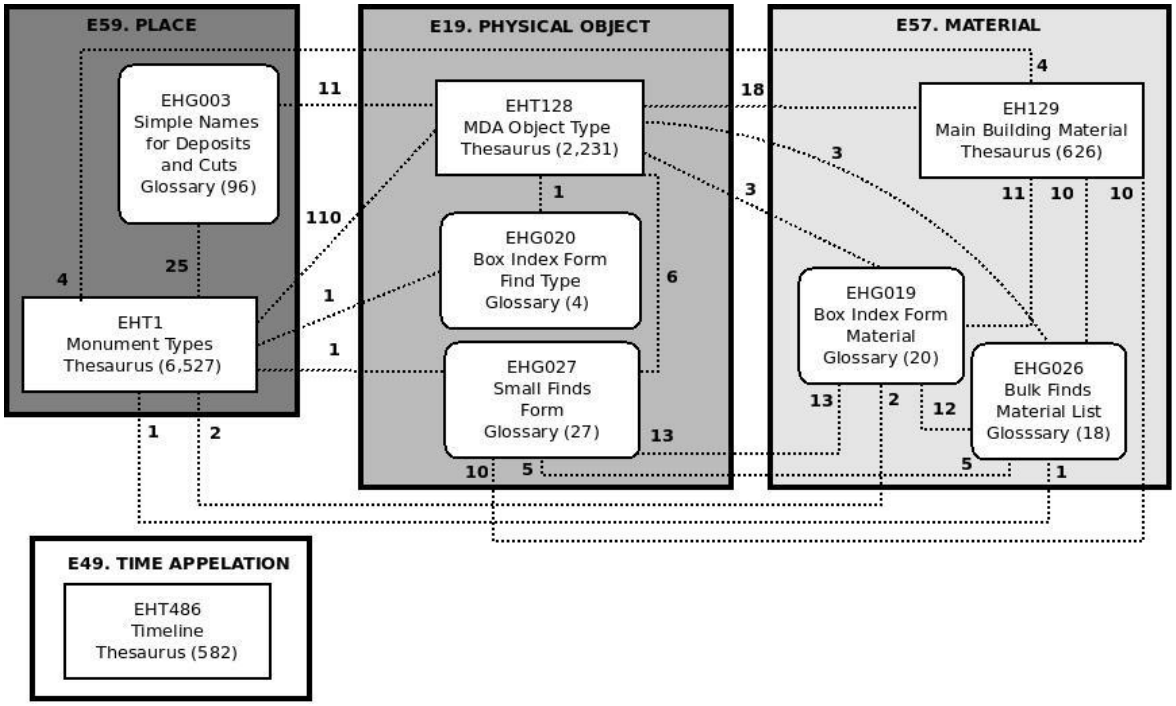


Figure 4.5: Diagram of overlapping terms between terminology resources.
The connecting dotted lines show the volume of overlapping terms between thesauri.

The Monument Types thesaurus as shown in table 4.1, is associated with the Place ontological entity. However, the resource presents a large number of 110 overlaps with the MDA Object Types thesaurus, which is associated with the Physical Object entity. Examining closer the range of overlapping terms, it is revealed that the vast majority of the 110 overlaps belong to the "Architecture" facet of the MDA Object Types thesaurus. The facet contains terms like "Floor", "Tomb", "Hearth" etc. Such terms are treated by the CRM-EH ontology as immobile places, which potentially can describe archaeological contexts and groupings .Thus, the "Architecture" facet of the MDA Object Types thesaurus is excluded from identifying *Physical Objec*t entities and it is aligned to deliver *Place* entities.

The Monument Types thesaurus also overlaps with the Simple Names for Deposits and Cuts glossary by 25 terms. The particular glossary is also associated with the *Place* entity and so overlaps do not conflict with the ontological alignment of the resource. On the other hand, the same glossary overlaps in 11 terms with the MDA Object Types thesaurus. The overlapping case is exactly the same as the case of the 110 overlapping terms (Architecture facet) described above.

The Monument Types thesaurus also presents some minor overlaps with the Main Building Materials thesaurus (4), the Box Index Form: Material (2) the Small Finds

Form(1), the Box Index Form: Find Type(1) and the Bulk Finds Material List (1). The overlaps concern the terms, "Stone", "Wood", "Mosaic", "Sarsen Stone" and "Skeleton"..

In the Monument Types thesaurus, the sense of "Skeleton" corresponds to "Inhumation" and the sense of "Wood" to "Woodland". Since the terms are only used as synonyms in these thesauri, they can be discarded to avoid confusion. Skeleton is an ambiguous case between an object/place sense affected by archaeology domain constrains. The IE system (OPTIMA) addresses *Physical Object<>Material* ambiguity but does not address *Physical Object<>Place* ambiguity. Therefore, "Inhumation" is treated as *Place* but "Skeleton" is always treated as a *Physical Object*. In addition, the terms "Stone" and "Mosaic", are treated either as *Material* or as *Physical Object,* with disambiguation techniques (section 5.6) defining their ontological alignment but they are never used to deliver *Place* entities.

Additionally to the overlaps discussed above, the MDA Object Types thesaurus presents a number of overlaps with *Material* aligned resources. The Object Thesaurus overlaps with the Main Building Material thesaurus by 18 terms, with the Box Index Form: Material glossary by 3 terms and with the Bulk Finds Material List glossary by 3 terms. The overlaps concern terms like "Brick", "Marble", "Paper", "Textile" etc., which have dual sense (i.e. they can be described as Material or as Physical Object terms). Similar overlaps (10 terms) occur between the Main Building Materials thesaurus and the Small Finds Form glossary. The overlap concerns again dual sense terms, such as "Glass", "Tile" and "Wood".

The above trend of overlaps is also evident between glossary resources associated with either Material or Physical Object entities. For example, there is an extensive overlap between the Small Finds Form glossary aligned with Physical Object and the Box Index Form: Material glossary aligned with Material. The overlap covers more than half of the glossary terms, 13 out of 20 terms overlap. A full list of the overlapping terms between Physical Object and Material aligned resources is available from the [Appendix A1].

Disambiguation techniques are employed by the NER phase to resolve when possible the appropriate sense of *Material* and *Physical Objec*t overlapping terms. The disambiguation techniques are targeted only at cases of *Physical Object<>Material* ambiguity, not at *Physical Object<>Place* ambiguity. Whenever, disambiguation is not achieved the overlapped terms maintain both senses. Disambiguation techniques, Negation Detection and Noun-phrase validation issues are discussed in more detail in chapter 5.

4.4 Compilation of Terminology Resources as Gazetteers

GATE gazetteers consist of a set of lists containing lexical entries like names, locations, measurements units, periods, etc. The lists can be invoked to support the task of NER by finding occurrences of gazetteer entries in text and delivering Lookup annotations. A gazetteer definition file is responsible for coordinating the cohesion of gazetteers (i.e. how many different lists contribute to the gazetteer), while runtime parameters control the operation of gazetteers, for example permitting or restricting case sensitive matching.

The gazetteer entries can be assigned to user defined parameters, which can be exploited by pattern matching rules. GATE by default allows the assignment of two parameters types on each list, the *Major* and *Minor* types, which are defined in the gazetteer definition file and are applied to the range of list entries. For example, a list of locations could have a *Major Type* "Location" and a *Minor Type* "City". Another list can also share the same *Major Type* but to have a different *Minor Type* for example "Country". Thus a more generic rule can exploit the *Major Type* parameter to find all locations including both cities and countries and a more specific rule to exploit just the *Minor Type* parameter for matching only cities or countries.

The prototype pipeline uses JAPE rules that exploit the *Major Type* parameter to support identification of CRM entities. This enabled the composition of matching rules that exploited the total range of entries of a gazetteer list. For example a JAPE grammar exploiting the *Major Type* "*Physical Object"* delivered matches from a gazetteer list which carried the range of MDA Object Type thesaurus entries. This particular functionality of exploiting the range of available entries of a list without restriction and control, delivered a low Precision rate for some entity types, as revealed from the prototype evaluation (section 3.4).

The *SKOS Type* parameter was used by the prototype for assigning a SKOS terminological reference to individual gazetteer entries. This particular assignment enabled annotations to carry a terminological reference, in addition to their ontological reference. Similar steps for the assignment of class references to terms of terminological resources have been reported in the field of document classification (Golub, Hamon and Ardö 2007). However, using the *SKOS Type* parameter for the composition of precise rules might be possible but also impractical. Rules would need to include every single SKOS reference for each gazetteer term participating in the grammar, cancelling the benefit of list parameterisation and resulting to the composition of verbose rules. On the other hand,

being able to define grammars that exploit SKOS references and relationships (broader - narrower term) is very desirable. Such grammars can be reusable due to the use of SKOS references and can enable controlled exploitation of gazetteers listings via SKOS relationships. The following section discusses a technique of gazetteer parameterisation that enables controlled exploitation of gazetteer listings via SKOS references.

4.4.1 Skosification of Gazetteers

The term "Skosification" is used here to denote the parameterisation of gazetteer lists, with respect to SKOS terminological references. As discussed previously, the level of overlap between the terminological resources is capable of introducing ambiguity during the NER task, especially for the entities, Physical Object and Material. Also the prototype system revealed problems with Precision when exploiting the total range of large gazetteer listings. Therefore, a mechanism should be put in place for enabling definition of rules, which are capable of exploiting selected parts of terminology resources. The Skosification of the gazetteer resources was seen as a parameterisation technique capable of providing a solution to the above issue of selective exploitation of gazetteer listings.

Skosification of the gazetteer entries enables exploitation of the hierarchical relationships (broader – narrower term) leading to automatic, concept expansion techniques, as discussed in section 4.4.3 and chapter 5 section 5.3. As discussed in section 3.2.3, thesaurus hierarchies are less formal than the ontological class-subclass relationship, however their semantics can be used to support document retrieval and concept identification in context. In addition, being able to maintain a clear distinction between ontological and terminological references can be beneficial, providing flexibility and specialisation on both the ontological and terminological level. For example, the term "Animal Remains" can have a terminological reference (skosConcept: 95074) and an ontological reference (E19.Physical_Object), allowing the annotation of concepts via terminological and ontological constructs without requiring the two to be combined under a unified knowledge base structure.

4.4.2 Transformation of Terminology Resources to GATE Gazetteers

The process of GATE Gazetteer Skosification is based on the use of XSL templates, which transform the selected terminology resources from SKOS/XML format to gazetteer listings. The resulting text files are not flat, meaning that each entry of a gazetteer list contains not only a thesaurus or a glossary term but also additional properties about the term, which can be exploited by JAPE rules. The additional properties store information about the terminological reference of terms (*skosConcept*) and their path to the top term of the structure. The latter property (*path*) is project specific and reflects the hierarchical arrangements, i.e. Broader – Narrower term relationships. This particular property is exploited by JAPE rules in order to enable exploitation of the hierarchical relationships.

The transformation process also considers preferred and alternative labels of terms, allowing exploitation of term synonyms via a single terminological reference. Both preferred and alternative terms have a common *skosConcept* reference. For example, the terms, "Animal Remains", "Antler", "Animal Skeleton", "Animal Tooth" and "Faunal Remains" all have the same terminological reference (`skosConcept=95074)`.

Glossaries do not enjoy hierarchies, apart from the glossary Simple Names for Deposits and Cuts, which contains a shallow hierarchy for describing relationships between archaeological contexts and groups. However, this particular hierarchical relationship is not exploited due to the higher level of generality of grey literature in reporting primitive contexts and their groupings (section 3.3.2). All glossary terms that overlap with thesaurus terms are assigned the additional property s*kosExactMatch*. The extra property provides links to the thesaurus terminological reference. The *skosConcept* property is also maintained. Thus, each glossary term that overlaps with a thesaurus has two terminological references, one pointing to a glossary and another to a thesaurus resource.

There are some rare cases where a glossary term overlaps with two thesauri, as for example the term "Brick". In such cases, which mainly concern ambiguity between Physical Object and Material, an additional entry (new line) is added in the glossary listing in order to enable linking to a second terminological resource. For example

```
Brick@skosConcept=ehg027.4@skosExactMatch=96010
Brick@skosConcept=ehg027.4@skosExactMatch=97777
```

There are no recorded cases of a term overlapping with more than two thesauri that have different ontological alignment but even in such an extreme scenario an additional glossary entry could be added to implement a linkage.

4.4.3 Gazetteer Skosification by Example

The following example reveals the mechanism of thesauri relationships exploitation via GATE gazetteers, based on SKOS parameterisation. Consider the following MDA Objects thesaurus structure from broader to narrower term*: Dress and Personal Accessories* > *Personal Ornament* > *Jewellery* > *Brooch* > *Bow Brooch* > *Caterpillar Brooch*

Each of the above terms enjoys a unique terminological reference. For example "Caterpillar Brooch" has the SKOS reference "141231". The same unique terminological reference is assigned to "Ansate Brooch", which is a synonym of the term "Caterpillar Brooch" (Figure 4.6).

A JAPE grammar can acquire matches that correspond to a particular SKOS reference. For example, *match Lookup annotations having SKOS "141231":*

```
Lookup.skosConcept == "141231"
```

The above line of code would match both "Caterpillar Brooch" and "Ansate Brooch". Similarly another JAPE grammar can match terms having SKOS reference "96665", which corresponds to "Brooch" and its synonym "Fibula" and "Brooch spring".

However, JAPE grammars that acquire matching on exact SKOS reference would only match specific terms and their synonyms. An additional parameter should be put in place in order to enable rules to match not only specific terms and synonyms but also their narrow terms and/or their broader terms. The (specific to the project) *path* parameter describing the path of terminological reference to the top of the hierarchy, enables rules to exploit narrower and broader term relationships.

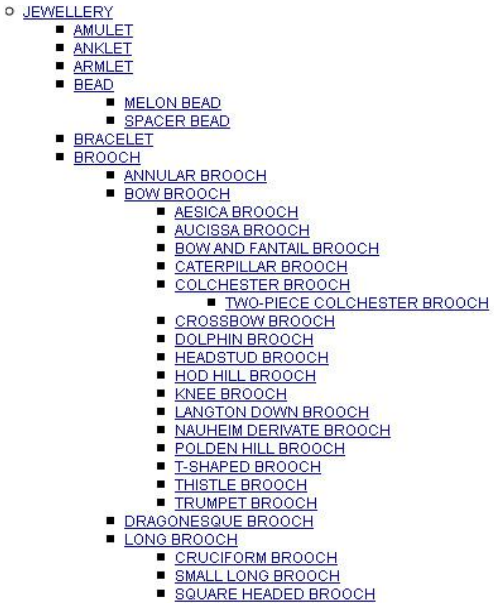


Figure 4.6: Hierarchy for term "Brooch" MDA Object Thesaurus

Considering the same example: *Dress and Personal Accessories* > *Personal Ornament* > *Jewellery* > *Brooch* > *Bow Brooch* > *Caterpillar Brooch*, the above terms were expressed as GATE gazetteer entries having two parameters skosConcept and path, as follows (parameters are separated by @).

```
Dress & Personal Accessories@skosConcept=100075@path=/100075
Personal Ornament@skosConcept=97111@path=/100075
Jewellery@skosConcept=96706@path=/97111/100075
Brooch@skosConcept=96665@path=/96706/97111/100075
Fibula@skosConcept=96665@path=/96706/97111/100075
Brooch Spring@skosConcept=96665@path=/96706/97111/100075
Bow Brooch@skosConcept=96662@path=/96665/96706/97111/100075
Caterpillar Brooch
@skosConcept=141231@path=/96662/96665/96706/97111/100075
- Ansate Brooch
@skosConcept=141231@path=/96662/96665/96706/97111/100075
```

JAPE grammars can make use of a range operators for testing equality (equal, non-equal), relational (greater than, less than) and partial matching (wildcard) of strings and parameter values. Therefore, grammars can match gazetteer entries that have a *skosConcept* equal to a specific value and/or a *path* parameter containing a partial value. The flexibility of JAPE operators to allow for partial matching is used to exploit narrower and broader term relationships that are expressed by the *path* parameter of each gazetteer entry. For example a JAPE grammar can be targeted at matching all narrower terms of the term "Bow Brooch" including "Caterpillar Brooch", "Dolphin Brooch", "Crossbow Brooch" etc. as seen in figure 4.6. The grammar of the rule would be:

```
{Lookup.skosConcept == "96662"}|{Lookup.path =~ "96662"}
```

The above grammar would match all gazetteer entries having a *skosConcept* equal to "96662", or their path parameter contains the value "96662", allowing for partial matches of the *path* parameter via the partial match operator (=~) .

Note a grammar targeted at "Brooch" (skosConcept:96665) not only matches the immediate narrower terms such as "Annular Brooch", "Bow Brooch", "Dragonesque Brooch" and "Long Brooch", but also matches sub-narrower terms (narrower terms of the narrower terms) such as "Caterpillar Brooch", "Dolphin Brooch", "Crossbow Brooch" etc. If it is required to exclude matching of sub-narrower terms the grammar can be altered to avoid for example matches that belong to the "Bow Brooch" (skosConcept:96662) category of terms . For example using the NOT operator the above rule can be modified as:

```
{Lookup.skosConcept == "96665"}|{Lookup.path =~ "96665",
Lookup.path !=~ "96662"}
```

The *path* parameter can also be used for matching broader terms similarly to matching narrower terms. For example a JAPE grammar can be defined for matching all broader terms of the term "Brooch", such as "Jewellery" and its narrower terms. Selecting the first SKOS reference of the path, which reflects the broader term of the hierarchy, the grammar below, will match all terms under "Jewellery", including all "Brooch" terms as well "Amulet", "Anklet", "Armlet", "Bead", "Bracelet" and their narrower terms etc.

```
{Lookup.skosConcept == "96706"}|{Lookup.path =~ "96706"}
```

If it is required to exclude any of the above branches the NOT operator can be employed. For example to exclude the "Amulet" and "Anklet" branches the rule can be modified as:

```
{Lookup.skosConcept == "96706"}|{Lookup.path =~ "96706",
Lookup.path !=~ "95896", Lookup.path !=~ "96650"}
```

The definition of the *path* parameter enables JAPE grammars to be flexible and to exploit specific areas of the gazetteer listings by referring to particular terminological references. This allows exploitation of narrow and broader term relationships. In addition, as discussed above, specific glossary concepts are given an extra attribute (*skos:exactMatch*) to accommodate linking between a glossary and a thesaurus.

The *skosExactMatch* property is used to assign a concept a second terminological reference to gazetteer entries, which can be later exploited at the retrieval level since all gazetteer parameters can be inherited to annotations.

4.5 Supportive GATE Gazetteer Listings

A small number of supplementary gazetteers are used to support the task of NER. The supplementary gazetteers are flat lists of terms that do not contain any parameters. Their aim is to support identification of entities by providing complementary vocabulary that is exploited by JAPE grammars. The complementary vocabulary is used in combination with Part of Speech (POS) input to support specialised NER tasks, which are targeted at Adjectival Conjunction and Negation Detection, as discussed in section 5.7 and section 5.8 respectively.

Four particular gazetteer listings available from the GATE framework, known as ANNIE gazetteers; the *Time prefix*, *Date Prefix*, *Date Suffix* and *Ordinal* are adopted to support expansion and conjunction of Time Appellation entities. The lists contain ordinals in their numerical and lexical form and terms which either precede or follow time related terms, such as "earlier", "later" or "period" and "century". The initial ANNIE gazetteers

were enhanced with additional lexical variations of terms during the gazetteer enhancement process which is discussed below (section 4.6). The *Date prefix* listing contains 72 entries, the *Date suffix* listing 36 and the *Ordinal* listing 79 entries [Appendix A2]. The Negation Detection task as discussed later (section 5.8), utilises a specialised set of complementary gazetteer listings [Appendix B]. Four lists support the operation of the Negation task: the *Pre-Negation* list, containing words that are usually found at the beginning of sentences denoting negation, such as "No", "Don't", "Absence of", etc., the *Post-Negation* list, containing phrases such as "is excluded", "is ruled out", which are usually found at the end of negation sentences, the *Negation-Verb* and the *Stopclause-Negation* lists that support the definition of archaeology related negation detection rules.

4.6 Enhancements of Gazetteer Listings

In this phase, thesauri, glossaries and supplementary gazetteer listings are enhanced to include mainly spelling variations and synonyms. The process of enhancement is informed by the list of Frequent Noun Phrases (FNP), provided by Dr. Renato Rocha Souza. Thesaurus and glossary gazetteer listings are enhanced with lexical variations of multi-worded entries.

There is no enhancement of resources with new concepts. Adding new concepts in thesaurus and glossary listings would require the declaration of new SKOS references for each new term included. However, modification of the EH terminology resources is a task undertaken by a specialised group of scholars. Even if the lists could be modified just for project purposes the resulted annotations would be of little use because SKOS references given to the added terms would only be known internally in the Information Extraction pipeline and not to any retrieval system.

The enhancement phase processed the FNP list (with a simple Information Extraction pipeline (Figure 4.7), which made use of a range of ANNIE modules, the English Heritage Gazetteers resources and a bespoke JAPE transducer.

In details, the pipeline used a Tokeniser for splitting text into tokens, a Sentence splitter configured to identify a new sentence in every new line of the document, a Part of Speech Tagger, a Morphological Analyser and a Flexible gazetteer. The configuration employed the Morphological Analyser for providing matches at the level of Token roots. This enabled the Flexible Gazetteer, exploiting the available terminological resources, to match all lexical variations of the FNP list. For example the gazetteer entry "Brick" matches both "Brick" (Singular Form) and "Bricks" (Plural form) since both words have a common

lemma.

The pipeline used a bespoke hand-crafted rule for annotating the lines of the FNP list which did not match any gazetteer entry. In addition, the Flexible exporter module was used for exporting the annotations in XML format, which then transformed via XSL into a simple list containing the unmatched terms.

The FNP list consists of 1724 lines, each one containing a noun phrase entry. For example:

```
animal bone
animal bone fragment
animal bone species
animal burials
animal husbandry
animal remains
 ……
another building
```

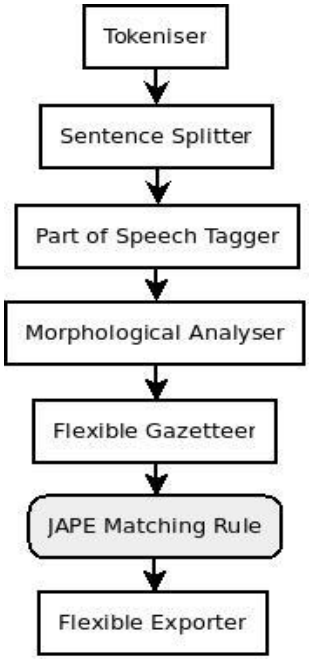


Figure 4.7: Pipeline used by the process of gazetteer enhancement

Each line that produced a match was excluded from annotation, even if matching was partial, for example “Burials” in “Animal Burials”. A total of 127 terms, just 7.4%, of terms were not matched [Appendix A3]. These terms were examined further and applicable terms were isolated and included in the EH gazetteer listings as synonyms of existing concepts. The examination of the unmatched terms also revealed variations in which multi-worded terms are spelled with regards to hyphenation. For example “Posthole” or “Post-hole” or even “Post - hole”.

The enhancement process based on the above observation added lexical variations to multi-worded entries of the glossaries and the Timeline thesaurus. The process did not enhance the Monument Type Thesaurus and the MDA Object type thesaurus because both thesauri are large containing mostly single worded entries. On the other hand, enhancing the Timeline thesaurus was more important since many of its entries use prefixes like “Early”, “Late”, “Mid”, “Post” etc. Variation in the composition of period terms which make use of such prefixes is common, for example “Early Roman” or “Early-Roman”. The above lexical variations were considered during the enhancement process of the timeline thesaurus. Each lexical variation was entered as a new entry in the gazetteer listing assigned to the same terminological reference with the main term.

For example:

```
Postmediaeval@path=/134715@skosConcept=134746
Post-Mediaeval@path=/134715@skosConcept=134746
Post - Mediaeval@path=/134715@skosConcept=134746
Post Mediaeval@path=/134715@skosConcept=134746
Postmedieval @path=/134715@skosConcept=134746
Post-Medieval@path=/134715@skosConcept=134746
Post - Medieval@path=/134715@skosConcept=134746
Post Medieval @path=/134715@skosConcept=134746
```

All glossaries were also enhanced with hyphenation variations. Glossaries also contain entries of the form "cesspit: fill" that denote a simple relationship between the two constituting words based on the use of the column character. It is unlikely that such written form is used when reporting. Therefore, such entries were elaborated to a written form, which is more commonly found, i.e. "cesspit fill", "fill of cesspit". However, multi-worded entries do not support matching on word root definition, as is the case with single worded entries. Thus, multi-worded glossary entries were also enhanced with their plural forms. An example of such enhancement can be found in [Appendix A5].

4.7 The Pre-processing Phase

A basic task of the preparation phase is the execution of the Pre-processing pipeline annotating the corpus of grey literature with a set of basic annotations, which are then utilised further by the NER phase. The Pre-processing pipeline was initially developed by the prototype, aimed at identifying rich discussion sections, heading spans and summary sections of documents. To deliver the annotation types, the pilot system made use of the ANNIE modules Tokeniser, Part of Speech Tagger (POS) and Verb Chunker.

The full scale NER phase adopts the pre-processing phase developed by the prototype while expanding it to support the requirements of the main NER phase. In detail, the full scale Pre-processing pipeline (Figure 4.8) employs a Tokeniser, a POS, a Morphological Analyser, a Noun Phrase Chunker, a Verb Chunker, and a Gazetteer listing of frequent noun phrases in order to deliver the annotation types; *Token*, *Heading*, *Section*, *Summary*, T*able of Contents (TOC*), *Noun-phrase* and *Verb-phrase*. Apart from the *Noun-phrase* annotation type, all other types had been previously targeted by the prototype pre-processing pipeline.

The main Pre-processing pipeline is little different from the equivalent prototype Pre-processing pipeline discussed in Section 3.3.1. It adopts all prototype rules and cascading order for generating the Pre-processing annotation types as discussed in chapter 3. For the generation of Nounphrase annotations, the pipeline uses the input of the NP chunker and

the bespoke gazetteer of frequent noun phrases (FNP list). A JAPE rule combined the two inputs and generated a unified noun phrase annotation, which was then utilised by the main NER pipeline to validate Lookup annotations.

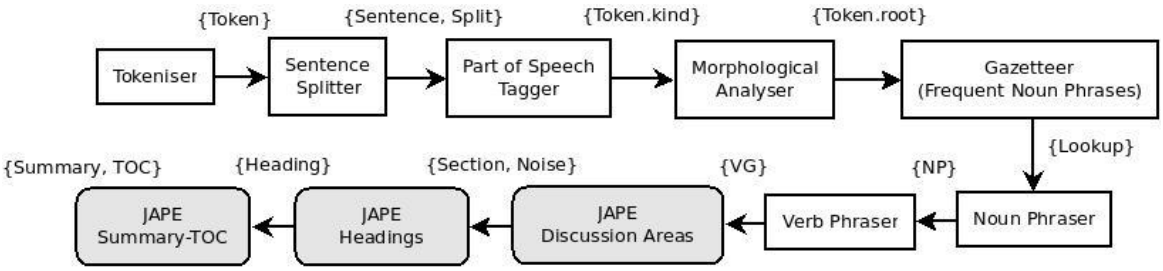


Figure 4.8: The Pre-processing pipeline, rounded grey boxes correspond to bespoke JAPE rules. The Annotation type produced at each stage of the pipeline are denoted within the curly brackets

4.7.1 Pre-processing Modules

In detail the full scale pre-processing pipeline employs the following modules:

Tokeniser: The Tokeniser splits text into small textual fragments known as Tokens. Tokens can be words, numbers, symbols, punctuations and white spaces referred as space tokens. In the case of words, the Tokeniser provides attributes which describe the case of tokens, such as; *upperInitial* (first letter in upper case the rest in lower case), allCaps (all upper case letters), *lowercase* (all lower case), *mixedCaps* (any mixture not included in the above categories)

POS: The Part of Speech tagger of GATE is a modified Brill tagger[4] employing supervised learning. The GATE POS is trained on a large corpus taken from Wall Street journals and provides around 50 tags reflecting different parts of speech. For example DT is used for determiners, IN for prepositions, JJ for adjectives etc. The POS tags, if not specified otherwise, reside as attributes of the Token annotation.

Morphological Analyser: The module analyses Tokens in order to identify the root base and affix of words. Tokens must be previously attributed by a part of speech tagger, in order for the module to operate successfully, since the analyser determines word lemmas. Thus, it relies on POS evidence for the identification of roots of words. The GATE morphological analyser is based on rules originally implemented by Kevin Humphreys in GATE version 1. The module allows adaptation of new rules and modification of the existing rules.

[4] The **Brill tagger** is a supervised machine learning method for doing part-of-speech tagging.

Noun Phrase (NP) Chunker: This module uses input from the POS module to define noun phrases in text. The module is a Java implementation of the Ramshaw and Marcus (1995) NP chunker that uses transformation-based learning to derive the noun phrases.

Verb Chunker: The Verb Group Chunker module uses 68 JAPE rules to identify verb group constructs. The module uses input from the POS module and delivers annotations containing attributes about the tense and voice (passive – direct) of the verb constructs.

Bespoke Gazetteer: The gazetteer list consists of 1724 frequent noun phrases that were extracted from a corpus of 550 OASIS documents. The list was composed by Dr. Renato Rocha Souza, professor at Fundação Getúlio Vargas (Brazil) who during 2010 was on a postdoctoral research fellowship at the University of Glamorgan. The frequently occurring noun phrases were extracted in GATE using the Multi-lingual Noun Phrase Extractor (MuNPEx). The output from MuNPEx is complementary to the Ramshaw and Marcus noun chunker. The gazetteer list is used to match noun phrases that are frequent within archaeological text but are not identified by the default NP chunker.

4.8 Summary

The chapter revealed the terminological resources that are employed by the full-scale system OPTIMA (2012) and their transformation from XML structures to Skosified GATE gazetteers listing. A thorough analysis of the terminological resources is given regarding their level of overlap and their ontological alignment with CIDOC CRM entities. The discussion revealed the rational supporting parameterisation of gazetteers listings with SKOS references and their potential to support the task of NER via exploitation of thesauri relationships (broader – narrower term). The method of gazetteer listing enhancement with synonyms and lexical variations was also revealed and the role of supportive gazetteer listings was also highlighted. The chapter also introduced the topic of NER providing background literature on relevant projects and schools of thought. The following chapter discusses the main NER phase of the OPTIMA pipeline, targeted at annotating the four CIDOC CRM entities, *Physical Object*, *Place*, *Time Appellation* and *Material*.

Chapter 5

Named Entity Recognition with CIDOC CRM

5.1 Introduction

The chapter discusses in depth the NER phase of the OPTIMA pipeline (Figure 5.1). The pipeline is focused 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 terminological resources, which have been expressed (skosification) as parameterised GATE gazetteer listings during the preparation phase (section 4.4). The discussion reveals an innovative technique of gazetteer exploitation via synonym, narrower and broader term relationships. The details of the technique are discussed and the role of the skosified gazetteer listings is revealed in the process of delivering semantic annotation with respect to ontological and terminological references.

The chapter also discusses the individual NLP stages that aim to improve the accuracy of the NER pipeline. In particular the pipeline addresses the issue of vocabulary polysemy, which has been identified by the terminological resources overlap study (section 4.3.3), with regards to the CRM entities, Physical Object and Material. In addition the role of negation detection is discussed and its algorithmic details are revealed. A number of JAPE grammars are also discussed, supporting the discussion with rich examples of hand-crafted rules employed by the various stages of the pipeline. Figure 5.1 presents the cascading order of the pipeline starting from the involvement of gazetteer listings in the process of NER, passing to semantic expansion via thesaurus relationships and ending with the bespoke NLP modules of validation, disambiguation, expansion and negation detection of semantic annotations.

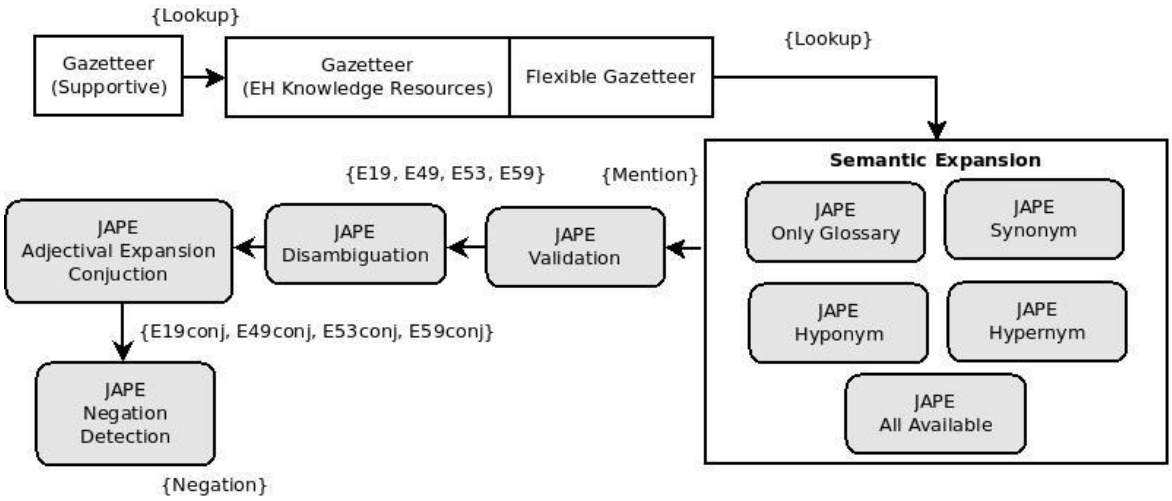


Figure 5.1: The NER phase of the OPTIMA pipeline. Curly brackets show the annotation types produced at the different stages of the pipeline. White boxes used for ANNIE modules, grey boxes used for bespoke rules and modules.

5.2 Configuration of Gazetteer Listings

The NER pipeline, as discussed in chapter 4, employs five glossary and four thesauri resources to support identification of CRM entities. A range of complementary gazetteer resources is also employed for equipping the NER task with supportive vocabulary. This vocabulary is utilised by hand-crafted rules for supporting the NLP tasks of adjectival expansion and negation detection.

The following section discusses the configuration of the above terminological resources as gazetteer resources. GATE allows a pipeline to use multiple gazetteer resources. Gazetteers may be composed of more than one listing, where each listing can accommodate multiple terms. The participating gazetteer listings of the NER pipeline are grouped under two distinct gazetteer resources. One gazetteer resource groups together the EH glossaries and thesauri resources, while another resource groups all the listings of supportive vocabulary. The first group will be referred to as the *SKOS gazetteer*, while the second group will be referred as the *Support gazetteer*.

5.2.1 Initialisation and Runtime Parameters of Gazetteers

Gazetteer resources, as most GATE modules, enjoy two kinds of parameters; runtime and initialisation. Runtime parameters are used during the execution of the pipeline, while initialisation parameters are declared once, during the instantiation of the module.

The gazetteer runtime parameters are concerned with a) the name of the Annotation Set where the resulting *Lookup*[5] annotations are created, b) whether the gazetteer will match whole words and c) whether the gazetteer will look for the longest matches. On the other hand, initialisation parameters are concerned with a) the text encoding type of the gazetteer resource, b) user-defined features, c) case sensitivity of matches and d) name of the gazetteer definition file which joins the various gazetteer listings under a single gazetteer resource.

Both SKOS and Support gazetteer use the *default* configuration of runtime parameters which means that the delivered Lookup annotations are assigned to the *default* Annotation Set for whole words and the longest matches. Setting the whole words parameter to true means that the gazetteer does not support partial matches, for example the entry "house" will not match "household". Setting the gazetteer for longest matches means that multi-word entries do not provide partial matches. For example, the gazetteer entry "animal remains" will not match "animal" or "remains" but only the complete match. The initialisation parameters for both SKOS and Support gazetteer resources use UTF-8 text encoding and case insensitive configuration, enabling matching regardless of the case of words.

5.2.1 Flexible Gazetteer Configuration

An important aspect of the configuration of the SKOS gazetteer is the use of the Flexible gazetteer. The Flexible gazetteer is a GATE plug-in that supports customised Lookup matching by using an external gazetteer and an annotation type feature. The Flexible gazetteer is used by the OPTIMA pipeline to enable matching of the SKOS gazetteer entries independent of singular or plural forms. By using the *root* Token feature, which corresponds to the lemma of a word and is assigned by the Morphological Analyser during the Pre-processing phase, the Flexible gazetteer employs the SKOS gazetteer to perform matching on the lemma feature of the words. Therefore, the gazetteer entry "bone" will match both singular and plural forms (bone, bones) since both forms share the same lemma.

This configuration is applicable only to single worded gazetteer entries since gazetteer resources are configured for longest matches. For example, the gazetteer entry "human bone" will not match "human bones" in text since the morphological analyser assigns

5 The annotations produced from gazetteers are known as Lookup annotations. Lookup annotations are simple matches of gazetteer terms in text.

lemmas on the token level but the gazetteer resources perform for the longest match, in this case two tokens. In order to support Lookup matching of such cases, the gazetteer enhancement process, discussed in section 4.6, has populated the glossary listings with the plural variation of multi-worded entries.

Another issue that affects the performance of the Flexible gazetteer is the matching of verbs that have a common lemma with nouns, for example "nail" and "nailing" share the same lemma. However, this particular problem would have occurred even if Flexible gazetteer configuration was not adopted since many gazetteer entries carry a dual sense, such as the terms "building", "find", and "cut". This particular problem of lexical ambiguity is addressed by the Lookup validation phase which is discussed in section 5.4.

5.3 Semantic Expansion for Information Extraction

Thesaurus structures have been successfully used in Information Retrieval for indexing and searching, as for example in the case of the FACET project (Tudhope et al. 2006). That project employed a semantic query expansion technique, which exploited the relationships of the Art & Architecture Thesaurus and a number of smaller specialist thesauri, in order to rank matching results by their semantic closeness, while providing a controlled vocabulary for search and indexing of datasets. However, similar approaches based on the exploitation of semantic relationships of thesaurus structures by Information Extraction systems remain unexplored.

5.3.1 Semantic Expansion for Information Extraction

The process of Semantic Expansion for Information Extraction introduces a novel approach which utilises Terminology-based resources for driving the task of NER. The novelty of the mechanism resides on its capability to invoke a controlled semantic expansion technique, which exploits synonym and hierarchical relationships of terminological resources for assigning distinct terminological and ontological definitions to the extracted results. The mechanism is capable of selective exploitation of gazetteer listings via synonyms, narrower and broader concepts relationships. Hence, the system can be configured to a range of different modes of semantic expansion depending on the aims of an IE task, i.e. being lenient and applying a generous semantic expansion or being strict and applying a limited semantic expansion.

The prototype system revealed the need of controlled exploitation of large gazetteer listing in order to improve precision. The full-scale system is capable of using three

different semantic expansion configurations, ranging from *synonym* expansion to *hyponym* and *hypernym* expansion modes, delivering diverse information extraction results. The results of the different semantic expansion modes are later used to evaluate the system configuration that delivers the best Precision and Recall results. Together with the above three expansion modes, two additional configurations are defined; one that does not use any semantic expansion and another that uses the entirety of the resources. The system configurations with regards to the semantic expansion modes are discussed below, while their IE results are revealed in the evaluation chapter.

5.3.3 Strict Mode of Semantic Expansion – Only Glossaries

The Strict mode is the simplest configuration of the pipeline. The mode does not make use of the semantic expansion mechanism; it simply uses the glossary gazetteer listings to identify Lookup annotations in text. The use of the glossary listings is based on the assumption that such glossary resources contain core archaeological terminology. The glossaries originate from the recording manuals developed by English Heritage to support field archaeologists to record excavation facts and findings. Hence, they contain controlled and consistent domain vocabulary forming a core knowledge resource of terms which is used by archaeologist when recording information of excavation results.

The mode employs the five selected glossaries and follows the alignment between glossary resources and ontological entities as described previously in table 4.1. Thus, the glossary *Simple Names for Deposits and Cuts* is aligned to the ontological entity *E53.Place*, the *Box Index Form (Material)* glossary is aligned to the entity *E57.Material*, the *Box Index Form (Find Type)* is aligned to the entity *E19.Physical_Object*. Similarly the *Small Finds Form* glossary is aligned to *E19.Physical_Object*, while the *Bulk Finds Material List* is aligned to *E57.Material*. The mode also employs the *EH Timeline Thesaurus*, which is aligned to the ontological entity *E49.Time_Appellation*. The thesaurus class *Political Period,* which is comprised of names of political leaders, Roman emperors, royal dynasties and Welsh kings is not used, since its contents are not relevant to the scope of the *E49.Time_Appellation* entity. This arrangement is followed by all configuration modes of semantic expansion that use the *Timelin*e thesaurus.

The Strict – Only Glossaries configuration is based on the use of JAPE rules for the definition of Lookup annotations. A single JAPE grammar per entity type is used for enabling annotation of entities.

For example, the following grammar `({Lookup.skosConcept =~"ehg019"}| {Lookup.skosConcept=~"ehg027"}),` exploiting the *skosConcept* parameter, will match all gazetteer entries of the *Index Form Material glossary* ("ehg019") or the *Bulk Finds Material glossary* ("ehg027"). All matches of the above condition are aligned to the *E57.Material* entity. Similar matching conditions are employed by the pipeline for annotations aligned to the other CRM entities, i.e. Physical Object, Place and Time Appellation entities.

Specifically the matching rule for Time Appellation used by the strict mode is used by all five modes of semantic expansion. The rule matches all gazetteer terms that contain in their SKOS path any of the following terminological references "134749", "134715", "136295", "134716", which correspond to the thesaurus top classes "Artistic Period", "Cultural Period", "Geological Period" and "Historic Period" respectively. Thus, the rule exploits the entirety of the EH Timeline thesaurus resource, excluding the entries belonging to the *Political Period* class.

The Time Appellation is the only CRM class that does not participate in the semantic expansion mechanism for the following reasons; a) All entries of the Timeline thesaurus are relevant to the NER of Time Appellation entities b) there is no glossary to distinguish a sub-set of Time Appellation terms relevant to the archaeological domain as is the case of the other three entities and c) the prototype development delivered good Precision and Recall rates for Time Appellation, which is considered indicative of the coverage of the Timeline thesaurus.

5.3.4 Synonym Mode of Semantic Expansion

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. Hence, the *Synonym* mode enhances the *Strict* mode by including both glossary terms and their synonyms from thesauri.

The study of Terminology Resources overlap (section 4.3.3) revealed the cases of overlapping terms between glossary and thesaurus resources. Based on the semantic alignment between ontological entities and terminology resources, the overlapping terms are assumed to have common word senses. For example the terms "pit" (terminological reference: ehg003.55) which originates from the glossary *Simple Names for Deposits and Cuts* share the same sense with the term "pit" (terminological reference: 70398), which

originates from the *Monument Type Thesaurus*. Therefore, a glossary term can inherit the same semantic relationships of its equivalent (overlapped) thesaurus term.

The Synonym expansion mode uses rules that exploit both glossary and thesaurus resources. Thus, the mode incorporates two sets of rules, those previously used by the *Strict* mode and a new set of rules which exploit the synonym relationships of the thesauri resources. Similarly to the Strict mode, a set of JAPE grammars is responsible for identifying Lookup annotations that correspond to the matching of glossary terms. For example, the grammar `{Lookup.skosConcept =~"ehg003"}` matches all terms of the glossary *Simple Names for Deposits and Cuts,* aligned to the ontological entity, E53.Place.

An additional set of grammars is also used for identifying matches that correspond to the synonyms of the overlapped terms. For example, `{Lookup.skosConcept ==="70398"}` matches the term "pit" and its synonym "pit dwelling". Note the double equality (==) operator is used by the rule for matching only those thesauri terms that have the specific *skosConcept* parameter.

The two sets of rules do not conflict with each other - neither produce verbose results but they are complimentary. Based on the "appelt"[6] mode of rule execution, the two rule sets do not annotate the same piece of text more than once. Although, the term "pit" is found both in glossary and in thesaurus and two different rules are executed, one targeted at exploiting glossary terms and another targeted at thesaurus terms, still the "appelt" mode of rule execution prevents the term "pit" from being matched more than once. On the other hand, the rules operate in a complementary manner. This means that the rule targeted at matching glossary terms will match the term "pit" while the rule targeted at matching thesaurus terms will match the term "pit dwelling", which is a synonym of "pit" that shares the same terminological SKOS reference.

6 GATE allows execution of rules in five different modes, *appelt, brill, first, once* and *all*. The default execution is *appelt* which is used extensively by the pipeline. The *appelt* mode produces the longest matches while when matching occurs the text passage is "consumed" and the parser moves to the next Token. The *first* mode is very similar to the *appelt* mode but produces the shortest match. Shortest or Longest match is dictated by the definition of matching patterns which make use of recursive operators (*,+) for matching zero or many and one or many tokens/annotation types. The *brill* mode does not consume the text passage upon matching, allowing other rules to produce annotation from the same text, while the *all* mode allows matching of all rules included in a JAPE transducer. The *once* mode allows execution of only one rule once, thus exhausting the JAPE transducer, prohibiting in this way matching of the rest of rules included in the set.

5.3.5 Hyponym Mode of Semantic Expansion

The Hyponym expansion mode exploits the *Narrower Term* semantic relationships of thesauri structures. The mode increments from the *Synonym* mode using the identified overlapping terms between glossary and thesauri as entry points to the thesauri structures. Additionally to the *Synonym* expansion, the *Hyponym* mode exploits both synonym and narrower term relationships. For example, executing the expansion mode for the term "pit", matches "pit" its synonyms, such as "pit dwelling", and the narrower terms, such as "ash pit", "fire pit", "latrine pit", "lime slaking pit", "lye pit". The mode also matches the synonyms of the narrower terms since such synonyms are also narrower terms of the targeted concept.

Similarly to the Synonym mode, the Hyponym mode uses two sets of rules, one targeted at matching glossary terms and another targeted at matching thesaurus terms. The definition of the rules is very similar to the definition of the Synonym mode, with the main difference that thesaurus matching rules exploit the SKOS path parameter instead of the SKOS concept parameter. The matching condition `({Lookup.path=~"70398"})` matches all thesaurus terms that contain the figure "70398" in their path. Thus, the rule will match "pit", its synonyms, its narrower terms and the synonyms of its narrower terms since all those terms contain the targeted figure "70398" in their path parameter.

Also note that the use of the tilde (~) operator enables transitive matching within a thesaurus hierarchy. This means that not only narrower terms and their synonyms are matched but also narrower terms of narrower terms and their synonyms are also matched. The *path* property describes the path of a thesaurus term to the top of the hierarchy using terminological references. Therefore, any term, regardless of how deep in the hierarchy it might be, that contains in its path the figure targeted by the rule is matched.

While the matching rule targets a single figure, the tilde (~) operator allows matching of many different terms which contain the targeted figure in their path. Such terms have different terminological (SKOS concept) references. However, the rule does not hamper or overwrite such references but passes them on as an annotation parameter of the matched concept. In this way, the annotations produced enjoy a clear terminological reference and a distinct ontological definition. As discussed, keeping terminological and ontological references separate enables semantic annotations to encapsulate information on both conceptual and terminological levels.

5.3.5 Hypernym Mode of Semantic Expansion

The *Hypernym* mode of semantic expansion enhances the *Hyponym* mode by including matching of broader terms. The mode allows transitive matching similarly to the *Hyponym* mode. Therefore, not only the broader semantic terms are matched but also the narrower terms of the broader term are also matched, including synonyms. For example "Archaeological Feature" is the broader term of "Pit" in the structure of Monument Types thesaurus. The *Hypernym* expansion matches all "Archaeological Feature" terms, such as "Buried Landscape", "Hearth", "Posthole", "Site" etc. In addition, the mode includes all narrower terms of narrower terms, similarly to Hyponym expansion. Therefore, "Occupation Site" which is a narrower term of "Site" is also included by the semantic expansion, together with synonyms of "Site"[7] such as "Crop Mark", and "Soil Mark".

The Hypernym mode uses two sets of rules, similarly to the Synonym and the Hyponym expansion modes, each set targeted at matching glossary and thesaurus terms respectively. The main difference of the Hypernym configuration from the other modes of semantic expansion is that the rules make use the broader term reference of the overlapped terms. Following the same example as previously, the term "pit" has a broader term "Archaeological Feature", which has the terminological reference "102912". In this case, the JAPE grammar is `({Lookup.path=~"102912"})`. The rule matches all gazetteer terms which contain in their path the above terminological reference, including synonyms, narrower terms and narrower of narrower terms of the "Archaeological Feature" term.

The *Hypernym* mode uses the same technique as the *Hyponym* mode in assigning terminological references to multiple matches produced by a single rule that exploits the *path* parameter of gazetteer entries. Therefore, matches enjoy terminological references which are passed on as annotation properties allowing for their dual standing as terminological and ontological entities.

[7] The CRM-EH ontology assigns to the concept of "archaeological site" the CIDOC-CRM class *E27.Site* extended by the CRM-EH class *EH0002.Archaeological Site*. This particular form of modelling is beyond the scope of the thesis.

5.3.6 All Available Terms Mode

The *All Available Terms* configuration does not make use of the semantic expansion mechanism. It uses the entirety of the available resources from the glossary and the thesaurus gazetteer listing. The definition of the matching rules is based on the exploitation of the *Major Type* property of gazetteers similarly to the rules used by the prototype system. The configuration is used to produce annotation results that are used by the evaluation stage for comparison with the semantic expansion modes.

5.4 Validation of Lookup Annotations

The main aim of the validation stage is to examine via contextual evidence the Lookup annotations produced by the Semantic Expansion. The validation task is addressed at *Place*, *Physical Object* and *Material* Lookup annotations, while *Time Appellation* annotations are excluded. This is because *Time Appellation* annotations are less ambiguous and frequently are not part of noun phrases due to the extensive use of adjectival moderators such as “Earlier”, “Mid”, “Late” etc.

Since the task of NER is targeted at recognising entities, in effect nouns, it is important to invoke a validation stage to drop out all verb instances that might have been identified as entity Lookups. The stage is required because of two main reasons. Firstly because in English, many verb and noun forms are spelled exactly the same, as for example the word “building” which may refer to the noun sense “Structure of a building”, or to the verb sense “ building a structure”. Secondly, as discussed in section 4.4.2, the configuration of gazetteer matching uses the Flexible gazetteer module for enabling matching at the level of word root (lemma). This allows matching of singular and plural forms from a single gazetteer entry. However, this configuration also causes matching of verbs which share the same word root with nouns, for example “nail/nailing”, “drain/draining” etc.

The matching grammar which implements the validation is simple and is based on the JAPE operator *within,* which matches annotations types which are within other annotation types. The *Nounphrase* annotation is generated during the pre-processing phase. This particular type of annotation is defined by the input of the Noun Chunker module and the gazetteer listing of frequent noun phrases. A JAPE rule is used by the validation process for matching Lookup annotations that are within *Nounphrase* annotations. The annotations which are matched by the rule qualify as valid Lookup annotations and are assigned a new property *validation* equal to NP. Those not matched are removed from a subsequent stage.

The stage also performs some additional validation tasks. A particular validation task is aimed at removing Lookup annotation of proper names. The NER is not targeted at recognising proper names but occurrences of concepts. Therefore, matching of "Church" in the phrase "…..Church of England…" is avoided. Note that the rule examines if capitalisation is a result of a new sentence commencing, in order to avoid removing valid annotations.

Another type of validation is specifically targeted at the word "(B/b)eaker", the only overlap case that concerns Time Appellation Lookups. The word when in capital "B" refers to a period and when in lowercase "b" in a physical object. The validation rule checks against word case and assigns a Lookup sense accordingly.

The validation stage also examines whether Lookup annotations are part of *Heading*, *Table of Content* and *Noise* sections of documents. Document sections are identified and appropriately annotated during the Pre-processing phase. The validation stage examines whether *Nounphrase* annotations, which consequently validate Lookup annotations are part of the above document sections. If they are part of such sections they are removed and consequently the contained Lookup annotations are also removed.

The Lookup annotations that pass the validation stage are finally assigned as *single sense* CRM annotations and are assigned the annotation types E19, E53 and E57 corresponding to the Physical Object, Place and Material concept respectively. Time Appellation Lookups are assigned the annotation type E49. Multi-sense Lookups i.e. annotations that share a *Physical Object* and a *Material sense*, are disambiguated by a dedicated disambiguation phase, which is discussed below.

5.6 Disambiguation Phase

5.6.1 Introduction to Word Sense Disambiguation

Word Sense Disambiguation (WSD) refers to the computational ability to identify the different meanings of a word that has multiple meanings (Navigli 2009).

Words that share the same spelling, or same pronunciation or both, but carry a different meaning are known as Homonymous. For example, "arms" and "left" are cases of Polysemous words with "arms" (as in plural of the body parts) being homonymous with "arms" (as in military force) and "left" (as in direction) being homonymous with "left" (as in past tense of leave). The Homonymous words that share the same spelling are called Homographs as for example "desert" (as in arid area or as in meal course), while

Homonyms that share the same pronunciation are called Homophones, as for example "to/too/two".

Early attempts to answer the problem of polysemy via computational means originate back to the 1950's. Initially the attempts were focused in limited domains or over small vocabularies. From the 1980's improvements in the scalability of WSD systems were made due to the advances of the available computational means and the progress of Machine Learning (ML) techniques, enabling disambiguation over larger heterogeneous resources.

WSD applications can be rule-based or Machine Learning (Sanderson 2000). When rule-based, a WSD task invokes hand-crafted rules which utilise contextual evidence and knowledge resources for determining the disambiguation results. Rule-based WSD approaches can exploit general context rules which assume that a word sense can be determined by particular words which appear near to the ambiguous word. Also rules can utilise templates which state that an ambiguous word has a certain sense when a particular word(s) appears in a specific location relative to that word. Machine learning approaches can be supervised, which require a training set for determining the disambiguation results, or they can be unsupervised. Knowledge based resources, such as dictionaries, glossaries, thesauri etc., can also be used by a WSD task for supporting inference of word senses in context.

Voorhees (1993) devised a WSD system based on WordNet, a well-known knowledge based resource. WordNet is described as a lexical database which organises words under a structure of *synsets*. A synset is a group of synonyms, hence the meaning of a synset is defined by its words and the sense of a word by the synset it belongs to. WordNet arranges synsets in a complex semantic network of Hypernym (broader terms), Hyponym (narrower terms), Meronym (has part), Holonym (is part of), and Antonym (is opposite) relations. Voorhees's system exploited synsets of nouns for defining what she called *hoods* which were used to determine the sense of ambiguous terms. Based on the assumption that a set of words that occur together in a context determine appropriate senses for one other, the system populated diverse *hoods* with words from different *synsets* for an ambiguous word in a given context. The *hood* with the largest number of occurrences determined the sense of an ambiguous word.

Voorhees's system was evaluated in the context of information retrieval and reported to perform worse than the original system that did not invoke the disambiguation technique. However, her approach signified the importance of contextual evidence in the disambiguation process, in line with (Resnik 1997) that linked the disambiguation process

with *Selectional Preference*. Selectional Preference is the tendency of words to co-occur with words that belong to specific semantic groups. Resnik used Selectional Preference for enabling disambiguation in an unsupervised machine learning environment. The model combined statistical and knowledge-based methods to determine the senses of ambiguous words by utilising syntactic relationships between words pairs and measuring their semantic association by frequency count.

WSD can be viewed as an automatic classification task that makes use of contextual evidence and external knowledge resources for applying an appropriate class (word sense) to ambiguous terms. The task of disambiguation can be focused on a particular set of words thus to be "targeted", or it can be applied to the vast range of all words in document. Typically a WSD task is configured as an intermediate task of a larger NLP application, either set up to execute as a standalone module or integrated within the system architecture. Although, use of ML approaches can improve the generalisability of a disambiguation method, still many WSD systems have inherited limitations in terms of their performance and generalisation, especially when fine-grained sense distinctions are employed by the disambiguation task (Navigli 2009).

5.6.2 Ontology Introduced Polysemy

The OPTIMA pipeline employs a WSD module aimed at resolving a particular type of polysemy. Polysemy in linguistic terms is defined above as the condition where a specific word carries multiple meanings. However, the adoption of the CIDOC CRM ontology for driving the NER task brought a specific form of polysemy, which is inflicted by the definition of ontology classes. It is a form of polysemy that is not purely based on the linguistic characteristic of words and their meanings but instead is dictated by the conceptual definitions of an ontological model.

A particular ambiguity that is inflicted by the CRM-EH ontology concerns the fine distinction between small find objects and materials. For example, a word that might not be polysemous in linguistic terms such as "pottery" can be ambiguous in a CRM-EH driven NER task. Pottery in CRM-EH can be classified as E19.Physical Object or as E57.Material with only one of the two classes considered to be correct in a given context. This particular problem is addressed by the disambiguation module of the OPTIMA pipeline.

More cases of ontological polysemy which do not fall within the scope of the thesis and in the general aims of the STAR project might be introduced when broadening the scope of the NER task. For example, the term "clay", a very frequently mentioned term in

archaeological reports, can be modelled in CRM-EH as the material of a find (EHE0030.Context Find Material) or as the physical material of a context (EHE0008 Context Stuff) when addressing soil types . Another case of ontological polysemy concerns the ambiguous cases between EHE0007.Context (modelled as place) and EHE0009.Context Find (modelled as physical object). In archaeological practice a large physical object such as a "skeleton" can be treated either as a context (thus a place) of excavation or a large find (thus a physical object). The aforementioned cases of ontological polysemy possibly can be resolved with the same disambiguation techniques that are described in section 5.6.3 but this remains to be tested by a future version of the pipeline.

The terminology resources overlap analysis (section 4.3.3) revealed that glossaries aligned to the ontological classes, Physical Object and Material, have a large number of overlapping terms [Appendix A1]. The overlapping terms were identified and handled by the NER task as ambiguous terms. The WSD module is targeted at automatically resolving the ontological polysemy of such words where possible and assigning them an appropriate ontological classification. For example, the term "pottery" can be part of the phrase "…ditch containing pottery and coins…" or "ditch containing pottery fragments". In the first case "pottery" refers to a physical object found in a ditch, while in the second case the same term refers to the material of fragments. It is important the NER task be able to deal with this form of ontological polysemy, in order to minimise the cases of incorrect classification of the recognised entities.

5.6.3 NER of Ambiguous Concepts

The range of the overlapping glossary and thesauri terms which do not have a clear ontological alignment are defined as ambiguous terms. The ambiguity tends to be between a Physical Object and a Material sense. A specific part of the NER pipeline is targeted at resolving their ambiguity (section 5.6.3). Although, the discussion on the modes of semantic expansion has not discussed ambiguity, all modes of semantic expansion take into account ambiguous terms during their matching process, as discussed below.

5.6.3.1 Initial Marking of Ambiguous Concepts

The very first step of the NER pipeline is to mark (annotate) all ambiguous terms. The ambiguous terms which were revealed from the overlap study are selected via their terminological reference (SKOS concept). Their terminological reference is then used by a JAPE grammar for annotating the terms as ambiguous. Consider the term "brick" which is

ambiguous and can refer either to a Physical Object or to a Material. The term is annotated as ambiguous by the matching condition `{Lookup.skosConcept =="96010"}`, where "960010" corresponds to the Physical Object sense of brick. This rule generates an annotation of the type *Mention,* assigning to the annotation the parameter "Ambiguous_Term". The same result could have been achieved by using the glossary terminological reference of "brick" (ehg027.4). It does not make any difference which terminological reference is used at this stage since the objective of the rule is to annotate the concept of "brick" as ambiguous.

Based on the assumption that the narrower terms of an ambiguous term are also ambiguous terms, the NER pipeline, depending on the mode of semantic expansion, employs a distinct set of rules for the annotation of ambiguous terms. The *Strict* and *Synonym* mode make use of a rule set that matches ambiguous terms by exploiting the SKOS concept parameter of gazetteer resources, as with the example of "brick" discussed above, in order to assign the "Ambiguous_Term" parameter.

The *Hyponym*, *Hypernym* and *All Terms* modes of semantic expansion employ a different set of rules which exploits the tilde (~) operator for annotating as ambiguous all the narrower terms of an ambiguous term. For example, the matching condition `{Lookup.path =~"96010"}` will annotate all synonyms and narrower terms of brick as ambiguous terms. Note that all the above modes make use of the same set of rules for the annotation of ambiguous terms. Broader terms of an ambiguous term are not considered to be ambiguous and so are not annotated as such. It would be a weak assumption to consider broader terms also as ambiguous since inheritance of attributes in hierarchical terms is normally passed from parent to children classes but not the other way around.

5.6.3.2 Lookup Annotation of Ambiguous Concepts

The marking (annotation) of ambiguous terms enables the NER pipeline to generate non-ambiguous and ambiguous Lookup annotations. The ambiguous annotations are processed further by the disambiguation phase of the pipeline. The modes of semantic expansion take into account the ambiguity of terms during the Lookup process. If a term is identified as ambiguous, it is not matched by any Lookup rule which deals with non-ambiguous terms in any of the modes of semantic expansion. This is implemented by including into the Lookup matching condition an additional statement which checks for term ambiguity.

Consider the above example for matching the term "pit". The matching condition for producing a Lookup annotation only if the match is not an ambiguous term translates as:

```
{Lookup.skosConcept =="70398",!Mention.type=="Ambiguous_Term"}
```

The above condition will match all gazetteer entries which have terminological reference equal to "70398" and are not ambiguous term. The NOT operator is declared by the use of exclamation mark (!) before the "Mention" annotation type, while the comma joins the two statements of the matching condition which is wrapped by the curly brackets.

The Lookup annotation of the ambiguous concepts is implemented by a set of rules which exploit the SKOS reference of gazetteer resources. Different rules are targeted at producing Lookup annotations of different senses while each Lookup rule has two versions. One version is used by the *Strict* and *Synonym* modes and another version by the *Hyponym*, *Hypernym* and *All Terms* modes of semantic expansion. The pipeline deals with two distinct senses of ambiguity, the sense of physical object and the sense of material. Therefore, in total there are 4 different rules which deliver the Lookup annotations of the ambiguous terms.

One rule is targeted at matching all the concepts (via their SKOS reference including synonyms) which are ambiguous and have a sense of a physical object. A variation of this rule implements matching of all concepts and their narrower terms. This latter rule is used when the pipeline executes in the *Hyponym*, *Hypernym* and *All terms* mode of semantic expansion while the first rule is used by the *Strict* and *Synonym* mode. Similarly two more rules are implemented for matching all the concepts (via their SKOS reference) which are ambiguous and have a material sense, one rule for the *Strict* and *Synonym* mode and another for the *Hyponym*, *Hypernym* and *All terms* mode.

Consider the example of the multi-sense term "Brick". The term has two senses, a physical object sense corresponding to the terminological reference "96010" and a material sense corresponding to the terminological reference "97777".

The responsible matching condition for producing Lookup annotations aligned to the physical object sense for the expansion modes *Strict* and *Synonym* is `{Lookup.skosConcept =="96010"}`. Similarly the matching condition for the same sense for matching term and narrower term used by the *Hyponym*, *Hypernym* and *All Available* mode of semantic expansion is `{Lookup.path =~"96010"}`, note the use of tilde. Both matching conditions produce a Lookup annotation of the kind *Mention*, having a *type* property equal to "Physical_Object" and a *multisense* property equal to "true".

Likewise the matching condition for producing Lookup annotations aligned to the material sense makes use of the same patterns. The terminological reference "97777" is used instead of "96010", while "Material" is assigned to the *type* property. Therefore, any textual instance of "brick" is assigned two annotations of the kind "Mention". Both have a

multisense property equal to "true" but one has *type* "Physical_Object" and the other "Material". The competing *Mention* annotations are used as input by the disambiguation phase, which resolves ambiguity and assigns the final sense, ontological and terminological reference, to a textual instance.

5.6.4 Techniques and Rules for Resolving Ontological Polysemy

The disambiguation technique for resolving ontological polysemy of physical object and material terms is based on the adoption of rules that implement grammatical templates. The rules assume that specific contextual collocation, expressed as templates, is capable of resolving ambiguity. Contextual collocation refers to the location of ambiguous terms in relative location to non-ambiguous terms. The module utilises three different groups of annotation types: (i) the non-ambiguous (single sense) annotations of E19.Physical Object, E49.Time Appellation and E57.Material types which are delivered by the previous stages of the NER pipeline; (ii) the (*Multisense*) annotations *Mention* of type "Physical Object" and "Material"; (iii) the *Token* input from the Part of Speech NLP module of the pipeline.

The disambiguation module resolves the appropriate terminological (SKOS) reference to ambiguous terms. For example when the term "brick" is disambiguated as material, the terminological reference "97777" originating from the *Main Building Material* thesaurus is assigned to the annotation. When the same term is resolved as physical object, the terminological reference "96010" originating from the *MDA Object Type* thesaurus is assigned instead. This is achieved by using specific Lookup rules for each particular sense of multi-sense concepts, as discussed in above (section 5.6.3.2).

The rules implement 16 different cases of contextual templates for the automatic disambiguation of physical object and material instances in text. The templates express grammatical rules of word pair, conjunction and other phrasal patterns that were empirically selected, based on the common use of English language. The list of rules is not exhaustive but covers common lexical patterns that can be invoked by the disambiguation process. Whenever the ambiguity of terms is not resolved due to limitations in the rules, annotation is assigned to both senses. This particular choice favours Recall rather than Precision 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.

5.6.4.1 Word Pair Pattern Rules

The word-pair rules define simple templates which examine the location of ambiguous terms in pair relation to other ambiguous and non-ambiguous terms. In total, the disambiguation module employs four JAPE grammars which are discussed below. The ambiguous terms targeted by the disambiguation rules are described in JAPE grammar as *Mention* annotations, while the non-ambiguous terms are described with their CIDOC CRM annotation type (E19, E49, and E57). Note that an ambiguous term can have either a physical object or a material ontological sense.

Grammar I: An ambiguous term followed by another ambiguous term. The grammar resolves the first ambiguous term as E57 (material) and the second ambiguous term as E19 (physical object), based on the use of English where the material of an object is stated first. For example in the phrase of "brick tile", "brick" is the material and "tile" the object.

```
({Mention.type=="Material", Mention.multisense=="true"
{Mention.type=="Physical_Object",Mention.multisense=="true"})
```

Note: that the property *Mention.type* is used by the rule in order to assign the correct terminological reference upon disambiguation resolution (see also section 5.6.3.2). The property *Mention.multisense* is used for matching an ambiguous term.

Grammar II: An ambiguous term followed by a non-ambiguous term of type Physical Object. The grammar resolves the ambiguous term as E57.Material based on the use of English as described in the previous rule. For example in the phrase "pottery fragment", pottery is resolved as material.

```
({Mention.type=="Material", Mention.multisense=="true"}{E19})
```

Grammar III: An ambiguous term followed by a non-ambiguous term of type Place. The grammar resolves the ambiguous term as E57.Material, based on the use of English where the material of a place is stated first. For example in the phrase "brick wall", "pottery" is resolved as material.

```
({Mention.type=="Material", Mention.multisense=="true"}{E53})
```

Grammar IV: A non-ambiguous term of type Material is followed by an ambiguous term. The grammar resolves the ambiguous term as E19 (physical object), based on the use of English as described in the previous rule. For example in the phrase "plaster tile", "tile" is resolved as physical object.

```
({E57}{Mention.type=="Physical_Object",
Mentionable=="true"})
```

5.6.4.2 Conjunction Pattern Rules

The patterns of this category target simple cases of conjunction between ambiguous and non-ambiguous terms for resolving ontological polysemy. The grammars are based on the assumption that co-ordinating conjunctions join terms of the same kind. The operators used for conjunction are comma ",",forward slash "/", the word "and" and the word "or".

Grammar I: An ambiguous term is conjunct with a non-ambiguous term of type Material. The grammar resolves the ambiguous term as E57.Material. For example in the phrase "brick and plaster", "brick" is resolved as material.

```
({Mention.type=="Material",
Mention.multisense=="true"}):match
({Token.string =="and"}|{Token.string =="or"}|
{Token.category ==","}|{Token.category =="/"})
{E57}
```

Grammar II: This grammar is the inverted version of the previous rule. A non-ambiguous term of type Material is conjunct with an ambiguous term. For example in the phrase "plaster and brick", "brick" is resolved as E57.Material.

```
{E57}
({Token.string =="and"}|{Token.string =="or"}|
{Token.category ==","}|{Token.category =="/"})
({Mention.type=="Material",
Mention.multisense=="true"}):match
```

Grammar III and IV: The rules are versions of the above rules (I and II) that examine conjunctions of terms of type physical object instead of material. For example the rules match cases like "coin and brick" or "brick and coin". The rules resolve the ambiguous term as E19.Physical object.

5.6.4.3 Phrasal Pattern Rules

A range of phrasal patterns is utilised by the module for addressing ontological polysemy between physical object and material terms. The list of phrases is not exhaustive but is representative of the kind of templates that can be employed for tackling term ambiguity. The templates were derived empirically by examining sample documents and abstracting patterns from phrases which carry clues for disambiguation. The rules use ambiguous, non-ambiguous terms and tokens which are parameterised with part of speech input. A full list of the part of speech Hepple tagger categories that are used by the rules below can be found in Appendix F1.

Grammar I: An ambiguous term of type Physical Object or non-ambiguous term of the same type followed by a preposition, followed by Time Appellation, followed by an ambiguous term of type Material or non-ambiguous term of the same type. The first ambiguous term of the phrase is resolved as E19.Physical_Object and the last term as material. For example, in the phrase "sherds of Iron Age pottery", "sherds" is annotated as physical object and pottery is resolved as material.

```
({Mention.type=="Physical_Object",
Mention.multisense=="true"}|{E19}):m1
({Token.category == IN}{E49})
({Mention.type=="Material", Mention.multisense=="true"}|
{E57}):m2
```

Grammar II: An ambiguous term of type Physical Object or non-ambiguous term of the same type followed by the string "of" or the string "made of", followed by an ambiguous term of type Material or non-ambiguous term of the same type. The first ambiguous term of the phrase is resolved as E19.Physical_Object and the second ambiguous term as E57.Material. For example in the phrase "artefacts made of wood", "artefacts" is annotated as physical object and wood is resolved as material.

```
({Mention.type=="Physical_Object",
Mention.multisense=="true"}|{E19}):m1
({Token.string=="of"}|({Token.string=="made"}
{Token.string=="of"}))
({Mention.type=="Material", Mention.multisense=="true"}|
{E57}):m2
```

Grammar III: A Time Appellation followed by an ambiguous term of type Material, followed by a noun or a proper noun. The rule resolves the ambiguous term as E57.Material. For example in "Roman pottery tile", "pottery" is resolved as material.

```
({E49})
({Mention.type=="Material",
Mention.multisense=="true"}):match
({Token.category ==NN}|{Token.category ==NNS}|
{Token.category ==NNP}|{Token.category ==NNPS})
```

Grammar IV: The grammar is a variation of the previous rule targeted at an ambiguous term of type Physical Object, instead of Material, which however is not followed by a noun or a proper noun. For example in the phrase "6th century pottery, at Puddlehill", "pottery" is resolved as physical object.

```
({E49})
({Mention.type=="Physical_Object",
Mention.multisense=="true"}):match
```

Grammar V: A determiner (a/an/the/some/few, etc.) followed by an ambiguous term of type Material, followed by a noun or a proper noun. The grammar resolves the ambiguous term as E57.Material. For example in the phrase "data that can be gained from the animal bone assemblage", "animal bone" is resolved as material.

```
({Token.category ==DT})
({Mention.type=="Material",Mention.multisense=="true"}):match
({Token.category ==NN}|{Token.category ==NNS}|
{Token.category ==NNP}|{Token.category ==NNPS})
```

Grammar VI: The grammar is a variation of the previous rule targeted at an ambiguous term of type Physical Object, instead of Material, which though is not followed by a noun or a proper noun. For example in the phrase "The top of the animal bone in pit was also observed at this level", "animal bone" is resolved as physical object.

```
({Token.category ==DT})
({Mention.type=="Physical_Object",
Mention.multisense=="true"}):match
```

Grammar VII: An adjective or a passive voice verb followed by an ambiguous term of type Material, followed by a noun or a proper noun. The rule resolves the ambiguous term as E57.Material. For example in the phrase "two well-stratified brick pieces" "brick" is resolved as material.

```
({Token.category ==JJ}|{Token.category ==VBN}|
{Token.category ==VBD})
({Mention.type=="Material",
Mention.multisense=="true"}):match
({Token.category ==NN}|{Token.category ==NNS}|
{Token.category ==NNP}|{Token.category ==NNPS})
```

Grammar VIII: The grammar is a variation of the previous rule targeted at an ambiguous term of type Physical Object, instead of Material, which however is not followed by a noun or a proper noun. For example in the phrase "The culvert was made of red hand clamped bricks measuring 0.07m", "bricks" is resolved as a physical object.

```
({Token.category ==JJ|{Token.category ==VBN}|
{Token.category ==VBD})
({Mention.type=="Physical_Object",
Mention.multisense=="true"}):match
```

5.7 Adjectival Expansion and Conjunction

The stage of adjectival expansion and entity conjunction has a dual purpose. It is aimed at enhancing existing annotations by combining entities of the same type in larger spans which are utilised by later stage of the pipeline, such as the negation detection stage, as well as expanding over phrasal moderators that add meaning. Expanded annotations support the user-centred focus of the semantic indexing and carry additional information about entities, for potential use during document retrieval. The stage invokes rules that make use of part of speech input and of simple patterns which deliver the enhanced and expanded annotation spans.

5.7.1 Adjectival Expansion

The adjectival Expansion stage is targeted at all entity types (Physical Object, Place, Material and Time Appellations). The stage mainly utilises part of speech (POS) input and gazetteer listings which are specifically used in the expansion of Time Appellation entities. The expansion technique is focused on enhancing existing entities with lexical moderators but it is not intended, within the scope of the thesis, to parameterise the annotations with moderators. This means that the lexical inputs that are used by the expansion pattern do not have unique terminological references which can be utilised by a semantic retrieval mechanism but instead they become part of the expanded annotation. For example the expanded annotation cases "burned pottery" and "broken pottery" would enjoy the same terminological and ontological reference since the expansion mechanism does not distinguish "burned" from "broken" but only expands to include the two different moderators. However, because the expanded versions became available as annotation spans the moderators can be potentially negotiated by an information retrieval scenario.

The rules of adjectival expansion expand the entities with their immediate linguistic moderator, if such moderator exists. Immediate is considered to be the main, most close to the entity, moderator. Therefore, if an entity enjoys more than one moderator the rules are designed to expand only to a single moderator. For example in the phrase "several worked flints" only the word "worked" is included by the expansion rule. There are some rare cases where a gazetteer entry is already making use of a moderator, for example "linear ditch", which is a different entry from "ditch", having a distinct SKOS terminological resource. In such cases, if a moderator exists then the annotation will look as if expanded beyond a single word.

Although the stage is named Adjectival Expansion, the rules do not consider only adjectival moderators but also match passive voice verbs and, in the case of Physical Object and Place entities, they also match cardinal numbers. In addition, in the case of Time Appellation, the rules make use of the complementary gazetteer listing for matching prefix and suffix parts of periods.

A full list of the part of speech Hepple tagger categories that are used by the rules below can be found in Appendix F1. The grammars for addressing the task of Adjectival Expansion are the following:

Grammar I: An adjective, a passive voice verb, or a cardinal number is preceding a E19.Physical_Object entity. The grammar matches the immediate moderator of the entity and includes it within an expanded entity annotation span. For example "worked flint", "ten sherds", "red brick" etc.

```
({Token.category=="JJ"}|{Token.category=="VBD"}|
{Token.category=="VBN"}|{Token.category=="CD"):A_Part
({E19}):B_Part
```

Grammar II: An adjective, a passive voice verb, or a cardinal number is preceding a E53.Place entity. The grammar matches the immediate moderator of the entity and includes it within an expanded entity annotation span. For example, "three layers", "alluvial deposits", "uncovered structures" etc.

```
({Token.category=="JJ"}|{Token.category=="VBD"}|
{Token.category=="VBN"}|{Token.category=="CD"):A_Part
({E53}):B_Part
```

Grammar III: A period prefix gazetteer entry precedes a E49.Time_Appellation entity, or a period suffix succeeds a E49.Time_Appellation entity, or both a period prefix and a period suffix combine with a E49.Time_Appellation entity. The grammar matches any of the above three cases and includes prefix and/or suffix terms within the expanded boundaries of an E49 annotation span, for example, "late Roman", "Roman period" and "late Roman period". Note that as with the previous expansion cases, Time Appellation related gazetteer entries might already contain a prefix operator, as in the case of "Post Medieval", which is a different entry from "Medieval", having a distinct SKOS reference. The expansion rule can expand to include prefixes of such gazetteer entries, as for example "Late Post Medieval".

```
(({Lookup.majorType=="Period Prefix"})?):Prefix
({E49}):TimeAppellation
(({Lookup.majorType=="Period Suffix"})?):Suffix
```

Grammar IV: An adjective or a passive voice verb is preceding a E57.Material entity. The grammar matches the immediate moderator of the entity and includes it within an expanded entity annotation span. For example, “large stone”, “tempered fabric” etc.

```
({Token.category=="JJ"}|{Token.category=="VBD"}|
{Token.category=="VBN"}:A_Part
({E57}):B_Part
```

5.7.2 Entity Conjunction

The rules of this stage address two distinct cases of entity conjunction. The first case concerns only Time Appellation entities, which are conjunct as a single time period span, for example “Iron Age to Early Roman”. In such cases, the rules produce one single “unified” annotation span that includes both previously defined Time Appellation entities. As a result, the new conjunct annotation overwrites the two single annotations, while it makes use of the same annotation type E49, which inherits both SKOS terminological references of the overwritten annotations. Hence, the above conjunct “unified” annotation can be retrieved by using any of the two applicable terminological references, either the SKOS concept of “Iron Age” or the concept of “Early Roman” in the example.

The second case of conjunction rules concerns all entity types (Physical Object, Place, Material and Time Appellations). It is targeted at conjunction phrases that make use of the words “and”, “or” and of “commas”, which are found between two or more entities of the same type. The annotation result of such conjunction cases is treated by the rules as non-unified, which mean that the conjunct entities are not overwritten. Instead both conjunct entities and the unique terminological references are maintained, while a new annotation type is produced which follows a naming convention that combines the conjunct entity type with the word “conjunction” (i.e. E19Conjunction, E49Conjunction, E53Conjunction, E57Conjunction).

The main purpose of this new type of annotation is to support the CRM-EH phase of the pipeline that follows the NER phase rather than to expose unified conjunction spans. Hence, the naming convention diverges slightly from CRM convention since the conjunction annotation type is only used internally by the OPTIMA pipeline. Terms which are syntactically conjunct with “and”, “or”, and “commas” are not assumed as a single entity. For example, the phrase “pottery, arrowhead and coins” refers to three different objects.

As fully discussed in a chapter 6, the CRM-EH specialisation phase employs elaborate patterns, which match CRM entities within the context of rich discussion phrases. For

example, the above objects might be mentioned in the phrase "ditch containing pottery, arrowhead and coins". In such a case, the CRM-EH specialisation should detect that all three objects are found in the same ditch. Thus, defining conjunct entities of the same type, allows the definition of simpler CRM-EH specialisation patterns, which do not have to address cases of conjunction between same type entities.

The grammars which address the task of Entity Conjunction are the following:

Grammar I: Two E49.Time_Appellation entities conjunct with punctuation other than a full-stop, or a comma, or conjunct with a maximum of two words other than "and" and "or". The grammar unifies the two Time Appellation entities under a single annotation span that carries two distinct SKOS terminological references. For example, the phrases "Medieval/Post Medieval", "Late Bronze Age – Iron Age", "Norman to Medieval" and "Mesolithic to the Post Medieval" produce unified Time Appellation spans which carry two SKOS references of the conjunct terms.

```
({E49}):TimeAppellation
({Token.kind == "punctuation", !Exclude})?
({Token.kind == "word", !Exclude})[0,2]
({E49}):TimeAppellation2
```

Note: the "!Exclude" directive is used for excluding the "comma", "full-stop", "and", and "or" cases.

Grammar II: A Time Appellation (E49) entity (which might be followed by a punctuation other than a full-stop, or a comma, or might be followed by a maximum of two words other than "and" and "or") is followed by an Ordinal and a Suffix Lookup, for example, "Post-Medieval to 19th century". The grammar produces a large annotation span including both Time Appellation and Ordinal and Suffix spans. When the latter spans do not enjoy unique terminological resources because they originate from flat gazetteer listings, the final annotation span has a single SKOS reference which originates from the Time Appellation (E49) entity.

```
({E49}):TimeAppellation
({Token.kind == "punctuation", !Exclude})?
({Token.kind == "word", !Exclude})[0,2]
(({Lookup.majorType == "Ordinal", !Entity}{Lookup.majorType
== "Period Suffix", !Entity})
```

Grammar III: This is a reverse version of the above grammar. Thus a Prefix period listing or an Ordinal (which might be followed by a punctuation other than a full-stop, or a comma, or might be followed by a maximum of two words other than "and" and "or") is followed by E40.Time_Appellation entity, for example, "early to middle Iron Age".

```
(({Lookup.majorType == "Ordinal",!Entity}|
{Lookup.majorType == "Period Prefix", !Entity})
({Lookup.majorType == "Period Suffix",
!Entity})?):TimeAppellation
({Token.kind == "punctuation",!Exclude})?
({Token.kind == "word",!Exclude})[0,2]
({E49}):TimeAppellation2
```

Grammar IV: A E49.Time _Appellation entity conjunct with one or more Time Appellation (E49) entities via "and", "or", and "comma", for example, "Iron Age, Roman and Medieval period". The grammar produces a new annotation span of type E49Conjunction which includes all E49 entities.

```
{E49}
(({Token.string =="and"}|{Token.string =="or"}|
{Token.category ==","})({Token.category=="DT"}|
{Token.category=="RB"})?
{E49})+
```

Note the use of the plus operator (+) for matching one or more conjunct terms. Also the grammar allows for matching adverbs or determiners that can be used between conjunct terms, for example, "Iron Age, the Roman period and possibly Medieval date".

Grammar V, VI, VII: The grammars are variations of the above IV grammar for matching conjunction phrases of E19.Physical_Objects, E53.Places and E57.Material entities. The rules produce new annotation spans of type E19Conjunction, E53Conjunction and E57Conjunction respectively. The only difference to grammar IV above is that the current grammar allows matching of materials which describe objects (together with adverbs and determiners) in the case of E19.Physical_Object conjunction, for example "brick, coins and iron arrowheads".

```
{E19}
(({Token.string =="and"}|{Token.string =="or"}|
{Token.category ==","})({Token.category=="DT"}|
{Token.category=="RB"})?
({E57})?
{E19})+
```

Note the use of ? operator for denoting that a material entity might be present.

5.8 Negation Detection

Negation is an integral part of any natural language system. It is a linguistic, cognitive and intellectual phenomenon, which enables the users of a language system to communicate erroneous messages, the truth value of a proposition, contradictions, irony and sarcasm (Horn 1989). From a philosophical stance, Aristotle defines negation as a system of oppositions between terms, such as *contrariety* (e.g. black vs white), *contradiction* (e.g. a modern antique), *privation* (e.g. dry vs wet) and *correlation* (whole vs half). Other philosophers, from Plato to Spencer Brown have independently approached negation as a *heteron* (other) not as *enantion* (contrary), defining negation as a "*positive assertion of the existence of a relevant difference"* (Westbury 2000)

The fact that "*relevant difference*" requires choice brings the study of negation beyond the limits of linguistics and into the study of human behaviour. A framework for ranking the complexity of negation in natural language classifies negation into 6 forms, from the simplest form "*negation as rejection*" to the most complex "*propositional negation*", which assumes fully-developed language production and comprehension. The in between classes from the less complex to the most complex form of negation are described as (i) *negation as a statement,* (ii) *negation as an imperative,* (iii) *negation of a self-generated or planned action, and* (iv) *scalar negation*" (Westbury 2000).

Whereas the topic of negation in natural language has been examined for millennia, enjoying much study and publication, computational negation methods are a rather recent study topic. Commercial keyword-based search engines and contemporary information retrieval systems allow the Boolean negation operator for specifying search queries but their ability to address negation at the natural language level is rather limited. A search keyword for example which contains the phrase "evidence of" is very much likely to return negation results relating to "no evidence of".

Attempts to address the issue of negation in an information retrieval context have been focused on the disambiguation of user queries. McQuire and Eastman (1998) describe a system which detects ambiguous queries and asks the user for clarification. The user is prompted with a list of choices for clarifying those elements of the search query that are negated. Other attempts originate from the medical domain where negation is used to describe many important facts about medical conditions. Rokach et al. (2008) and Mutalik et al. (2001) describe information retrieval systems which mainly employ machine learning methods for detecting negation in medical text.

In the context of Information Extraction, the task of negation is mainly employed by applications, which are targeted at opinion mining and sentiment analysis (Maynard & Funk 2011, Popescu & Etzioni 2005). Sentiment detection techniques can be divided into rule-based (lexicon-based) and machine-learning methods, broadly following the earlier classification of general information extraction methods. Discussion of the literature on sentiment analysis applications is not within the scope of this thesis.

Two specific opinion mining projects are mentioned as examples of the wide range of sentiment analysis projects that are available today. The OPINE system (Popescu & Etzioni 2005) is an unsupervised machine-learning system built on the KnowItAll web information-extraction system, which extracts opinions from reviews for building a model of important product features. The system addresses negation by detecting the semantic orientation and polarity for the lexical head of the various opinion phrases identified by the machine-learning algorithm.

Maynard and Funk (2011) describe a rule-based method for detecting political opinion in Tweets. The system uses GATE and a range of precision focused, hand-crafted rules for extracting triples of the form "<Person, Opinion, Political Party> e.g. <Bob Smith, pro, Labour>", The system addresses negation via a gazetteer list, which contains negative words such as "not", "couldn't", "never", etc. Whenever, such a negative word is detected the opinion is reverted from "pro" to "anti" and vice versa.

Both systems have reported the address of negation detection with some success, with the rule-based system enjoying better Precision than Recall and the machine-learning system delivering high precision and recall regarding customer opinions and their polarity.

5.8.1 The Role of NegEx Algorithm for Identifying Negated Findings

NegEx (Chapman et al. 2001) is a specific algorithm targeted at the identification of negated findings in medical documents. The algorithm determines whether Unified Medical Language System (UMLS) terms of findings and diseases are negated in the context of medical reports. In such reports, there is an extensive use of negation utterances which clarify the presence of findings and conditions. Being able to identify negated terms in a medical context is highly desirable for the performance of retrieval systems serving the domain (Rokach et al 2008, Chapman et al. 2001, Mutalik et al. 2001). The NegEx is one of the many available negation detection algorithms operating in the medical domain and is particularly relevant to the scope of the OPTIMA negation detection, due to its rule-based design, the use of pattern matching mechanism and the employment of vocabulary listings.

The design of the algorithm is based on the use of offset patterns that utilise negation related vocabulary. The vocabulary contains terms and phrases that denote negation, which are invoked by a set of rules. In detail, the algorithm makes use of two specific patterns. The first pattern identifies negated UMLS terms in phrases which commence with a negation phrase that is followed by up to five tokens before a UMLS term, i.e. <negation phrase> * <UMLS Term>. The second pattern is a reversed version of the above, which matches negated UMLS terms which are up to five token prior to a negation phrase, i.e. <UMLS Term> * <negation phrase>. The asterisk indicates that up to five tokens may fall between the negation phrase and UMLS term.

The algorithm employs 24 negation words and phrases, such as "no", "without", "absence of", etc., for supporting the operation of the first pattern and 2 words, "declined" and "unlikely", to support the operation of the second pattern. In addition, the algorithm makes use of a third list of 10 terms which are named "pseudo-negation phrases". Such phrases carry false negation triggers and double negatives like "without any further", "not certain if", "not rule out", etc. which diminish the meaning of a negation. Upon matching such phrases an identified negation is bypassed.

There are two main parallels which support the adoption of the NegEx approach by the OPTIMA negation mechanism. Firstly, the use of pattern matching rules and vocabulary terms allows a smooth integration of the algorithm within OPTIMA cascading order of the pipeline. The second is the good performance of the algorithm in detecting negations about findings. In archaeological reports, as in medical reports, authors frequently negate facts about findings. For example in the medical domain, a typical phrase might be "extremities showed no cyanosis (medical)", whereas in the archaeological domain a negation case might be "there is no evidence of Roman pottery (archaeological)". This particular parallel formed the foundation for adopting and generalising the NegEx algorithm in the context of archaeological reports.

The NegEx adaptation by the OPTIMA pipeline also improves some reported limitations of the algorithm. The algorithm is reported to have a limited performance with regards to conjunct terms and to falsely negating terms, due to its limitation on correctly adjusting the scope of a negation phrase. For example, negating both UMLS terms of the following phrase "no *cyanosis* and positive *edema*", or failing to negate a conjunct term which exceeds beyond the threshold of 5 tokens. The adapted OPTIMA version addresses the above issues by introducing additional vocabulary to the algorithm and by utilising the defined entity conjunction spans. Also the OPTIMA version does not make use of the

"pseudo-negation phrases" that did not bring the anticipated results when applied but instead prevented the adapted OPTIMA algorithm from negating valid negation cases.

5.8.2 Negation Detection of the OPTIMA Pipeline

The OPTIMA pipeline implements a negation detection mechanism, which is based on the technique of NegEx algorithm for using specialised vocabulary in combination with phrasal offset. The implemented mechanism enhances NegEx vocabulary and improves the algorithm for addressing known limitations and domain related issues. The vocabulary is enhanced with additional terms and with new listings, while the algorithm is modified to enable negation detection within the context of archaeology reports.

In detail, the OPTIMA negation detection mechanism adopts the two vocabulary listings, which are utilised by the pattern matching of phrases that commence or end with a negation term/phrase. The list which supports the commencement of negation phrases is given the name *Pre-negation* and the list which supports matching of phrases that end a negation description is named *Post-negation.* The Pre-negation listing adopts all NegEx vocabulary of the similar list and enhances it with additional terms. This enhancement process intellectually utilised WordNet synsets (synonym relationships of the existing NegEx vocabulary), in order to identify additional terms that could be included in the list. The enhancement was also based on simple English grammar syntax rules and empirical knowledge relating to negation. The process populated the list with 15 additional terms. The same technique was followed for enhancing the Post-negation listing. The complete set of terms of both listings is made available in [Appendix B].

The negation mechanism also utilises two additional listings which are not part of the NegEx algorithm, the *Stopclause-negation* and the *Negation_verbs* list. The *Stopclause-negation* list aims to help the negation detection mechanism overcome the known limitation of the NegEx approach in addressing larger than 5 Tokens negation phrases. The Negation_verbs list aims to enhance the pattern matching mechanism with a set of verbs, which when combined with a negative moderator, denote lack of evidence. Both listings aim to enable the negation algorithm to address the writing style of archaeological reports.

While medical reports are written in a very clear and concise manner, archaeological reports tend to follow a more creative writing style than medical reports. Hence negation of facts and findings might be found beyond the threshold of 5 tokens, or negations might be expressed in passive voice. For example; "absence of any other occupation evidence such as structures" or "deposits were not encountered at the machined level" respectively. The

construction of the Stopclause-negation and Negation-verbs listings is based on the outcome of a bottom-corpus analysis described below.

5.8.3 Negation Detection Corpus Analysis

The general aims of the bottom up corpus analysis was to reveal vocabulary evidence which could be used by the negation detection mechanism (a) to address the known limitations of NegEx regarding the length of negation span and (b) to improve adaptation of the algorithm in the context of the writing style of archaeological reports. In order to reach the above objectives, the corpus analysis exercise evolved into two successive information extraction stages.

The first stage extracted from a volume of 2460 archaeological reports, phrases of a maximum of 11 tokens which contained negation moderators and EH Thesauri terms. This particular extraction approach was based on the assumption that larger spans would reveal enough evidence of contextual vocabulary, which could be analysed and then refined into gazetteer listings. Such gazetteer listings could then be invoked by JAPE grammars for improving the performance of the negation algorithm, with regards to negation spans and the writing style of archaeological reports.

The extraction of large negation spans was based on the use of the *Pre-negation* and *Post-negation* gazetteer listings, in combination with two simple matching grammars as seen below.

Grammar I

```
({Token.string!="."})[0,5]
{Lookup.majorType == PreNeg}
({Token.string!="."})[0,5]
```

Grammar II

```
({Token.string!="."})[0,5]
{Lookup.majorType == PostNeg}
({Token.string!="."})[0,5]
```

The rules are almost identical; they only differ on the listing type which they invoke, either *PreNeg* for the Pre-negation gazetteer list, or *PostNeg* for the Post-negation listing. The rules translate as: match a span which expands 5 tokens before a gazetteer match and 5 tokens after a gazetteer match excluding full stops, in order to prevent the rule expanding beyond the limits of a sentence.

A succeeding rule was invoked for filtering out those phrases matched by the above rules but not containing any EH Thesauri terms. The filtering process was based on the

principle of selecting phrases that were closer to the four main entities targeted by the OPTIMA pipeline. The four EH Thesauri resources relate to the four main entities targeted by the pipeline. Matching only the phrases which contained at least one thesaurus entry supports the assumption that the phrase relates to at least one of the four CRM entities. Although, this approach might be somewhat simplistic, it satisfies a basic criterion for selecting phrases which have a strong potential for carrying instances of CRM entities. The filtering grammar implements a simple matching condition which is carried out by the "contains" JAPE operator as shown below.

Grammar III

```
{NegSentence contains Lookup.majorType=="Monument Type"}|
{NegSentence contains Lookup.majorType=="MDA ObjectType"}|
{NegSentence contains Lookup.majorType=="Building Material"}|
{NegSentence contains Lookup.majorType=="Timeline"}
```

The rule translates as: match all previously identified negation phrases (NegSentence) which contain a term (Lookup) from any of the four main EH Thesauri gazetteers (Monument Type, MDA Object Type, Building Material and Timeline). The filtering rule has managed to identify 15732 phrases which made use of a negation operator found in the two listings (Pre-negation and Post-negation) containing at list one EH Thesauri entry.

The second stage implemented a simple extraction pipeline which processed the 15732 negation phrases identified by the first stage. The pipeline of the second stage used a cascading pipeline of the following order a) Tokeniser b) Part of Speech Tagger c) Noun Phraser and e) Verb Phraser. The aim of this particular pipeline was to reveal from the volume of negation phrases the most commonly occurring Noun and Verb phrases. Such commonly appearing phrases were then examined further to decide whether they should be intellectually selected as specialised vocabulary, which could be utilised by the negation detection algorithm.

In total, 29040 noun phrases and 14794 verb phrases were identified. From them 14686 were unique noun phrases and 2564 were unique verb phrases. As might be expected, the most commonly occurring noun phrase was "No" with 1027 hits and the most commonly occurring verb phrase was "is" with 914 hits. Examining the list of noun phrases there was not an obvious observation that could be made since there was a large volume of unique noun phrases (about half of the list), while the commonly occurring phrases did not present any particular interest. Among the most common noun phrases, apart from "it", "which", "there" etc, were the phrases "no evidence", "the absence", "the lack", and "no find", which were already part of the *Pre-negation* list. Thus, there was not a clear emerging

pattern of noun phrases which could be used as specialised vocabulary to enhance the already defined gazetteer vocabulary and the noun phrase list was not utilised further.

On the other hand, the verb list revealed some very interesting vocabulary patterns. By examining the most commonly occurring verbs, a pattern emerged relating to passive voice phrases. For example the phrase "should not be considered" occurred 134 times, the phrase "was not excavated" 67 times, the phrase "were not encountered" 39 times, etc. Although, NegEx covered some cases of "backward" pattern matching via the *Post-negation* vocabulary for phrases where a gazetteer entry is found at the end of a negation phrase, the algorithm did not consider the use of passive voice utterances. This is possibly due to the direct style of medical report writing. However, the corpus analysis revealed a range of passive voice phrases that are found frequently in archaeological reports.

An intellectual examination of the list of the frequently occurring verb phrases isolated a set of passive voice verbs that could be used to negate archaeological findings and facts. The lists of verbs [Appendix B3] constitute a specialised vocabulary of 31 entries that was defined as a gazetteer listing, with the name *Negation-verbs*. The gazetteer resource was then used by pattern matching rules, which are discussed below, for identifying negation in phrases such as "deposits were not encountered at the machined level", "further postholes were not observed in the field", "ridge and furrow was not recorded during the present exercise", etc.

The corpus analysis of negation phrases revealed the use of passive voice verbs as a writing style method of negation which is frequently found in archaeological reports. However, one of the objectives of the analysis was to reveal specialised vocabulary, which could be used for the expansion of the negation span beyond the threshold of 5 tokens. The noun phrase analysis did not reveal any evidence which could have been used in that direction. By examining closer the bulk of 15732 negation phrases, it was observed that long sentences introduce new clauses when they make use of particular vocabulary, such as "but", "and", "however", etc. and punctuation such as "comma". For example, "no set of foundation postholes but other archaeological remains were observed", "no floor which was removed, other underlying features could be identified", "not rebuilt, although substantial traces", etc.

A specialised gazetteer vocabulary was defined [Appendix B4], which is used by the negation rules for controlling the negation span. The entries of the specialised vocabulary were selected by using WordNet and looking for the words "and", "but", "however" and "although" and following the links to sister and synonym terms. The gazetteer resource

named *Stopclause-negation* contains 38 entries. Although, the resource contains the word "and", it does not prevent rules from negating conjunct entities, since negation rules match both single CRM and conjunct CRM entities. For example, in the phrase "no evidence of pottery and tile" both "pottery" and "tile" are negated.

5.8.4 Negation Detection Rules

The stage of negation detection of the OPTIMA pipeline utilises the four gazetteer listings, Pre-negation, Post-negation, Negation-verbs and Stopclause-negation, together with a range of pattern matching rules targeted at the negation of the CRM entities, Place, Physical Object, Material and Time Appellation. This stage uses a set of three pattern matching rules which are varied four times. Each variation corresponds to a different CRM entity.

The arrangement of the negation rules is such as to avoid negation of the same phrase more than once, even if more than one CRM entities are mentioned. Based on the use of intermediate negation annotation types which are aggregated under a final negation annotation, the pipeline delivers a single negation span, which covers all CRM entities involved in a phrase. For example in the phrase "no evidence of Roman pottery", the pipeline delivers a single negation that spans the whole phrase, not two annotations one spanning until "Roman" and another until "pottery". Similarly when conjunction of entities is present, the negation span covers all conjunct entities under a single annotation span, for example "no evidence of Roman pottery and tile".

The JAPE grammars which address the task of Negation Detection are the following:

Grammar I: A Physical Object entity or conjunct Physical Object terms which are followed by up to 10 Tokens, which do not contain a stop word (StopNeg), which are followed by "not" which might be followed by no token or one token, followed by a Lookup annotation of Negation-verbs gazetteer list, for example, "pottery and tile remains were not observed".

```
({E19}|{E19_Conjuction})
({Token,!StopNeg})[0,10]
{Token.string=="not"}
({Token})?
{Lookup.majorType==VerbNeg}
```

Note the use of the span [0,10] zero to ten tokens which are not StopNeg. This particular annotation type (StopNeg) is defined by a rule which matches Lookup from the *Stopclause-negation* gazetteer listing, or a punctuation, or a symbol character, which are considered as elements that terminate a clause sentence.

Grammar II: A Lookup annotation of the Pre-negation gazetteer list followed by up to 10 Tokens which do not contain a stop word (StopNeg), which are followed by a Physical Object entity or conjunct Physical Object terms, for example, "absence of any datable small finds or artefacts".

```
{Lookup.majorType==PreNeg}
({Token, !StopNeg})[0,10]
({E19}|{E19_Conjuction})
```

Grammar III: Physical Object entity or conjunct Physical Object terms which are followed by up to 10 Tokens which do not contain a stop word (StopNeg), which are followed by a Lookup annotation of the Post-negation gazetteer. For example, "wares such as tea bowl are particularly unlikely to exist".

```
({E19}|{E19_Conjuction})
({Token, !StopNeg})[0,10]
{Lookup.majorType==PostNeg}
```

Note, the above three rules exist in four variations, each targeted at one of the four CRM entities. The only line of code which changes from the above grammars is the one that matches E19 or E19_Conjuction annotations. Changing this line to E49, E53, E57 alters the rules to match Time Appellation, Place and Material entities respectively.

The last step of the Negation Detection stage is to discard all previously identified CRM entities which belong to negation phrases. The pipeline line uses four Negation Filter rules, each targeted at one of the four CRM entities. The grammar is simple and makes use of the "within" JAPE operator. All matches are removed (filtered out) for the Annotation Set.

Grammar IV: Match a Physical Object entity which is within a Negation annotation

```
{E19 within Negation}
```

Note the E19 instance is removed by the RHS part of the JAPE rule. Altering the above rule to E49, E53 and E57 adjusts the rule to the remaining CRM entity types.

5.8.4.1 Special Cases of Negation Detection (Well and Post)

The negation detection stage utilised an additional set of rules which are targeted at negating the special cases of "well" and "post" which widely used in archaeology reports and proved problematic due to polysemy. Negation of these cases actually refers to their sense disambiguation but since the non-desired senses of terms are discarded, this particular form of disambiguation is treated by convention as a negation detection task. Both terms are polysemous (having more than one meaning), which makes them ambiguous and prone to deliver false positive matches.

Well for example can refer to a shaft into the ground from which water can be obtained but also "well" is used as an adverb ("to be treated well"), as an adjective ("well preserved") or as an interjection ("Well, who is next"). On the other hand, "Post" can refer to a structural element of a building like "a four post structure", synonym to pillar and pile, or can refer to a position of duty "managerial post", or to the mail post, or can be used in a time definition context, i.e. "post excavation, post war, post modern etc".

With regards to "well", part of the disambiguation process is handled via the noun phrase validation stage. Hence, "well" senses which fall in the category of adverb, for example "as well", are filtered out by the validation stage. However, the adjective senses of the term such as "well preserved, well-situated etc" remain as part of the noun phrases and thus can be confused with the sense of "shaft", producing false positive matches. Two specific rules were defined to overcome the above problem and to negate the non desired senses.

Grammar V: Match all those cases of "well" which are followed by a past tense verb or a hyphen. For example "well arranged" or "well - supported". The rule annotates the matches as negated phrases. The instances of "well" which are contained within such phrase are removed from the Annotation Set.

```
({Token.string=="Well"}|{Token.string=="well"})
({Token.category==VBN}|{Token.string=="-"})
```

Rule VI: Match all Tokens which contain "well-" or "Well-". This is a simple rule for matching all those tokens that make use of hyphenated "well" without using spaces. For example, "well-supported, well-arranged". The rule assumes that hyphenated "well" cases do not correspond to the Place entity sense.

```
{Token.string=~"[Ww]ell-"}
```

With regards to the non desirable sense of "post", the use of the noun phrase validation did not help since "post" is always recognised by the POS tagger as a noun. The term is contained in the Time Prefix gazetteer list and is part of time related Lookup annotations, such as "Post Medieval". Hence, any other sense of "post" which does not relate to a Time Appellation entity is discarded, with the exception "Post-hole(s)", which is the only sense of "post" which can be found in a Place entity. The rule is simple and implements the above assumption.

Rule VII: Match all cases of Place entity which are equal to the hyphenated version of "Post-hole(s)" or "post-hole(s)" which are synonym to "posthole". These are the only valid cases of a place entity sense that are allowed to contain "post".

```
{E53, Token.string =="[Pp]ost-hole"}|
{E53, Token.string =="[Pp]ost-holes"}
```

Note all other cases of Place entity containing "post" which do not fall under the above rule are removed from the Annotation Set.

5.9 Summary

The chapter concludes a major part of the information extraction effort delivering named-entity output with regards to the four CRM entities E19, E49, E53 and E57.Material. The NER phase of the OPTIMA pipeline used an innovative technique of exploiting gazetteer listing via three distinct thesaurus relationships; synonym, narrower and broader concept. This enables the IE system to gain a flexible control over the volume of vocabulary that contributes to the NER task. In addition, the technique enables JAPE grammars to invoke input from gazetteer listings using SKOS enabled features and to deliver semantic annotation output that is enhanced with terminological as well as with ontological references.

The chapter also discussed word sense disambiguation techniques and in particular disambiguation between Physical Object and Material senses that influenced by ontological definition. Thus, this form of ambiguity is coined as *ontological polysemy*. The technique is based on the use of linguistic patterns for resolving the appropriate sense to polysemous terms while assigning the corresponding SKOS reference. Also Lookup validation techniques via noun-phrase input were revealed, while adjectival expansion techniques based on part of speech input were employed for enhancing the span of semantic annotations. Moreover, the NER pipeline implemented a negation detection module capable of identifying negation phrases in archaeological text using an adjusted version of the NegEx algorithm combined with vocabulary listings.

The aforementioned NLP techniques of Adjectival Expansion, Word Sense Named Entity Recognition Disambiguation and Negation Detection are novel contributions in the archaeology domain. The discussion concludes the NER effort, the following chapter discusses the CRM-EH specialisation phase of the pipeline using event detection techniques.

Chapter 6

Relation Extraction with CRM-EH

6.1 Introduction

The chapter discusses the second phase of the OPTIMA pipeline targeted at the detection and recognition of relations between CRM entities previously identified by the NER phase. The role and background information with regards to the task of Relation Extraction (RE) is revealed. The discussion highlights relevant literature findings and reveals the details of a RE task driven by the ontological arrangements of the CRM-EH model. In particular, the scope of rule-based RE task is discussed with regards to a set of selected CRM-EH event and property entities. The results of a corpus analysis study are presented which informed the construction of the hand-crafted JAPE grammars of the RE pipeline. The significant contribution of Zipfian distribution principles in the analysis of the corpus "bottom-up" data, which led to the abstraction and formulation of RE JAPE grammars, is also revealed. The discussion also presents the total range of the JAPE grammars that are employed by the pipeline for extracting three CRM-EH event and one property entities.

6.2 Information Extraction of Relations and Events

Early attempts at the evaluation of relation extraction and event detection applications can be traced back to the Message Understanding Conference (MUC 7) 1997. The conference called for the extraction of 3 relations (employee_of, product_of, location_of) and 1 event (air vehicle launches). The successor of MUC, the Automatic Content Extraction (ACE) programme, formulated the task of relation detection and recognition during its second phase that commenced in 2003. The evaluation programme defined relation extraction as an inference task addressing explicit relations between two entities that occur within a common syntactic construction. The participating entities of a relation are called arguments. In total five relation types (At, Near, Part, Role and Social) and twenty four sub-types were included by ACE 2003 (US-NIST 2003).

The Event Detection and Recognition task was addressed a year later by ACE 2004 (US-NIST 2004). The evaluation programme defined event detection as an inference task between zero or more entities, values and time expressions mentioned in the source text.

The entities involved in an event are called participants and every participant was characterised by a role. The programme evaluated the recognition of five event types; destruction/damage, creation/improvement, transfer of possession or control, movement and interaction of agents. Similarly to relation extraction, event detection was targeted at recognizing syntactic constructions, i.e. phrases or sentences containing the aforementioned event types. The succeeding programmes ACE 2005 (US-NIST 2005) and ACE 2007 (US-NIST 2007) introduced new types of events and enhanced event recognition with attributes, such as type, subtype, modality, polarity, genericity and tense.

6.2.1 Applications of Relation Extraction and Event Recognition

Extraction of semantic relations between entities is a significant step towards the development of sophisticated NLP applications that explore the potential of natural language understanding. As discussed in literature, during the last decade a range of projects have addressed the task of relation extraction and recognition (Bach and Badaskar 2007). Efforts have been mainly directed at the recognition of binary relations between entities employing supervised and unsupervised machine learning methods, with the latter method the less popular. Some influential supervised methods of relation extraction are based on the computation of kernel functions between shallow parse trees and on feature-based relation extraction of lexical, syntactical and semantic knowledge using Support Vector Machines (SVM). A SVM based, GATE application (Wang et al. 2006) employed several features including part-of-speech tag, entity subtype, entity class, entity role, semantic representation of sentence and WordNet synonym sets to address the task of relation extraction with regards to the ACE 2004 relation types.

Byrne (2009) on the other hand, in her thesis discusses a supervised ML application (txt2rdf) that uses the maximum entropy model for detecting archaeological entities (NER) and relations (RE). The application was trained to detect archaeological events that are often mentioned via verb phrases such as “site X was visited on a [date]”, “site Y has been recorded by [an agent]”, etc. Such events are expressed as a collection of binary relations between entities, where each participating entity defines an event attribute, such as agent, role, date, patient and place. For example, in the phrase “The following *were found* in *Unst* by Mr *A T Cluness*: a *steatite dish*”, the term “were found” represents the event, “Unst” is the event location, “AT Cluness” is the event agent and “steatite dish” is the event patient.

Byrne (2009) also discusses detection of inter-sentence relations that expand beyond the limits of a single sentence which do not include cases of co-reference. The detection of

such inter-sentence relations is doable due to the peculiar characteristics of the processed text, such as brief document length (maximum half a page) and synoptic writing style. Some parallels can be drawn with her work mainly because the pipeline performs NER and RE over archaeological text. However, the use of probabilistic ML techniques and absence of ontology, in particular CRM or CRM-EH, yields a significantly different method of IE than the one adopted by the OPTIMA pipeline.

The role of event detection and relation extraction systems has also been explored by applications targeted at the biomedicine domain (Ananiadou et al. 2010). The domain is particularly interested in the detection and deeper semantic analysis of relationships and interactions between biological entities. A number of supervised and rule-based approaches have been applied to the automatic detection of protein interactions, gene regulatory events, etc. Annotation event corpora such as GENIA and BioInfer have been constructed to encapsulate domain knowledge in the form of manual annotations that can be used to train supervised ML systems.

Rule-based systems have been also developed, such as GENIES and GenIE. The first system uses a full parsing strategy combined with sub-language grammar constraints and domain dictionaries, while the later is an ontology-driven system that uses linguistic analysis and semantic representation formalisms to extract information on biochemical pathways and functions of proteins. The above systems demonstrate the capacity of rule-based, ontology guided systems to tackle the task of RE with some success, scoring higher Precision (90-96%) than Recall (53-63%) (Cimiano, Reyle and Saricet 2005; Friedman et al. 2001). However, the capability of rule-based, ontology guided systems to tackle the task of RE in the archaeology domain is not yet fully explored. The following sections discussed the efforts of answering the problem of RE in the archaeology domain by a rule-based, ontology guided system.

6.2.2 CRM-EH Relation Extraction and Event Recognition

The second phase of the OPTIMA pipeline is targeted at detecting “rich” textual phrases that connect, (previously identified by the NER phase) CRM entity types in a meaningful way. The aim of the pipeline is to detect and to annotate such phrases as CRM-EH event or property entities. An initial effort on the detection of such phrases was made by the prototype development, which used rules that identified coexistence of entities in phrases, delivering some encouraging results. However, simple coexistence of entities in phrases does not necessarily constitute detection of relations or events (Ananiadou et al. 2010).

Thus the full scale-system aims to improve the prototype development.

The full-scale pipeline uses hand-crafted rules that employ syntactical structures for the detection of "rich" textual phrases. The extraction method follows a shallow parsing strategy based on the input of part of speech tag, entity types and domain dictionaries. Other projects have also found shallow parsing useful for tackling the task of relation extraction (Zelenko, Aone and Richardella 2003). The annotation technique is influenced by the ACE definition of Relation Detection and Recognition tasks (US-NIST 2004). The pipeline adopts the ACE definition of relation mention as phrases or sentences that express a relation. The binary definition of relation is adopted where each relation phrase consists of two arguments identified by a unique ID and a role

The pipeline also detects phrases that can be modelled as CRM-EH events. However the definition of such events is different to the ACE definition. ACE events involve zero or more entities, values and time expressions, whereas the CRM-EH events which are targeted by the pipeline, connect CRM entities in a binary form. The focus is to detect "rich" phrases which can be modelled as CRM-EH events or properties. Such events or properties can be explicitly or implicitly mentioned in text. Thus, neither the extent nor complexity 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 "[ditch contains {pottery] of the Roman period}" delivers two CRM-EH events. One event connects "ditch" and "pottery" and another event connects the same "pottery" with the "Roman period", both events having "pottery" as a common argument. The first event can be modelled in CRM-EH terms as a *deposition* event (EHE1004.ContextFindDepositionEvent) while the second event can be modelled as a *production* event (EHE1002.ContextFindProductionEvent). Both events are implicitly defined in this example since there is no clear mention of an event that deposited the pottery in the ditch or how the pottery had been produced. However, it can be assumed that since the pottery has been found in the ditch it must have been deposited in that place and since the pottery is described as Roman it must have been produced during the Roman period.

The above modelling technique is significantly different than the technique followed by Byrne and Klein (2010). Primarily, the event detection adopts a rule-based IE technique instead of a ML, thus it does not require a training set. Also the event detection is driven by the CRM-EH ontological arrangements and event mentions are phrases or sentences. In

contrast the Byrne and Klein (2010) approach detects events as mentions of verb phrases carrying a single event type which though might contain more than two arguments. On the other hand, the OPTIMA pipeline detects more than one event type and detects binary relationships between entities which can overlap as discussed in the example above. Thus, events types are specialised and well defined by ontological definitions which can be exploited by retrieval applications.

The same applies to the detection of CRM-EH properties which are specific binary relationships between entities, such as a physical object *consists of* material. Since both CRM-EH event and property detection are driven by binary relationships of entities, their detection and recognition process follows the same IE technique. In practical terms, detection and recognition of CRM-EH events and properties by the OPTIMA pipeline can be understood as an ACE Relation Detection and Recognition task (RDR), with the only difference that the task is specific to the CRM-EH ontology and that relations can be implicitly and explicitly mentioned in source text.

6.2.2.1 The Event-Based models, CRM and CRM-EH

The CIDOC CRM, as discussed in section 2.4.3, is an event-based conceptual model, which aims to enable semantic interoperability of cultural heritage information by providing a framework of metadata that promotes a shared understanding (Doerr 2003). Its extension model, CRM-EH, specialises in the domain of archaeology and describes entities and relationships for a range of archaeological events (Cripps et al. 2004). A core element of both models is the use of events as hubs for connecting together various entities, such as actors, places, objects etc. Thus, the models are defined as event-based or as event-centric. This approach is particularly useful for modelling events in a cultural and heritage setting, either by describing events in their broader sense, such as historical developments, or as recording events of restoration and change of ownership (Shaw et al. 2009).

In the CRM model, *E2 Temporal Entity* is the super class of all event entities, which "*comprises all phenomena, such as the instances of E4 Periods, E5 Events and states"*. It is an abstract class with no direct instances and is specialised into *E3 Condition State*, which "*comprises the states of objects characterised by a certain condition over a time-span*" and *E4 Period* which "*comprises sets of coherent phenomena or cultural manifestations bounded in time and space*". *E4 Period* is specialised further into *E5 Event*, which "comprises changes of states in cultural, social or physical systems, regardless of scale", which is also specialised further into *E7 Activity*, *E63 Beginning of Existence* and *E64 End of Existence* and so forth. Overall, the model defines 35 Event entities, all

descending from the abstract class, *E2 Temporal Entity.*

The CRM-EH model specialises CRM events, entities and properties, by declaring abstractions that are closer to the scope of the archaeology domain. Thus, CRM entities and events are extended to enable integration of heterogeneous archaeological resources. For example, the CRM event *E9 Move* which "*comprises changes of the physical location of the instances of E19 Physical Object*" is specialised into *EHE1004 Context Find Deposition Event,* which defines "e*vents which are often hypothesized to explain how a finds object came to end up in a context*". Similarly, entities such as *E19 Physical Object* and *E57 Material,* are extended to *EHE0009 Contex*t *Find* and *EHE0030 Context Find Material* respectively. As previously discussed (section 2.3.3.1), the CRM-EH comprises 125 extended CRM events and entities.

CRM and CRM-EH also model properties which define specific kinds of binary relationships between conceptual classes. The scope of properties is described by the *scope note* and their usage is declared by the *intention* of use. Every property is defined with a reference to its *domain* and its *range.* As an analogy, consider *property* as a verb, the *domain* as an object and the *range* a subject of a sentence. For example, the property *P45 consists of has* as domain the *E19 Physical Object* class and as range the *E57 Material* class. Properties can be interpreted in passive or active voice, thus they are bidirectional and so it is arbitrary which class is the domain and which is the range of the property. For example *consists of* and *is incorporated in* implement the same P45 property.

6.2.2.2 Scope of the OPTIMA Relation Extraction

The NER phase of the OPTIMA pipeline is targeted at identifying the four CRM entities, E19.Physical_Object, E53.Place, E49.Time_Appellation and E57.Material. A succeeding, second phase of the pipeline, which is described as the CRM-EH Relations Extraction phase, is aimed at identifying "rich phrases" that can be modelled as CRM-EH event and properties entities. Such Event and property entities connect in a meaningful pair, CRM entities that previously are identified by the NER phase.

The decision on the scope of events and property entities that should be targeted by the Relation Extraction phase of the OPTIMA pipeline has been informed by available use case scenarios [Appendix D1] and project discussions with archaeology experts. Since the outcome of the semantic annotation effort contributes to the STAR project, the selected CRM-EH entities and properties should support the project aims for semantic interoperability and cross searching with the excavation datasets participating in the STAR architecture. In addition, the selected events and properties should be based on the type of

CRM entities delivered by the NER phase and should connect such entities in the CRM-EH model by binary relations.

Detection of "rich phrases" also serves the CRM-EH specialisation effort of the pipeline with regards to CRM entities previously extracted by the NER phase. The employment of syntactic and semantic evidence in the form of hand-crafted JAPE grammars targeted at phrase level, can reveal sufficient contextual evidence regarding event and property argument-entities leading to their CRM-EH specialisation. For example, the entity-argument "pottery" previously annotated as E19.Physical_Object, when matched in the phrase "ditch contain pottery" can qualify as the CRM-EH entity EHE0009.Context_*Find*. Similarly the entity-argument "ditch" previously annotated as E53.Place can qualify as the CRM-EH entity EHE0007.Context, while the whole phrase itself can be modelled as *EHE1004.Context_Find_Deposition_Event*. The CRM-EH Relation Extraction pipeline is targeted at identifying the following three event and one property entities

- *EHE1001.ContextEvent*: The event entity associates an *E53.Place* with an *E49.Time_Appellation*, specialising such entities as *EHE007.Context* and *EHE0026.ContextEventTimeSpanAppellation*, respectively. It is defined as "*context formation event in which events took place and build a stratigraphic understanding of the site*". For grey literature the event is adopted to model phrases which associate a place with a time appellation in an archaeological setting, for example "pit dates to Prehistoric period", "Roman ditch" etc. The event is a specialisation of the CRM *E63. Beginning_of_Existence* event, which is defined as "*event that brings into existence any E77.Persistent_Item*", where *E77.Persistent_Item* is the super class of the *E19.Physical_Object*. The event may be used for temporal reasoning about things (intellectual products, physical items, groups of people, living beings) beginning to exist, such as the construction of a building or the birth of a person.

- *EHE1002.ContextFindProductionEvent*: The event entity associates an *E19.Physical_Object* with an *E49.Time_Appellation*, specialising the entities as *EHE009.ContextFind* and *EHE0039ContextFindProductionEventTimeAppellation*, respectively. It is defined as "*an event that resulted in the production of artefacts which later came to be deposited and eventually excavated*". For grey literature the event is adopted to model phrases which associate a physical object with a time appellation in an archaeological setting, for example "pottery dates to Prehistoric period", "Roman coin" etc. The event is a specialisation of the *E12.Production_Event,* which is defined as "*activity that is designed to, and succeeds in, creating one or more new items*". The event is used to describe the production of "new" items as well as the modification of existing items to "new" uses, as for example the modification of a potsherd to a voting token.
- *EHE1004.ContextFindDepositionEvent*: The event associates an *E19.Physical_Object* with an E53.Place, specialising the entities as *EHE009.ContextFind* and *EHE0007.Context*, respectively. It is defined as "*an event for explaining how a finds object came to end up in a context, enabling archaeologists to record information and interpretations about the deposition of the object*". For grey literature the event is adapted to model phrases which associate a physical object with a place in an archaeological setting, for example "ditch contains pottery". The event is a specialisation of the *E9. Move* event, which is defined as "*changes of the physical location of the instances of E19 Physical Object*". The event is used to document movement and relocation of objects such as the "*relocation of the London bridge from the UK to the USA* " or the movement of an exhibition to a new place .
- *P45.consists_of*: This is the only property that is targeted by the pipeline. It associates an *E19.Physical_Object* with an *E57.Material* and is used to associate an *EHE009.ContextFind* with *EHE030.ContextFindMaterial*. It is used to define "t*he instances of E57.Materials of which an instance of E18 Physical Thing is composed*", where *E18 Physical Thing* is the super class of *E19 Physical Object*. For grey literature the property is adopted to model phrases which associate a physical object with a material in an archaeological setting, for example "iron arrowheads", "sherds of pottery" etc.

6.3 Tracking Ontology Events via Corpus Analysis

A corpus analysis, "bottom-up" study was conducted aimed at revealing linguistic evidence to be utilised by the Relation Extraction (RE) pipeline. The corpus analysis task was based on the experience gained from a previous, corpus analysis, which was employed by the Negation Detection task (section 5.8.2). In total, 2460 documents participated in the analysis originating from the OASIS corpus.

Corpus analysis is a well-known technique adopted in adaptive NLP approaches for enabling the training of machine learning algorithms (Jurafsky and Martin 2000). It has been used to support a range of NLP problems including part-of-speech tagging, prepositional phrase attachment disambiguation, and syntactic parsing (Brill 1995). However, the current study is influenced by the work of George Kingsley Zipf and his study on corpus analysis for detecting the frequency rate of individual words.

6.3.1 Zipf's Law

Zipf's law (1935) states that, given a natural language corpus of sufficient volume, the frequency of any single word is inversely proportional to the word's rank associated with the frequencies. Thus, for a specific corpus the most frequent word is twice as frequent as the second in rank word and three times more frequent than the third in rank and so forth. The law is reflected by a probabilistic distribution (Zipfian), which is governed by a power-law behaviour expressed as a decay exponential function.

Zipf's experiment investigated the frequency with which words occur by examining three corpora from three different languages, Chinese, Latin and English. The Chinese corpus originated from colloquial samples of Chinese (Peiping dialect) totalling 13248 words in length and representing the occurrences of 3332 unique words, the Latin from four Playtime plays (Aulularia, Mostellaria, Pseudolus and Trinummus) totalling 33094 words in length and representing the occurrences of 8437 unique words and the English from a sample of American newspapers totalling 43989 words in length and representing the occurrences of 6002 unique words. The investigation focused on the occurrences of individual words not lexical units, where for example "child*"* and "children*"* are counted as two different words, although they encompass a single lexical *unit*.

Zipf observed that in all three corpora a few words occur with a very high frequency, while many words occur less frequently, describing a "*strikingly evident phenomenon that, as the number of occurrences increases, the number of different words possessing that*

number of occurrences decreases" (Zipf 1935). Even more significant was the orderliness with which numbers increased and decreased. To objectify the orderliness of the distribution, Zipf projected frequencies and occurrences of those words that occurred from 1 to 45 times on double logarithmic graph-paper (Figure 6.1). The line drawn approximately through the centre of the points is represented the by the formula $ab^2=k$ where *a represents the number of words of a given occurrence and b the number of occurrences.*

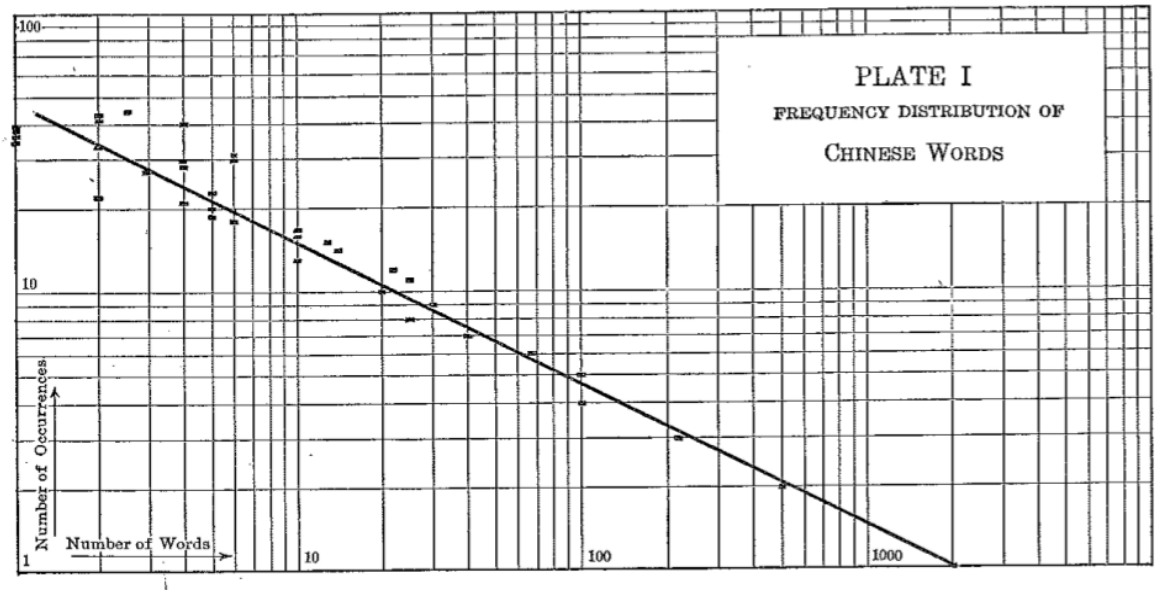


Figure 6.1: Frequency Distribution of Chinese Words, double algorithmic projection of occurrences (graph taken from Zipf 1935)

The Zipf law is equivalent to the power-law distribution, as is the Pareto distribution which differs from the Zipfian distribution only by the way it plots the cumulative distribution on the x,y axis (Newman 2006). The Pareto principle, which is also known as the 80%-20% rule, states that for many phenomena 80% of the effects are induced by 20% of the causes. Similarly to the Zipf example, where a small number of words are responsible for producing the largest amount of occurrences, the same observation can be made with respect to other phenomena. Examples include the population of cities where 20% of larger cities counts for 80% of a country's population and the distribution of wealth tends to follow the same principle, which was actually the observation made by Pareto regarding land ownership in Italy in the early 20th century. Other phenomena, which can be characterised by the 80-20 rule, include intensity of solar flares, diameter of moon craters, citation of scientific papers and web visits (Li 2002; Newman 2005).

6.3.2 Corpus Analysis Pipeline

The first stage of the corpus analysis task developed an IE pipeline (Figure 6.2) which extracted a vast amount of spans (146,008 spans), which were analysed further by a subsequent intellectual stage as discussed in section 6.3.3. The IE pipeline aimed at extracting spans in which pairs of CRM entities coexisted. The pairs of entities corresponded to the binary relations of CRM-EH events and properties (section 6.2.2.2).

Figure 6.2: Corpus analysis pipeline, JAPE grammars are shown in grey boxes, white boxes are used for GATE modules

The corpus analysis pipeline employed eight rules, two versions of the same rule for each pair, covering both directions of the relationship, i.e. Arg1 – Arg2 and Arg2 – Arg1. The four spans which were targeted by the pipeline were given project specific annotation labels. The labels scheme was adopted to ease processing and documentation of the extracted spans during the intellectual analysis stage. The span types which were targeted by the corpus analysis pipeline are the following.

PlaceObject: The spans that belong to this category include an *E19.Physical_Object* entity and an *E53.Place* entity. The extracted spans support the development of JAPE grammars targeted at extracting phrases of context find deposition events. Two versions of the same rule were used to extract the spans which were both combined in a single JAPE rule file running in appelt mode. The first JAPE grammar translates as {E19}{E53}, and the second as {E53}{E19}[8].

ObjectTime: The spans that belong to this category include an *E19.Physical_Object* entity and an *E49.Time_Appellation* entity. The extracted spans support the development of JAPE grammars targeted at extracting phrases of context find production events. Two versions of the same rule were used to extract the spans which were both combined in a single JAPE rule file running in appelt mode. The first pattern translates as {E19}{E49}, and the second pattern as {E49}{E19}.

8 As already discussed in section 3.3.1.2 (Summary section extraction), by introducing to a JAPE rule only the necessary annotation types, we control which types are transparent and processable by the rule. Thus, we manage to annotate large chunks of text using simple rules that bypass annotation types which are found within the text chunk but are not visible to the rule

PlaceTime: The spans that belong to this category include an *E53.Place* entity and an *E49.Time_Appellation* entity. The extracted spans support the development of JAPE grammars targeted at extracting phrases of context events. Two versions of the same rule were used to extract the spans which were both combined in a single JAPE rule file running in appelt mode. The first pattern translates as {E53}{E49}, and the second pattern as {E49}{E53}.

MaterialObject: The spans that belong to this category include an *E19.Physical_Object* entity and an *E57.Material* entity. The extracted spans support the development of JAPE grammars targeted at extracting phrases of find consists of material property. Two versions of the same rule were used to extract the spans which were both combined in a single JAPE rule file running in appelt mode. The first pattern translates as `{E19}{E57}`, and the second pattern as `{E57}{E19}`.

The last stage of the corpus analysis pipeline invoked the flexible exporter which was configured to export the resulted span annotations in XML output . The exporter was also configured to export *Token* annotations which are generated by the pre-processing phase of the pipeline. Token annotations are included by the XML output because they contained part of speech information that is used to reveal the linguistic patterns of the extracted spans.

6.3.2.1 Extracting Frequent Verbs

The corpus analysis proceeded to a subsequent stage where the extracted spans were processed further to reveal the most frequent verbs of each span type. The process followed the same technique as the extraction of verb phrases conducted during the phase of Negation Detection. In detail, the exported XML files were transformed via XSLT templates to text files containing the text of all spans. Four files resulted from the transformation, each file containing a span type of the four different types. The files were then processed further by a simple information extraction pipeline containing the following modules: a) Tokeniser b) Part of Speech Tagger c) Verb Phraser, and d) Flexible Exporter. The pipeline identified the verb phrases of each span while the Flexible Exporter exported the resulted verb phrase annotations in XML output. The exported XML files were transformed further via XSLT producing four CVS files, which were used to import the verbs phrases into a spreadsheet.

The process of transformation resulted in the construction of four lists each one containing the verb phrases of a different span type. In detail, the list of the *PlaceObject* span contained in total 19117 verb phrases having 1066 unique phrases, the *ObjectTime* list

contained 4622 verb phrases having 442 unique phrases and the *PlaceTime* list contained 13181 verb phrases having 1065 unique phrases. The unique phrases counts were calculated using the COUNTIF spreadsheet function for the range of available verb phrases. The resulting occurrences for each unique verb assembled a list similar to the Zipfian distribution, where the number of occurrences declines following a nearly exponential behaviour (Table 6.1).

PlaceObject		**ObjectTime**		**PlaceTime**	
Verb Phrase	***No.***	***Verb Phrase***	***No.***	***Verb Phrase***	***No.***
was	1815	is	524	is	1446
is	1537	are	356	was	1360
watching	1522	were	348	contained	865
contained	889	was	316	dating	736
were	793	dating	305	are	697
are	698	including	270	were	670
containing	698	finds	263	to be	541
were recovered	651	watching	257	lies	446
ditching	633	to be	168	containing	424
finds	553	worked	159	produced	373
looking	456	cut	142	were recovered	353
showing	360	recovered	140	cut	345
including	337	containing	137	associated	342
produced	335	were recovered	120	was recovered	331
cut	323	contained	117	including	316
was recovered	322	dated	104	dated	308
facing	321	associated	99	finds	257
has	312	struck	98	remains	249
associated	289	suggests	97	suggests	228
recovered	282	made	93	revealed	205

Table 6.1: The 20 most frequent verb phrases and their occurrences for the 3 different span types. Projecting the occurrences of verbs in (x,y) axis, the resulting graphs follow a Pareto distribution; similar graphs are produced by all four different types of spans. The graph below (Figure 6.3) shows the distribution of verb occurrences for the PlaceObject span type.

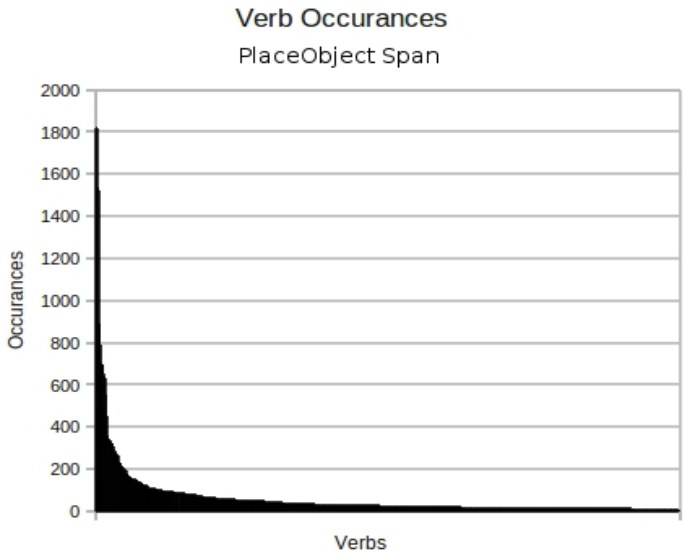


Figure 6.3: Verb occurrences distribution of the span type PlaceObject

A particular version of the system, which was soon abandoned, used the lists of frequent verbs and defined JAPE rules for extracting CRM-EH events. The system assumed that use of frequent verbs, in combination with token offsets, would be sufficient for extracting phrases which contained the desired CRM-EH specialisation. The rules were based on patterns which extracted phrases that commenced with a CRM entity, followed by a number of tokens (zero to five), followed by a frequent verb, followed by a number of tokens, followed by another CRM entity. For example the pattern `{E19}({Token}[0,5]){E12Verb}({Token}[0,5]){E49}` was targeted at extracting phrases containing an *EHE1002.ContextFindProductionEvent*.

The system managed to deliver phrases which contained the targeted CRM-EH event but also delivered a considerable amount of false positive phrasal annotations. Many of the frequent verbs appear in more than one list, with the verb to "be" (in its various forms) the most frequent. Due to this extensive overlap between lists, it is not practicable simply to define a specialised vocabulary of verbs for each type of span.

The above approach was abandoned in favour of part of the speech patterns approach discussed below in section 6.3.2.2. However, a single specialised vocabulary of verbs [Appendix A6] was defined, which was utilised by JAPE grammars for identifying context find deposition events. The particular list was constructed by examining the resulted verbs lists and empirically selecting verbs which could support the definition of deposition events. The grammars of the *Move Event*, *Production Event* and *consists of* material do not make use of specialised verb vocabulary, since as discussed below, the identified patterns were capable of supporting the definition of grammars that did not require employment of

a specialised verb vocabulary.

6.3.2.2 Exposing the Part of Speech Patterns

As discussed in section 6.3.2, the Flexible Exporter of the Corpus Analysis pipeline was configured to export *Token* annotations. The *Tokens* were parameterised by a Part-of-speech (POS) module to carry tags which corresponded to the grammatical class of words, i.e. noun (NN), adjective (JJ), adverb (RB), determiner (DT) etc. The part of speech tags followed the Hepple Tagger notation (Hepple 2000). The inclusion of *Tokens* in the XML output enabled a grammatical view of the identified spans, which supported a further analysis of the emerging patterns. Such patterns were analysed and abstracted into JAPE grammars, which were targeted at identifying textual spans corresponding to CRM-EH events.

The exposure of patterns was conducted via a PHP- DOM XML transformation of the XML files. A simple bespoke programme was written which recorded each identified span into a CSV file containing the span string, its POS pattern and the number of tokens involved. For example the span "deposits dated from the 11th century" has the POS pattern "NNS VBN IN DT JJ NN" and it contains 6 Tokens figure 6.4.

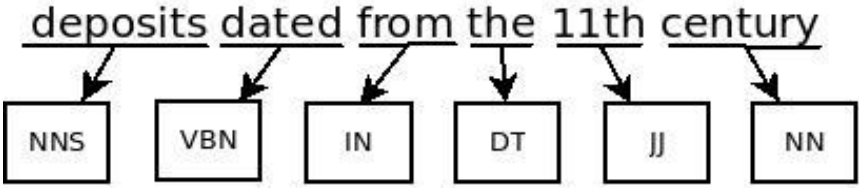


Figure 6.4: The emerging pattern for a particular span of type

Four individual CSV files were generated, each one containing spans of a particular kind. The *PlaceObjec*t CSV file contained 45524 spans, the *ObjectTime* CSV file contained 15707 spans, the *PlaceTime* CSV file contained 40895 spans and the *MaterialObject* CSV file contained 43882 spans (overall 146,008 spans). The resulting files were imported into spreadsheets which were processed further by the intellectual analysis phase.

6.3.3 Intellectual Analysis of the Extracted Spans

The intellectual analysis phase aimed to process the statistics of span patterns, in order to reveal commonly occurring pattern behaviour which could be abstracted as JAPE grammars. The phase processed the four individual spreadsheets and generated statistics, with regards to the pattern occurrences. The results of the occurrences were projected in graphs, which depicted the distribution of patterns with regards to the frequency of span length. The graphs revealed the trend of occurrences and assisted in the further analysis of the data. The data, as discussed below, revealed a Zipf like distribution of the occurrences, which formed the basis for determining the patterns selected and abstracted as JAPE grammars.

6.3.3.1 Analysis of the "PlaceObject" Event Spans

The first stage in the analysis of spans revealed occurrences of different span length, in terms of tokens. The spans were grouped according to their size (2 tokens span, 3 tokens, 4 tokens etc.) and the spreadsheet function COUNTIF used for calculating the number of occupancies within each group. In addition, the logarithmic values of token size and occurrences were calculated, which were projected in a double logarithmic chart (logarithmic scale) following Zipf's example. Table 6.2 below presents the first 30 spans, their occurrences and their logarithmic values. Figure 6.5 depicts the distribution of token occurrences for the first 49 pairs, i.e. span size from 2 to 50 while figure 6.6 shows the same distribution on a logarithmic scale.

Number of Tokens	Occurrences	Log Num. of Tokens	Log Occurrences
2	1097	0.30	3.04
3	3302	0.48	3.52
4	3922	0.60	3.59
5	3823	0.70	3.58
6	3702	0.78	3.57
7	3471	0.85	3.54
8	3006	0.90	3.48
9	2843	0.95	3.45
10	2402	1.00	3.38
11	2021	1.04	3.31
12	1803	1.08	3.26
13	1656	1.11	3.22
14	1418	1.15	3.15
15	1109	1.18	3.04
16	1061	1.20	3.03
17	979	1.23	2.99
18	829	1.26	2.92
19	820	1.28	2.91
20	690	1.30	2.84
21	554	1.32	2.74
22	485	1.34	2.69
23	414	1.36	2.62
24	334	1.38	2.52
25	328	1.40	2.52
26	292	1.41	2.47
27	240	1.43	2.38
28	211	1.45	2.32
29	173	1.46	2.24
30	170	1.48	2.23

Table 6.2: Pairs of Span size in number of Tokens in actual and logarithmic

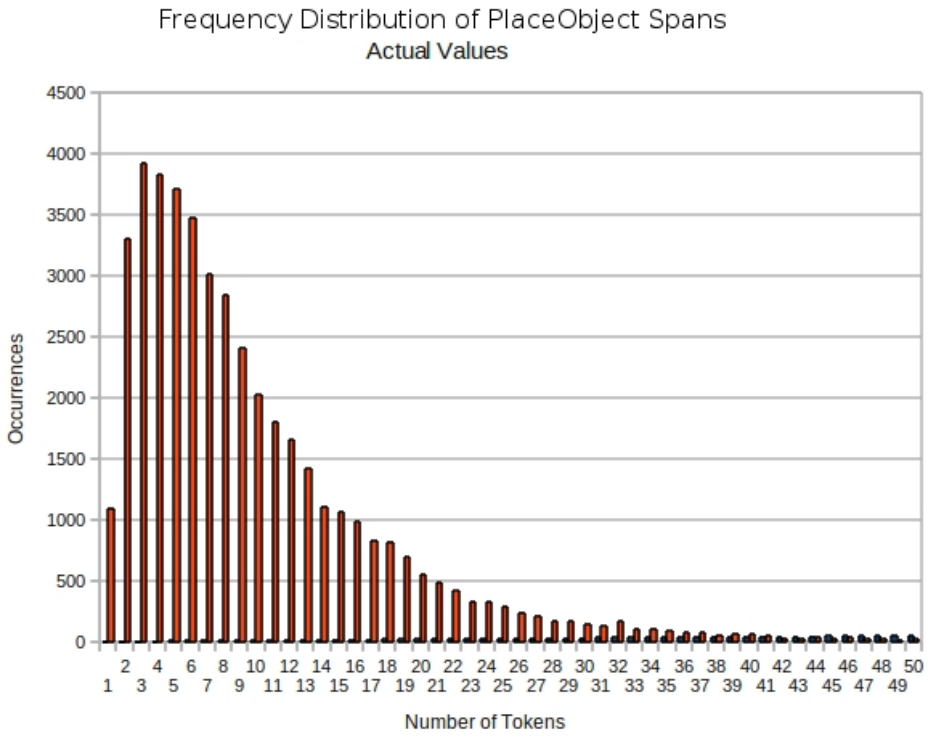


Figure 6.5: Span distribution actual values

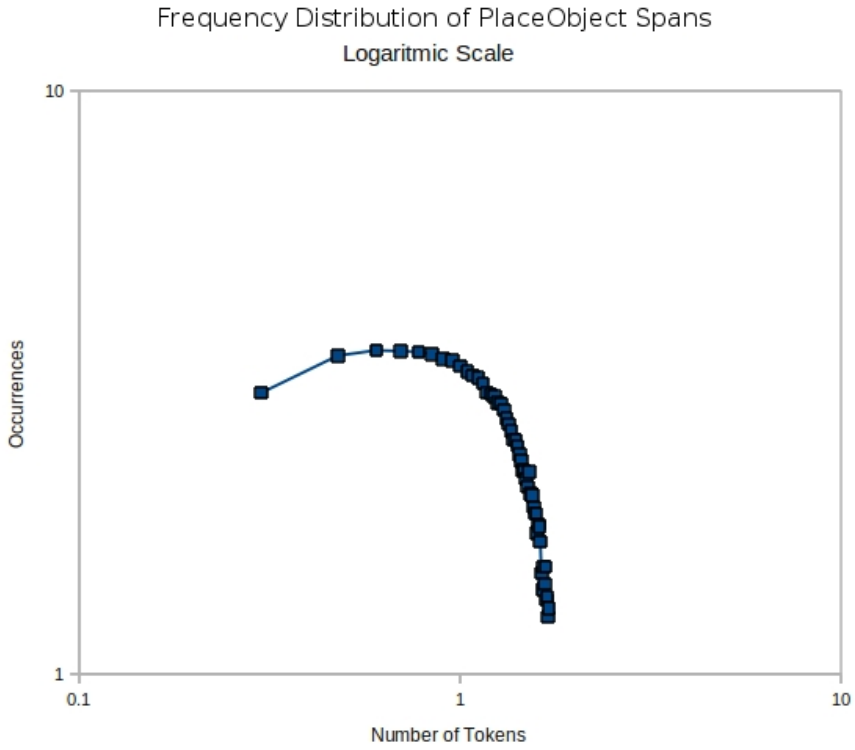


Figure 6.6: Span distribution on the logarithmic scale

The charts reveal a nearly exponential distribution of span occurrences where short spans containing few tokens are far more frequent than longer spans containing many tokens. This particular distribution is reflected by the left skewed graph of frequencies as shown in figure 6.6, where the vast majority of occurrences are observed between spans having 2 to 10 tokens, with the graph following the Pareto principle where 20% of spans are responsible for 80% of occurrences (36345 out of the overall 45524).

An alternative view of this particular behaviour is observed in the logarithmic scale distribution, where a steep decline of the number of occurrences is observed in spans that exceed 10 tokens, shown on the x-axis as log 1. Examining the graphs closer, it is also observed that the number of occurrences for some of the very small spans (2-3 tokens) is smaller than some larger spans (4-10 tokens). This is a particular behaviour of the *PlaceObject* span type and is not observed for the other span types that are discussed below. This behaviour can possibly be explained by the syntactical arrangement of Places and Physical Objects in phrases. Based on the above observations regarding the distribution of occurrences and following the Pareto principle, it was decided that the analysis of patterns would be focused on spans of maximum 10 tokens.

The next stage of the analysis produced lists which grouped the emerged part of speech patterns by span length. Nine lists were produced for the span lengths 2 to 10 tokens. The occurrences of unique patterns were generated by applying the spreadsheet function COUNTIF on the string value of each available pattern for the range of spans in the particular list. For example in the list of span length 2 tokens, the pattern "NN NN" (two nouns in a row) occurred 342 times, whereas the pattern "JJ NN" (an adjective followed by a noun) 81 times. In total, 37 unique patterns were found having a span length of 2 tokens. The technique of finding the occurrences of unique patterns was applied to all nine lists table 6.3. It was observed that the larger the span the bigger the number of unique patterns, since more tokens are involved in the span and so more lexical combinations can be made. Although for span length 8, 9, and 10 the number of unique patterns seems to slightly decrease, this is due to the decrease of the overall number of patterns in the list. For example, the list of span length 8 tokens totalling 3006 entries has 2365 unique patterns, while the list of span length 10 tokens totalling 2402 entries has 2060 unique patterns.

Span Length	2	3	4	5	6	7	8	9	10
Unique Patterns	37	212	798	1568	2262	2462	2365	2265	2060

Table 6.3: The number of unique patterns for 9 different span lengths

The number of occurrences of each unique pattern approximates a Zipfian distribution similar to the trend of occurrences of span lengths as described before. Therefore, in the list of two tokens length, the most frequent pattern occurs 342 times, the second most frequent 220, the third 182, the fourth 81, the fifth 55, the tenth 8 so on so forth. 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 in the list of five tokens length, the most frequent pattern occurs 134 times, the second most frequent 74, the third 71, so on so forth, while in the list of ten tokens length, the most frequent pattern occurs 81 times, the second most frequent 46 and the third 19 times.

Based on the above distribution of frequencies, it is clear that the Pareto principle can be applied for selecting the 20% of the most frequent patterns of each span size up to ten tokens to inform the definition of JAPE rules. However, the frequency aspect alone could not guarantee that the patterns reflected occurrences of CRM-EH events. For example, in the list of span length 2 tokens, although it contains many spans (1097) and a range of unique patterns (37), none actually reflected a move event which could be understood as a deposition event even implicitly. For example "hearth stone" or "coffin nails" are cases where the two CRM entities are found next to each other but the phrases do not denote a deposition event, instead one entity acts as moderator to the other. Therefore, a qualitative, intellectual analysis of the patterns was also required, together with the quantitative analysis, in order to select the patterns implemented as JAPE grammars.

The intellectual analysis and selection of patterns stage examined the patterns as to their merit for supporting identification of CRM-EH events. Hence, only a small number of patterns were selected from the vast range available, since many of the examples were arbitrary spans containing the two CRM entities, which did not denote a CRM-EH event or even make any sense as phrases. Although, the intellectual analysis focused on the most frequent patterns (considering the Pareto principle), the selection was not based only the statistical input. Additionally, some patterns were examined beyond the limit of the 20% most frequent, while very frequent patterns were not included if they did not denote a CRM-EH event. Overall the intellectual analysis for the *PlaceObject* span, selected 91 patterns from the 9 lists which were then abstracted into JAPE grammars, as discussed in section 6.4.3 below. A sample of those 91 patterns, such as "ditch containing burnt flint" (pattern *NN VBG JJ NN)* and "artefacts retrieved from these deposits" (*pattern NNS VBD IN DT NNS*) can be found in [Appendix C4].

6.3.3.2 Analysis of the "ObjectTime" Event Spans

The analysis of the *ObjectTime* spans followed the same methodology as the analysis of *PlaceObject* Spans. The first stage calculated the occurrences of the different span sizes and the logarithmic values for span sizes and their occurrences. The resulting data were projected on a decimal and a logarithmic scale chart for depicting the distribution of span occurrences [Appendix C1]. Based on the charts projection, a judgement was made for the maximum length of spans included for further pattern analysis. It was decided that the maximum span size to be 10 tokens long. The patterns of span length 2-10 tokens were grouped under 9 different lists and the unique patterns of each list were calculated. The analysis concluded with the intellectual selection of 75 patterns, a sample of which can be found in [Appendix C5]. The selected patterns were then abstracted into JAPE grammars (section 6.4.2).

The number of unique patterns presents a very similar frequency distribution to the distribution of the *PlaceObject* spans. The only difference between the two distributions concerns the number of frequencies of the very small spans of 2 to 3 tokens. In the *ObjectTime* distribution, such span lengths have the highest number of occurrences of the entire range but in the *ObjectPlace* distribution the highest frequency is for span length of 4 tokens.

Another slight diversion from the Zipfian distribution is shown by the 10 token length spans which a have higher frequency than spans of length 5, 6, 7, 8 and 9 tokens. However, this behaviour is not repeated in the rest of the spans after the 10^{th} position and so it is considered as an irregularity which did not affect the threshold of selection. As observed in the charts, the frequencies beyond the span size of 10 tokens drop significantly and do not present any particular interest.

6.3.3.3 Analysis of the "PlaceTime" Event Spans

The analysis of the *PlaceTime* spans followed the same method as the other two types of spans as described above. The occurrences of the various span sizes, the frequency distribution in actual values and in logarithmic scale are shown in [Appendix C2] . The behaviour of the frequency distribution is very similar to the *ObjectTime* span type, where the frequency of spans is inversely proportional to their size, with the 2 Token long span being the most frequent. As with the previous span types, the logarithmic chart depicts a significant decline of frequencies beyond the length of 10 Tokens, which is the maximum size of spans that is considered by the intellectual analysis. The analysis concluded with

the intellectual selection of 105 patterns, a sample of which can be found in [Appendix C6]. The selected patterns were then abstracted into JAPE grammars (section 6.4.1).

6.3.3.4 Analysis of the "MaterialObject" Spans

The last type of span analysis concerned a CRM-EH property rather than a CRM-EH event. However, the same methodology was followed for analysing the behaviour of span frequency as the event type spans above. Appendix C3 presents the data of the analysis, as previously discussed for the other span types.

This particular distribution presented some unique characteristic with regards to span size of highest occurrences and threshold of span sizes considered by the intellectual analysis stage. The distribution presented a great difference between the number of occurrences between spans of two token length and spans of three token length. The distribution of the *PlaceObject* span type presented a similar frequency behaviour. However, in the case of *MaterialObject*, the difference between occurrences is sharper and the three token long spans are 5 times more frequent than the two tokens.

This behaviour is explained by a close examination of 3, 4 and 5 token long spans, while considering that there is an extensive term overlap between the terminological resource used to match physical objects and materials. Many of the patterns describe physical objects or materials which are conjunct by a comma or the word "and", such as the phrases "iron, lead", "brick and tile", "bone or artefact", etc. Such patterns do not describe a "consists of" material property and were rejected by the intellectual analysis.

The distribution figures also revealed that spans which are greater the 6 tokens long are less frequent and so the intellectual analysis did not consider spans beyond this length. The analysis concluded with the intellectual selection of 27 patterns, a sample of which can be found in [Appendix C7]. The selected patterns were abstracted into JAPE grammars as discussed below (section 6.4.4).

6.4 Rules for Extracting CRM-EH Events

The formulation of JAPE grammars is a task that requires expert skills for the definition of regular expression patterns. The analysis stage delivered an extensive range of part of speech patterns, as shown in [Appendix C], which were then translated into JAPE grammars. The translation stage from linguistic patterns to the JAPE grammars requires abstraction skills since it is tedious and also unnecessary to define as many grammars as the number of available linguistic patterns. In order to abstract the rules, the volume of the selected part-of-speech patterns was examined and divided into two large groups, one

group formulated for the patterns having size up to 5 tokens and another group for the patterns having length 6 to 10 token. Then the two groups were further divided into two sub-groups for those patterns containing a verb phrase and for those that did not contain a verb phrase. This technique allowed grouping of phrases that shared common pattern characteristics, which were then abstracted into JAPE grammars.

The use of JAPE operators enabled the formulation of complex expressions that matched a range of different linguist patterns. For example the phrases "deposit is prehistoric", "deposits were clearly modern" and "deposit was of post-medieval date" can all be matched by a single grammar that uses the logical OR and the ? operator (zero or 1 occurrences). Thus, the JAPE rules translates as: a Place which is followed by a verb phrase, which might be followed by zero-or-one tokens of the kind preposition or adverb, which is followed by a Time Appellation entity, i.e.

```
{E53}{VG}({Token.kind==IN}|{Token.kind==RB})?{E49}
```

JAPE syntax supports the following operators, which were extensively used during the formulation of the rules, logical OR (|), the recursive Kleene operator (*) for matching zero or more occurrences of the same expression. The Kleene operator has two additional versions the (+) for matching one or more and the (?) for matching zero or one occurrences of the same expression. Also JAPE allows the definition of dynamic spans, for example, match 2 to 5 [2,5] occurrences of the same expression. Note that operators are applicable to expressions. An expression can be as simple as matching a single token or more complex for matching a whole JAPE grammar phrase. Such phrases are wrapped in brackets and so the operator is applicable to a matching expression. For example the expression

```
{E19}({Token.string== ",”}{E19})*({Token.string== ",”}|
{Token.string== "and”}}){E19}
```

will match a long list of physical object entities separated by commas, such as "coin, nail, pottery, brooch, arrowhead and bowl". JAPE grammars do not need to state the logical AND operator which is implicit between expressions.

The JAPE grammars that resulted from the formulation stage, as described below, are incorporated into the CRM-EH Relation Extraction pipeline (Figure 6.7). The pipeline is divided into 5 distinct sets of rules.

The first set detects relations between CRM *E53.Place* and *E49.Time_Appellation* entities annotating textual phrases as *EHE1001.ContextEvent.*

The second set detects relations between CRM *E19.Physical_Object* and *E49.Time_Appellation* entities annotating textual phrases a *EHE1002.ContextFindProduction Event*.

The third set detects relations between CRM *E53.Place* and *E19.Physical_Object* entities annotating textual phrases as *EHE1004.ContextFindDepositionEvent*. The fourth set detects relations between CRM *E19.Physical_Object* and *E57.Material* entities annotating textual phrases as *P45.consists_of*

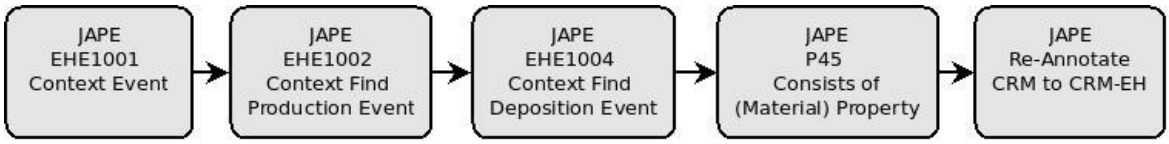


Figure 6.7: The CRM-EH event pipeline, grey boxes indicate set of rules

Using four distinct set of rules in a cascading order, where each set is targeted at identifying a particular event type and each set succeeds the other, enables annotation "overloading". This means that if a phrase or part of a phrase qualifies for more than one event/property annotation then all applicable annotations are generated. Consider the following phrase: "Ditch containing Roman pottery sherds". The phrase generates 3 different event/property annotations, an EHE1004 associating "Ditch" with "sherds", an EHE1002 associating "Roman" with "sherds" and a P45 associating "pottery" with "sherds". All three annotation types are generated due to the cascading order in which the different sets of rules are invoked.

The fifth transducer is a particular case which re-annotates all CRM entity annotations of CRM-EH event phrases to their CRM-EH specialisation. Therefore, if a *E53 Place* annotation is found within a CRM-EH *EHE1001.ContextEvent* or an *EHE1004.ContextFindDepositionEvent* then it is specialised as *EHE0007.Context*. Similarly, an *E19.Physical_Object* which is found within an *EHE1002.ContextFindProductionEvent* or an *EHE1004.ContextFindDepositionEvent* is specialised to an *EHE0009.ContextFind*.

An *E57.Material* and an *E19.Physical_Object* entity which are found within a *P45 consists of* phrase are also specialised as *EHE0030.ContextFindMaterial* and *EHE0009.ContextFind* respectively, since the CRM-EH model does not specialise the P45 property, which is used by the ontology to associate finds with their materials. Also time appellation entities, which are found within the event phrases EHE1001 and EHE1002, are specialised to the CRM-EH entities *EHE0026 Context Event Time Span Appellation* and *EHE0039 Context Find Time Span Appellation*, respectively.

6.4.1 EHE1001 Context Event Rules

This set consists of 16 JAPE grammars that detect and annotate phrases as *EHE1001.ContextEvent*. The rules use the internal to the pipeline annotation types *E53plus* and *E49plus*. The *CRMplus* annotation types join the CRM (E19, E49, E43 and E57) annotations with their conjunction annotations (E19Conjuction, E49Conjuction, E43Conjuction and E57Conjuction) under a single annotation type. Although the rules could have used, for example, the expression ({E53}|{E59Conjuction}) instead of {E53plus} the configuration of the rule to run on the *first* match mode would have caused the rule to exclude the conjunction annotation from the match since the mode favours of the shortest match. The *first* mode is used for restricting the dynamic matching of spans from delivering a match from a larger sentence clause when a shorter clause is present. The shortest clause match is assumed to be more precise than a larger clause and is prioritised. The same configuration applies to all CRM-EH Relation Extraction rules. A full list of the part of speech Hepple tagger categories that are used by the rules below can be found in Appendix F1.

Grammar I: A pair of a Time Appellation and a Place; for example "prehistoric pit".

```
{E49plus}{E53plus}
```

Grammar II: A phrase of three words where between a Time Appellation and a Place entity, there is a token which is not punctuation or verb; for example "medieval moated settlement", "Modern 74 Deposit".

```
{E49plus}
({Token.kind != punctuation, !Token.category == VB}|
{Token.kind != punctuation, !Token.category == VBG}|
{Token.kind != punctuation, !Token.category == VBP}|
{Token.kind != punctuation, !Token.category == VBD})
{E53plus}
```

Grammar III: A phrase of maximum four words where between a Place and a Time Appellation and entity, there is a verb, excluding "contain", which might be followed by a determiner, an adverb or a preposition; for example "Road is post-medieval", "deposits were clearly modern", "deposit was of postmedieval date".

```
{E53plus}
{VG, Token.root!="contain"}
({Token.category == DT}|{Token.category == IN}|
{Token.category == RB})?
{E49plus}
```

Grammar IV: A phrase of maximum five words where between a Time Appellation and a Place entity, there are up to 3 words, adjectives or nouns; for example "Roman ceramic land drain", "Saxon Sunken feature building".

```
{E49plus}
({Token.category == JJ}|
{Token.category == NN}|
{Token.category == NNS}|
{Token.category == NNP}|
{Token.category == NNPS}|
{Token.category == NP}|
{Token.category == NPS})[1,3]
{E53plus}
```

Grammar V: A phrase of maximum five words where between a Place and a Time Appellation entity, there might be a comma, which is followed by a preposition, which might be followed by a determiner or an adjective or an adverb; for example "deposits, of Roman Date", "pit of the Iron Age".

```
{E53plus}
({Token.string == ","})?
{Token.category == IN}
({Token.category == DT}|{Token.category == JJ}|
{Token.category == RB})?
{E49plus}
```

Grammar VI: A phrase of maximum seven words where between a Time Appellation and a Place entity, there are up to 3 tokens, which are not Physical Object entities or full-stops, followed by a preposition, which might be followed by a word; for example "early 2nd century AD system of ditched enclosures".

```
{E49plus}
({!E19, Token.string != "."})[0,3]
{Token.category == IN}
({Token.kind == word})?
{E53plus}
```

Grammar VII: A phrase of maximum ten words where between a Time Appellation and a Place entity, there are up to 3 words which are not Physical Object entities or full-stops, followed by a verb excluding "contain", which might be followed by an adjective, which is followed by a preposition, followed by a determiner, which might be followed by a word; for example "Saxon or medieval date are present on the site"

```
{E49plus}
({!E19, Token.string != "."})[0,3]
({VG, Token.root!="contain"})?
({Token.category == JJ})?
{Token.category == IN}
{Token.category == DT}
({Token.kind == word})?
{E53plus}
```

Grammar VIII: A phrase of maximum ten words where between a Time Appellation and a Place entity, there are up to 4 words which are not Physical Object entities or full-stops, followed by a noun, which is followed by "To", followed by a determiner, which might be followed by a word; for example "Medieval references to a ditch", "Medieval references to a town ditch".

```
{E49plus}
({!E19, Token.string != "."})[0,4]
({Token.category == NN}|{Token.category == NNS})
{Token.category == TO}
{Token.category == DT}
({Token.kind == word})?
{E53plus}
```

Grammar IX: A phrase of maximum ten words where between a Place entity and a Time Appellation, there are up to 4 words which are not Physical Object entities or full-stops, which might be followed by a verb or an adverb, which might be followed by another verb, or preposition or "To", which is followed by a preposition and a determiner for example; "ditch may well lie within the Saxon phase", "deposits dating from after the 18th century", "contexts dating to before the late 13th century".

```
{E53plus}
({!E19, Token.string != "."})[0,4]
({VG}|{Token.category == RB})?
({VG}|{Token.category == IN}|{Token.category == TO})?
{Token.category == IN}
{Token.category == DT}
{E49plus}
```

Grammar X: A phrase of maximum ten words where between a Place entity and a Time Appellation, there are up to 3 tokens, which are not Physical Object entities or full-stops, which might be followed by a verb or an adverb, which are followed by another verb, or adjective or adverb, which is followed by "To" and a determiner for example; "deposits relating to the medieval", " human burials dating back to the Bronze Age", "ditch probably dating to the Late Bronze Age".

{E53plus}

```
({!E19, Token.string != "."})[0,4]
({VG}|{Token.category == RB})?
({VG}|{Token.category == JJ}|{Token.category == RB})
{Token.category == TO}
{Token.category == DT}
{E49plus}
```

Grammar XI: A phrase of maximum ten words where between a Place entity and a Time Appellation, there are up to 6 words which are not Physical Object entities or full-stops, which are followed by "To" or "have", that is followed by a verb, for example; "pit was shown to be modern", "pits at Davyshiel are likely to be post-medieval", "ditch, which must have been Late Iron Age/Roman or later".

```
{E53plus}
({!E19, Token.string != "."})[0,6]
({Token.category == TO}|{Token.root == "have"})
({Token.category == VB}|{Token.category == VBN}|
{Token.category == VBP}|{Token.category == VBD}|
{Token.category == VBG})
{E49plus}
```

Grammar XII: A phrase of maximum ten words where between a Place entity and a Time Appellation, there are up to 5 tokens, which are not Physical Object entities or full-stops, which are followed by a verb excluding "contain", which is followed be an adverb or an adjective or a determiner which might be followed by an adverb for example; "layer suggests a 13th century date", "site is of relatively modern date", "ditch is probably Roman".

```
{E53plus}
({!E19, Token.string != "."})[0,5]
{VG, Token.root!="contain"}
({Token.category == RB}|{Token.category == IN}|
{Token.category == DT})
({Token.category == RB})?
{E49plus}
```

Grammar XIII: A phrase of maximum seven words where between a Place entity and a Time Appellation, there are up to 3 tokens, which are not Physical Object entities or full-stops, which are followed by an adverb which is followed by a preposition for example; "archaeological deposits, particularly of prehistoric date".

```
{E53plus}
({!E19, Token.string != "."})[0,3]
({Token.category == RB})
{Token.category == IN}
{E49plus}
```

Grammar XIV: A phrase of maximum ten words where between a Place entity and a Time Appellation, there are up to 4 tokens, which are not Physical Object entities or full-stops, followed by a verb excluding "contain", which is followed by "To", a determiner, an adjective and a "To" for example; "pit 37 belonging to an early to mid-13th century".

```
{E53plus}
({!E19, Token.string != "."})[0,4]
{VG, Token.root!="contain"}
{Token.category == TO}
{Token.category == DT}
{Token.category == JJ}
{Token.category == TO}
{E49plus}
```

Grammar XV: A phrase of maximum fifteen words (exception of a large span due to the use of the phrase "dating to the ...") where between a Place entity and a Time Appellation, there are up to 10 tokens, which are not Physical Object entities or full-stops, followed by a token of root "date", followed by "To" and determiner for example; "building of the late type, probably dating to the 18th century".

```
{E53plus}
({!E19, Token.string != "."})[0,10]
{Token.root=="date"}
{Token.category == TO}
{Token.category == DT}
{E49plus}
```

Grammar XVI: Phrase of a Time Appellation and a Place entity where a sub clause surrounded by commas exists between the two entities. The sub clause consists of adverb, verb of past tense/passive voice and a coordinating conjunction; for example "Roman, heavily silted and blocked, land drains".

```
{E49plus}
{Token.string == ","}
({Token.category == RB}|{Token.category == VBN}|
{Token.category == CC})+
{Token.string == ","}
{E53plus}
```

6.4.2 EHE1002 Context Find Production Event Rules

This set consists of 16 JAPE grammars that detect and annotate phrases as *EHE1002.ContextFindProductionEvent*. The rules use the internal to the pipeline annotation type *E19plus* and *E49plus* which joins the CRM and CRMConjuction annotations under a single annotation description.

Grammar I: A pair of a Time Appellation and a Physical Object entity; for example "medieval find".

```
{E49plus}{E19plus}
```

Grammar II: A phrase of three words where between a Time Appellation and a Physical Object, there is a token which is not punctuation or verb; for example "Prehistoric 406 Flint", "Neolithic worked flint".

```
{E49plus}
({Token.kind != punctuation, !Token.category == VB}|
{Token.kind != punctuation, !Token.category == VBG}|
{Token.kind != punctuation, !Token.category == VBP}|
{Token.kind != punctuation, !Token.category == VBD})
{E19plus}
```

Grammar III: A phrase of maximum 5 words where between a Physical Object and Time Appellation, there might be a comma which is followed by a verb or a preposition, which might be followed by a determiner, or an adverb, or an adjective or a preposition; for example "coins are Roman", "finds of Roman period", "coin of the Roman period", "coin is clearly modern", "coins, of Roman Date".

```
{E19plus}
({Token.string == ","})?
({VG}|{Token.category == IN})
({Token.category == DT}|{Token.category == JJ}|
{Token.category == RB}|{Token.category == IN})?
{E49plus}
```

Grammar IV: A phrase of maximum five words where between a Time Appellation and a Place entity, there are up to 3 words, adjectives or nouns; for example "Roman domestic pottery find".

```
{E49plus}
({Token.category == JJ}|
{Token.category == NN}|
{Token.category == NNS}|
{Token.category == NNP}|
{Token.category == NNPS}|
{Token.category == NP}|
{Token.category == NPS})[1,3]
{E19plus}
```

Grammar V: A phrase of maximum ten words where between a Time Appellation and a Physical Object entity, there are up to 3 tokens, which are not Place entities or full-stops, which might be followed by a verb, which might be followed by an adjective, which is followed by a preposition, followed by a determiner, which might be followed by a word; for example "post-medieval date were identified on the coin", "coins dating from the 18th century".

```
{E49plus}
({!E53, Token.string != "."})[0,3]
({VG})?
({Token.category == JJ})?
{Token.category == IN}
{Token.category == DT}
({Token.kind == word})?
{E19plus}
```

Grammar VI: A phrase of maximum ten words where between a Time Appellation and a Physical Object entity, there are up to 4 tokens, which are not Place entities or full-stops, that are followed by a noun, which is followed by "To", followed by a determiner, which might be followed by a word; for example "medieval references to a coin", "medieval references to a silver coin".

```
{E49plus}
({!E53, Token.string != "."})[0,4]
({Token.category == NN}|{Token.category == NNS})
{Token.category == TO}
{Token.category == DT}
({Token.kind == word})?
{E19plus}
```

Grammar VII: A phrase of maximum ten words where between a Physical Object and a Time Appellation entity, there are up to 3 tokens, which are not Place entities or full-stops, which might be followed by a verb or an adverb, which might be followed by a preposition, or "To" or a verb, which is followed by a preposition, followed by a determiner; for example "coins dating from after the 18th century".

```
{E19plus}
({!E53, Token.string != "."})[0,3]
({VG}|{Token.category == RB})?
({VG}|{Token.category == IN}|{Token.category == TO})?
{Token.category == IN}
{Token.category == DT}
{E49plus}
```

Grammar VIII: A phrase of maximum ten words where between a Physical Object and a Time Appellation entity, there are up to 3 tokens, which are not Place entities or full-stops, which might be followed by a verb or an adverb, which might be followed by an adjective, or an adverb or a verb, which is followed by "To" which might be followed by a

preposition, followed by a determiner; for example "coin probably relates to the 20th century", "animal remains dating back to the Bronze Age".

```
{E19plus}
({!E53, Token.string != "."})[0,3]
({VG}|{Token.category == RB})?
({VG}|{Token.category == JJ}|{Token.category == RB})
{Token.category == TO}
({Token.category == IN})?
{Token.category == DT}
{E49plus}
```

Grammar IX: A phrase of maximum ten words where between a Physical Object and a Time Appellation entity, there are up to 6 tokens, which are not Place entities or full-stops, which are followed by "To", a verb or an adverb, which might be followed by an adjective, or an adverb or a verb, which is followed by "To", followed by a verb; for example "coin was shown to be modern", "animal remains at Davyshiel are likely to be post-medieval".

```
{E19plus}
({!E53, Token.string != "."})[0,6]
{Token.category == TO}
({Token.category == VB}|{Token.category == VBN}|
{Token.category == VBP}|{Token.category == VBD}|
{Token.category == VBG})
{E49plus}
```

Grammar X: A phrase of maximum ten words where between a Physical Object and a Time Appellation entity, there are up to 5 tokens, which are not Place entities or full-stops, which are followed by a verb, which is followed by an adverb or a preposition or a determiner, which might be followed by an adverb; for example "artefact fragments associated with post-medieval date".

```
{E19plus}
({!E53, Token.string != "."})[0,5]
{VG}
({Token.category == RB}|{Token.category == IN}|
{Token.category == DT})
({Token.category == RB})?
{E49plus}
```

Grammar XI: A phrase of maximum ten words where between a Physical Object and a Time Appellation entity, there are up to 6 tokens, which are not Place entities or full-stops, which are followed by an adverb, which is followed by a determiner; for example "archaeological remains, particularly of prehistoric date".

```
{E19plus}
({!E53, Token.string != "."})[0,6]
({Token.category == RB})
{Token.category == IN}
({E49plus}
```

Grammar XII: A phrase of maximum ten words where between a Physical Object and a Time Appellation, entity there are up to 3 tokens, which are not Place entities or full-stops, which are followed by a verb, followed by "To", which is followed by a determiner, followed by an adjective, followed by "To"; for example "coin 37 belonging to an early to mid-13th century".

```
{E19plus}
({!E53, Token.string != "."})[0,3]
{VG}
{Token.category == TO}
{Token.category == DT}
{Token.category == JJ}
{Token.category == TO}
{E49plus}
```

Grammar XIII: A phrase of maximum seven words where between a Physical Object and a Time Appellation entity, there is a comma followed by a number, which might be followed by a noun, which might be followed by an adjective, followed by a comma; for example "Brick, 30mm thick, Roman".

```
{E19plus}
({Token.string == ","})
({Token.kind == number})
({Token.category == NN}|{Token.category == NNS})?
({Token.category == JJ})?
({Token.string == ","})
{E49plus}
```

Grammar XIV: A phrase of maximum ten words where between a Physical Object and a Time Appellation entity, there are up to 5 tokens which are not Place entities or full-stops, which are followed by a verb, which is followed by a noun phrase, which is followed by a preposition; for example "Find were fragments of late medieval/post-medieval".

```
{E19plus}
({!E53, Token.string != "."})[0,5]
{VG}
{NP}
{Token.category == IN}
{E49plus}
```

Grammar XV: A phrase of maximum fifteen words (exception of a large span due to the use of the phrase "dating to the ...") where between a Physical Object entity and a Time Appellation, there are up to 10 tokens, which are not Place entities or full-stops, followed by a token of root "date", followed by "To" and a determiner for example; "find of the late type, probably dating to the 18th century".

```
{E19plus}
({!E53, Token.string != "."})[0,10]
{Token.root=="date"}
{Token.category == TO}
{Token.category == DT}
{E49plus}
```

Grammar XVI: A phrase of a Time Appellation and a Physical Object entity, where a sub-clause surrounded by commas exists between the two entities. The sub-clause consists of adverb, verb of past tense/passive voice and a coordinating conjunction.

```
{E49plus}
{Token.string == ","}
({Token.category == RB}|{Token.category == VBN}|
{Token.category == CC})+
{Token.string == ","}
{E19plus}
```

6.4.3 EHE1004 Context Find Deposition Event Rules

This set consists of 7 JAPE grammars s that detect and annotate phrases as *EHE1004.ContextFindDepositionEvent*. The rules use the internal to the pipeline annotation type *E19plus* and *E53plus* which joins the CRM and CRMConjuction annotation under a single annotation description.

Grammar I: A phrase of maximum four words where between a Place and a Physical Object entity, there is a verb phrase, which might be followed by a determiner or a preposition or an adverb; for example "fills contained limestone", "cists contained a pot", "walls contained later brick".

```
{E53plus}
{VG}
({Token.category == DT}|
{Token.category == IN}|
{Token.category == RB})?
{E19plus}
```

Grammar II: A phrase of maximum five words where between a Place and a Physical Object entity, there is a preposition, which is followed by up to 2 tokens, which are not Place entities or full-stops; for example "findspots of prehistoric worked flint".

```
{E53plus}
{Token.category == IN}
({!E53, Token.string != "."})[0,2]
{E19plus}
```

Grammar III: A phrase of maximum five words where between a Physical Object and a Place entity, there is a preposition, which is followed by up to 2 tokens, which are not Place entities or full-stops; for example "pottery from Phase 1 contexts".

```
{E19plus}
{Token.category == IN}
({!E53, Token.string != "."})[0,2]
{E53plus}
```

Grammar IV: A phrase of four words where between a Physical Object and a Place entity, there is a verb phrase, which is followed by a preposition; for example "Amphora incorporated into context".

```
{E19plus}
{VG}
{Token.category == IN}
{E53plus}
```

Grammar V: A phrase of maximum four words where between a Physical Object and a Place entity, there is a preposition, which might be followed by a determiner; for example "slag in contexts", "finds from this layer".

```
{E19plus}
{Token.category == IN}
({Token.category == DT})?
{E53plus}
```

Grammar VI: A phrase of maximum eleven words where between a Physical Object and a Place entity, there are up to 4 tokens, which are not Physical Object entities or full-stops which are followed by a verb of the list E9_Verb [Appendix A6], which is followed by up to 4 tokens which are not Physical Object entities or full-stops; for example "The bottle were collected from levelling layer", "The animal bone was collected from two contexts", "The animal bone fragments were collected from seven contexts".

```
{E19plus}
({!E19, Token.string != "."})[0,4]
{Lookup.majorType == E9_Verb}
({!E19, Token.string != "."})[0,4]
{E53plus}
```

Grammar VII: A phrase of maximum eleven words where between a Place and a Physical Object entity, there are up to 4 tokens, which are not Place entities or full-stops, which are followed by a verb of the list E9_Verb [Appendix A6], which is followed by up to 4 tokens which are not Place entities or full-stops; for example "The bottle were collected from levelling layer", "The animal bone was collected from two contexts", "The animal bone fragments were collected from seven contexts".

```
{E53plus}
({!E53, Token.string != "."})[0,4]
{Lookup.majorType == E9_Verb}
({!E53, Token.string != "."})[0,4]
{E19plus}
```

6.4.4 P45 Consists of Property Rules

This set consists of 4 JAPE rules that detect and annotate phrases as *P45.consists_of* property. The rules use the internal to the pipeline annotation type *E19plus* and *E57plus* which joins the CRM and CRMConjuction annotation under a single annotation description.

Grammar I: A pair of a Material and a Physical Object entity; for example "ceramic artefacts".

```
{E57plus}{E19plus}
```

Grammar II: A phrase of three words where between a Physical Object and a Material entity, there is a preposition, or a verb phrase, or a Time Appellation entity; for example "artefacts of gold", "floor is stone", "artefacts comprising pottery", "bronze Roman vessels"

```
{E19plus}
({Token.category == IN}|{VG}|{E49plus})
{E57plus}
```

Grammar III: A phrase of maximum 5 words where between a Physical Object and a Material entity, there might be a verb phrase, followed by a preposition, which is might be followed by a Time Appellation entity; for example "pin was of iron", "wall is of stone", "sherd of postmedieval pottery"

```
{E19plus}
({VG})?
{Token.category == IN}
({E49plus})?
{E57plus}
```

Grammar IV: A phrase of maximum 5 words where between a Material and a Physical Object entity, there might be a verb phrase, followed by a preposition, which might be followed by a Time Appellation entity; for example "pottery from Anglo Saxon urn", "Anglo Saxon urn consists of pottery"

```
{E57plus}
({VG})?
{Token.category == IN}
({E49plus})?
{E19plus}
```

6.4.5 CRM-EH Entities Re-Annotation Rules

The re-annotation of CRM annotation to their CRM-EH specialisation is based on rules that use the *within* JAPE operator. The operator produces a match whenever a specified annotation is found within the boundaries of another annotation type. This particular matching operation is employed for finding the CRM annotations which are within the span of CRM-EH event or property annotation. The re-annotation stage uses 8 JAPE grammars in total.

Grammar I: A Place entity within a Context Event span is re-annotated as Context. This rule is presented here in full, including the Right Hand Side, which is omitted from the rules that follow.

```
Rule: EHE1001_ContextEvent_I
      Priority: 10
      (
      {E53 within EHE1001}
      )
      :Mention -->
      {
      try {
            gate.AnnotationSet annMention =
      ((gate.AnnotationSet)bindings.get("Mention"));
            gate.Annotation cleanKey =
      (gate.Annotation)annMention.iterator().next();
            outputAS.add(cleanKey.getStartNode(),
      cleanKey.getEndNode(),"EHE0007", cleanKey.getFeatures() );
            }
            catch (Exception ex) {
            Out.println("Exception in RHS rule
      EHE1001_ContextEvent_I
      ");
            ex.printStackTrace(Out.getPrintWriter());
            }
      }
```

Grammar II: A Place entity within a Context Find Deposition Event span is re-annotated as Context.

```
({E53 within EHE1004})
```

Grammar III: A Physical Object entity within a Context Find Production Event span is re-annotated as Context Find.

```
({E19 within EHE1002})
```

Grammar IV: A Physical Object entity within a Context Find Deposition Event span is re-annotated as Context Find.

```
({E19 within EHE1004})
```

Grammar V: A Physical Object entity within a Consists of Property span is re-annotated as Context Find.

```
({E19 within P45})
```

Grammar VI: A Material entity within a Consists of Property span is re-annotated as Context Find.

```
({E57 within P45})
```

Grammar VII: A Time Appellation entity within a Context Event span is re-annotated as Context Event Timespan Appellation.

```
({E49 within EHE1001})
```

Grammar VII: A Time Appellation entity within a Context Find Production Event span is re-annotated as Context Event Timespan Appellation.

```
({E49 within EHE1002})
```

6.5 Summary

The chapter revealed the rule-based method and techniques targeted at the task of relation extraction of CRM and CRM-EH entities. In particular, an innovative approach is revealed based on the application of the Zipfian distribution principles for the selection of candidate linguistic patterns that can be abstracted to hand-crafted JAPE grammars. Such linguistic patterns are the result of a shallow processing pipeline that uses part-of-speech tags, entities and domain dictionaries which is employed by a "bottom-up" corpus analysis study. Although, the discussed JAPE patterns are targeted at the recognition of specific CRM-EH event and property entities, the method itself based on a corpus analysis study and application of Zipfian distribution principles is not domain specific and possibly can be generalised. The chapter concludes a series of OPTIMA pipeline phases (pre-processing, NER, RE) executed within the GATE environment and targeted at detection and recognition of a range of entity types. The following chapter discusses the further manipulation of GATE semantic annotation into semantic indices and their usage in an information retrieval setting.

Chapter 7

Semantic Indices: Formats and Usage

7.1 Introduction

The chapter discusses the delivery and usage of semantic indices of grey literature. Being able to utilise and exploit the results of semantic annotation in a document retrieval and inspection environment is highly desirable. Thus, transformation of semantic annotations to interoperable abstractions (indices) enables use of semantic annotations by third party, non-GATE applications that support real-world information seeking activities.

The discussion reveals the transformation process of GATE annotations to interoperable outputs that are used by third party applications. In particular, the chapter discusses the transformation of semantic annotations to XML files, which couple semantic annotations with content and the transformation of semantic annotations to RDF triples that decouple annotation from content. Both XML and RDF outputs are utilised by web-based applications. The Andronikos web-portal utilises the XML output for presenting document abstractions of all semantic annotation phases (Pre-process, NER, RE). The STAR demonstrator exploits the RDF output of the CRM-EH RE phase to enable document retrieval and cross-searching between grey-literature documents and archaeological datasets.

The first part of the chapter discusses the transformation process of annotations to interoperable formats based on the arrangement of the CRM-EH ontology. The second part discusses uses of the XML output by the Andronikos web-portal and presents a range of examples that demonstrate use of semantic annotation in document inspection and fact finding. The third part discusses uses of the semantic annotation RDF triples by the STAR demonstrator in a document retrieval environment and presents a range of search scenarios that exploit the semantic properties of document indices. The chapter concludes with a brief discussion on the issue of *false positive* results and a summary of the chapter contributions.

7.2 Transforming Semantic Annotations to Indices

The transformation process of the semantic annotations to semantic indices delivered two separate outputs. The first output contained both annotation and context, coupled together in XML files, while the second contained the annotations decoupled from content, expressed as RDF graphs. The RDF (Resource Definition Framework) uses XML syntax for representing information in the form of *triples* (RDF graphs), which are statements that use a subject-predicate-object expression to describe resources. For example, expressing the notion that Cardiff is the capital city of Wales in RDF triple translates as; the *object* is "Wales", the *predicate* is "has capital" the *subject* is "Cardiff".

The GATE module, *Flexible Exporter,* was used for exporting the semantic annotations from the GATE environment as basic XML files. The module exports a list of user defined annotation types as XML tags. For example a textual instance "Roman", which is recognised as a CRM Time Appellation (E49) entity, is tagged as `<E49>Roman</E49>`. The Flexible Exporter does not export a well-formed XML output; it only provides the XML tags of annotations but it does not include any root tag information, nor does it write an XML header for declaring document type and encoding.

A bespoke PHP script was written for adding to the output of the Flexible exporter the XML header information, name-spaces and the root tag definition. The script adds the following lines at the beginning of each file exported from GATE by the Flexible Exporter.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<?xml-stylesheet type="text/css" href="single.css"?>
<doc xmlns:gate="http://www.gate.ac.uk"
xmlns:skos="http://www.w3.org/2004/02/skos/core#">
```

The script adds the header definition and document encoding, as well as a link to a CSS file which is utilised by the Andronikos web portal for displaying the coloured annotation spans. The root tag `<doc>` uses the name-spaces of GATE and SKOS. The PHP script also appends the closing tag (`</doc>`) at the end of each document.

7.2.1 XML Output

The Flexible Exporter of the OPTIMA pipeline is configured to deliver three distinct XML sets of semantic annotation output. The sets correspond to the Pre-process, NER (CRM) and RE (CRM-EH) phases of the pipeline (Figure 4.1), where each phase delivers a particular group of annotations. The first phase delivers the domain neutral document section annotations, the second phase delivers the four CRM oriented annotations (Place,

Physical Object, Material and Time Appellation), while the last phase delivers the CRM-EH events and CRM-EH specialisation of the annotations. All three output sets are processed by the bespoke script described above, which adds to the files root tag and XML header information.

In detail, the Pre-process configuration of the Flexible Exporter produces XML output including the following tags; *Heading, TOC* (Table Of Contents) and *Summary*. The tags mark textual instances of the aforementioned annotation types. For example, the textual instance "1. Introduction", recognised as a heading by the IE pipeline, is exported as an XML tag of the following format: `<heading>1. Introduction</heading>`. Similarly, *Summary* and *TOC* tags mark textual instances larger than heading phrases, normally containing multiple sentences.

The CRM configuration of the Flexible Exporter produces XML output including the tags *E19.Physical Objec*t, *E49.Time Appellation*, *E53.Place* and *E57.Material*. The tags of this particular configuration have attributes that hold additional information with regards to the textual elements. Four particular attributes are used in each tag type:

- The *gate:gateID* which is an auto-generated number produced by GATE during the IE process, used for the unique identification of each annotation
- The *skos:Concept* holding the unique terminological reference of each annotation
- The *note* attribute which is produced by JAPE rules during the IE process and holds a contextual snapshot of the phrase within which the annotation appears. The attribute stores an offset of 75 characters before and 75 characters after the textual instance of the annotation.
- The *rule* attribute which stores the name of the rule that produced the annotation. The attribute is used by the debugging process.

The CRM-EH configuration of the Flexible Exporter produces XML output of the following tags; *EHE0007.Context, EHE0009.ContextFind, EHE0026.ContextEventTimeSpanAppellation, EHE0030. ContextFindMaterial, EHE0039.ContextFindTimeSpanAppellation,* and the CRM-EH event tags *EHE1001.ContextEvent, EHE1002.ContextFindProdictionEvent, EHE1004.ContextFindDepositionEvent* and the CRM property tag *P45.consists_of*.
The tags mark textual instances of CRM-EH entities and Events. The CRM-EH event tags mark phrases that connect CRM-EH entities together. For example the phrase "Roman coin" denoting a production event, relates a context find with a time appellation. The above phrase is tagged as (omitting attributes and full tag description for simplicity):

```
<EHE1002>
      <EHE0039>Roman</EHE0039>
      <EHE0009>coin</EHE0009>
</EHE1002>
```

Moreover, the same phrase might be part of a larger phrase which denotes a deposition event, for example “hearth containing Roman coin”. In such a case, tags are nested under the deposition event tag as seen below:

```
<EHE1004>
      <EHE0007>hearth</EHE0007>contains
      <EHE1002>
            <EHE0039>Roman</EHE0039>
            <EHE0009>coin</EHE0009>
      </EHE1002>
<EHE1004>
```

The XML output of the above three configurations is utilised by the Andronikos web portal as discussed below (section 7.3). In addition, the same XML output is employed by the RDF transformation process (section7.2) for delivering semantic indices of RDF triples.

7.2.2 RDF Output

The transformation of the semantic annotations to RDF triples is based on further manipulation of the CRM-EH XML output. Although it is possible to deliver RDF triples from all three sets of XML output, the CRM-EH was selected as the most appropriate set for supporting the cross searching functionality of the STAR demonstrator.

The transformation technique employs bespoke PHP scripts that exploit the DOM XML object definition for accessing the contents of the semantic annotation XML files. PHP is used instead of XSLT due to its capabilities as a programming language to deal with string manipulation and array definition more efficiently and flexibly than the transformation templates of XSLT. The scripts define a range of templates which transform the XML structures to new, decoupled from content, RDF graphs of semantic annotation.

During transformation the PHP scripts utilise the file name of the processed documents to produce unique identifiers for each semantic annotation that is transformed to an RDF resource. The grey literature documents, which are processed by the OPTIMA pipeline, hold unique file names assigned by the OASIS system. In addition, each semantic annotation delivered by GATE is assigned an auto-generated ID number. The two (document name and auto-generated ID) are used as a compound key which uniquely identifies each semantic annotation turned into an RDF resource. In addition, the unique

document filename is also used as a resource by the *EHE0001.Project* entity, which is the CRM-EH entity that links all triples originating from the same document. Using the *EHE0001.Project* entity as a hub connecting all RDF triples of a document, allows the origin of document information to be made available to the user during retrieval. The above unique identification techniques are applied to the semantic annotation indices which are produced by the RDF graphs as discussed.

7.2.2.1 EHE0007.Context Graph

The figure below describes the triples of an example *EHE0007.Context* resource. The resource links to *EHE0001.Project* entity which represents the document from which it originates. The resource contains an *RDF value*, a *SKOS* terminological reference and a *note* of type *String*.

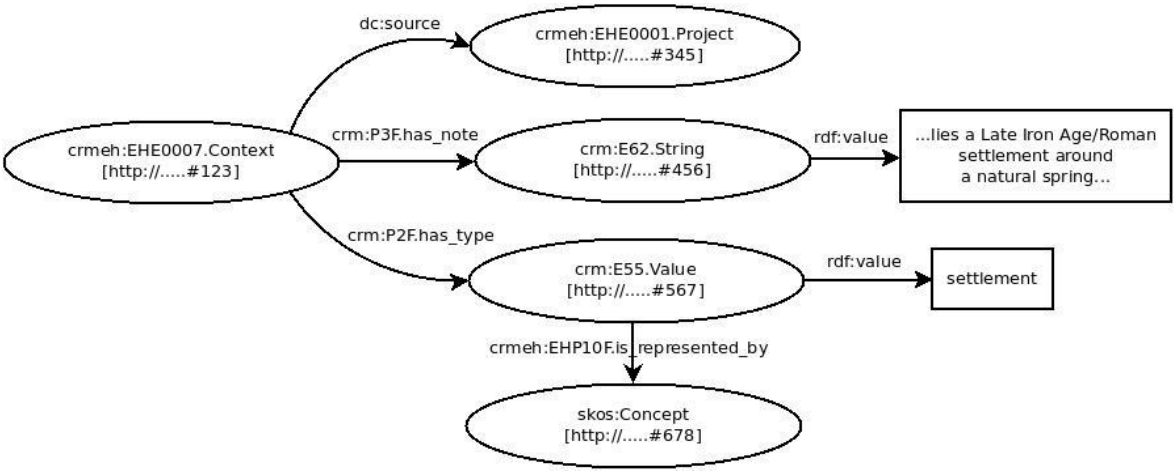


Figure 7.1: EHE0007.Context graph

The above example figure produces the following RDF code.

```
<crmeh:EHE0007.Context
     rdf:about="http://tempuri/star/base#suffolkc1-6115.286732">
     <dc:source rdf:resource="http://tempuri/star/base#suffolkc1-
     6115" />
     <crm:P2F.has_type>
          <crm:E55.Type>
          <rdf:value>settlement</rdf:value>
          <crmeh:EXP10F.is_represented_by
          rdf:resource="http://tempuri/star/concept#68977"/>
          </crm:E55.Type>
     </crm:P2F.has_type>
     <crm:P3F.has_note>
          <crm:E62.String>
          <rdf:value>...to the north-east lies a Late Iron Age /
          Roman settlement around a natural spring...</rdf:value>
          </crm:E62.String>
     </crm:P3F.has_note>
</crmeh:EHE0007.Context>
```

7.2.2.2 EHE0009.ContextFind Graph

The graph describes the triples of a *EHE0009.ContextFind* resource. The resource links to a *EHE0001.Project* entity which represents the document from which it originates. The resource contains an *RDF value*, a *SKOS* terminological reference and a *note* of type *String*. Note that in dotted lines is the triple which links EHE0009.Context with the *EHE0030.ContextFindMaterial*. Because not all *EHE0009.ContextFind* instances have a *P45F.consists_of* property the triple graph is shown in dotted lines.

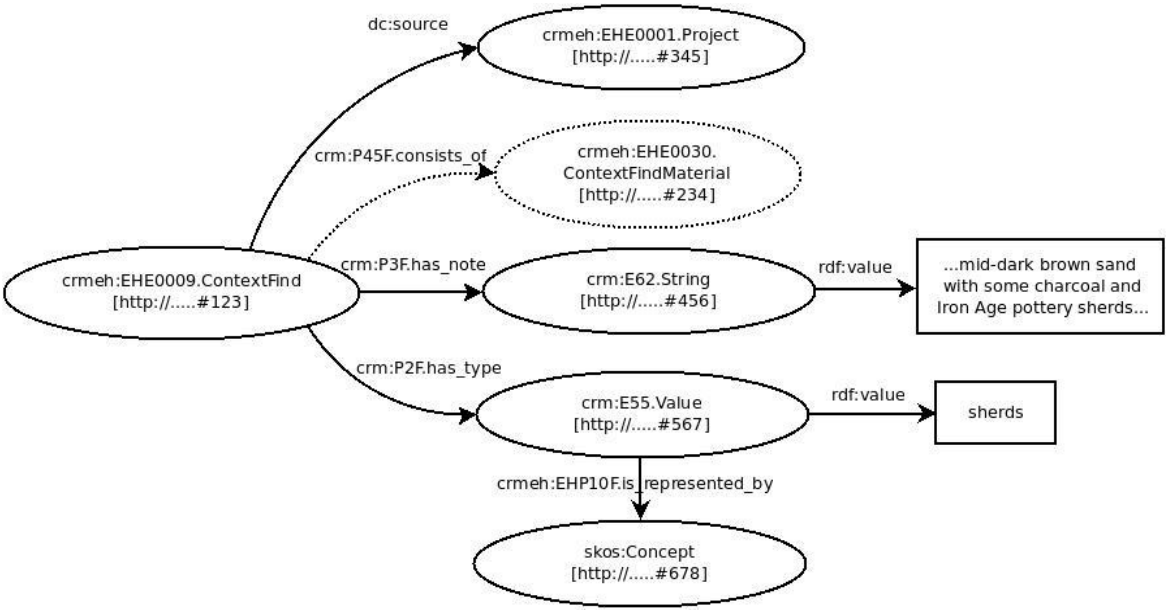


Figure 7.2: EHE0009.ContextFind graph

The above example figure produces the following RDF code.

```
<crmeh:EHE0009.ContextFind
      rdf:about="http://tempuri/star/base#suffolkc1-6115.286777">
      <dc:source rdf:resource="http://tempuri/star/base#suffolkc1-
      6115" />
      <crm:P2F.has_type>
            <crm:E55.Type>
            <rdf:value>sherds</rdf:value>
            <crmeh:EXP10F.is_represented_by
            rdf:resource="http://tempuri/star/concept#137051"/>
            </crm:E55.Type>
      </crm:P2F.has_type>
      <crm:P3F.has_note>
            <crm:E62.String>
            <rdf:value>...a mid-dark brown sand with some charcoal
            and Iron Age pottery sherds. 0008 was a circular
            pit...</rdf:value>
            </crm:E62.String>
      </crm:P3F.has_note>
      <crm:P45F.consists_of
      rdf:resource="http://tempuri/star/base#suffolkc1-
      6115.286776"/>
</crmeh:EHE0009.ContextFind>
```

7.2.2.3 EHE0030.ContextFindMaterial Graph

The graph describes the triples of a *EHE0030.ContextFindMaterial* resource. The resource links to a *EHE0001.Project* entity which represents the document from which it originates. The CRM entity *E57.Material* is a subclass of the *E55.Type* class, thus it has a value (without requiring any link to a *E55.Type* which has a value), which is represented by a SKOS terminological reference. In addition, the resource has a note of type *String*.

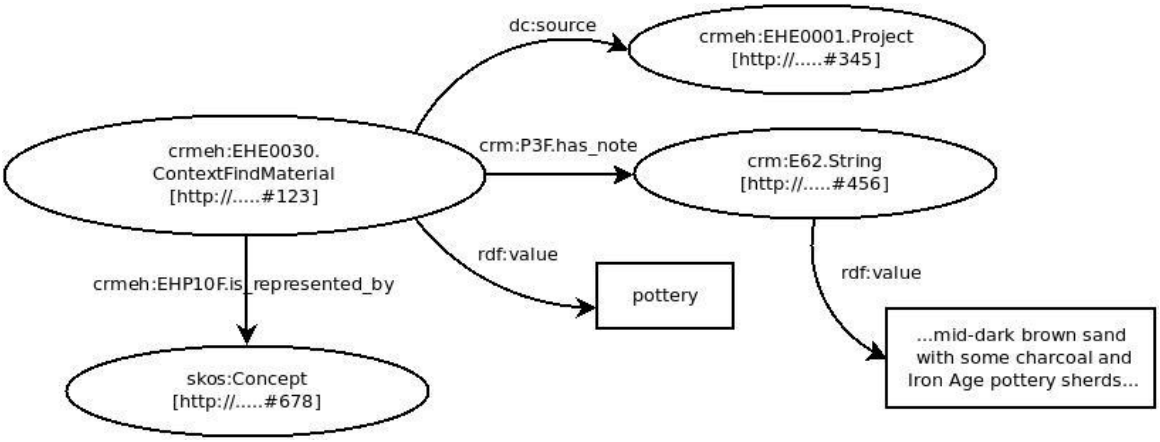


Figure 7.3: EHE0030.ContextFindMaterial graph

The above example figure produces the following RDF code. Note the unique identifiers assigned in each class. The unique identifier of this resource, "suffolkc1-6115.286776", is used above (in the RDF for ContextFind) to implement the link (consists_of Material) between EHE0009 and EHE0030.

```
<crmeh:EHE0030.ContextFindMaterial
      rdf:about="http://tempuri/star/base#suffolkc1-6115.286776">
      <dc:source rdf:resource="http://tempuri/star/base#suffolkc1-
      6115" />
      <rdf:value>pottery</rdf:value>
      <crmeh:EXP10F.is_represented_by
      rdf:resource="http://tempuri/star/concept#ehg027.2"/>
      <crm:P3F.has_note>
            <crm:E62.String>
            <rdf:value>...a mid-dark brown sand with some charcoal
            and Iron Age pottery sherds. 0008 was a circular
            pit...</rdf:value>
            </crm:E62.String>
     </crm:P3F.has_note>
</crmeh:EHE0030.ContextFindMaterial>
```

7.2.2.4 EHE0026.ContextEventTimeSpanAppellation

The graph describes the triples of a E*HE0026.ContextEventTimeSpanAppellation* resource. The resource links to a *EHE0001.Project* entity which represents the document from which it originates, also the resource has two values, which might be represented by up to two SKOS terminological references. The resource also has a note which is of type String.

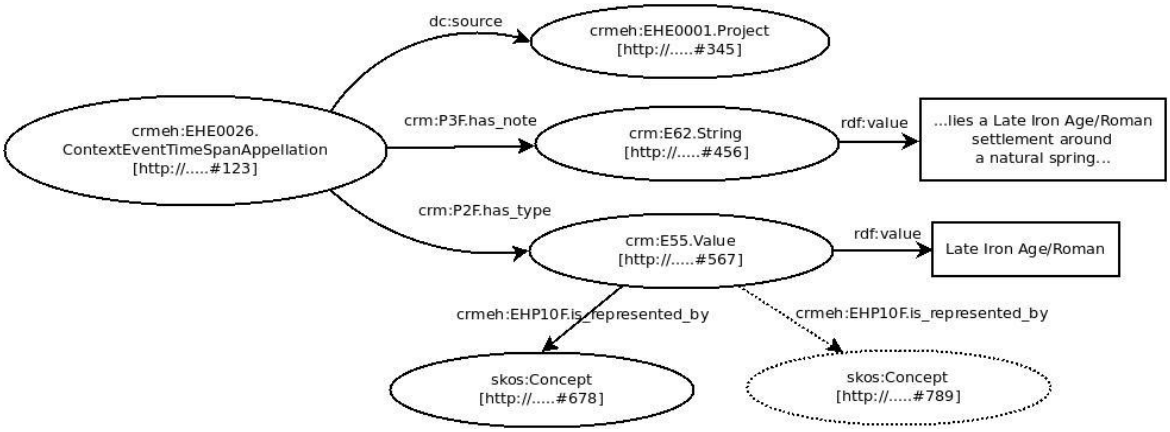


Figure 7.4: EHE0026.ContextEventTimeSpanAppellation graph

The above figure produces the following RDF code. Note that the resource is linked to (represented_by) two distinct terminological references, one linking to "Iron Age" and another to "Roman". This modelling choice is consistent with the use of dual reference for time appellation spans that are conjunct, as previously explained in section 3.2.4

```
<crmeh:EHE0026.ContextEventTimeSpanAppellation
      rdf:about="http://tempuri/star/base#suffolkc1-6115.286956">
      <dc:source rdf:resource=
      "http://tempuri/star/base#suffolkc1-6115" />
      <crm:P2F.has_type>
            <crm:E55.Type>
            <rdf:value>Late Iron Age/Roman</rdf:value>
            <crmeh:EXP10F.is_represented_by
            rdf:resource="http://tempuri/star/concept#134738"/>
            <crmeh:EXP10F.is_represented_by
            rdf:resource="http://tempuri/star/concept#134737"/>
            </crm:E55.Type>
      </crm:P2F.has_type>
       <crm:P3F.has_note>
            <crm:E62.String>
            <rdf:value>...throughout the vicinity of ERL 120.
            A kilometre to the north-east lays a Late Iron
            Age/Roman settlement around a natural spring at Caudle
            Head mere and   three ...</rdf:value>
           </crm:E62.String>
      </crm:P3F.has_note>
</crmeh:EHE0026.ContextEventTimeSpanAppellation>
```

7.2.2.5 EHE0039. ContextFindProductionEventTimeSpanAppellation

The graph describes the triples of a E*HE0039.ContextFindProductionEventTimeSpanAppellation* resource. The resource links to an *EHE0001.Project* entity which represents the document from which it originates. Also the resource has a value, which might be represented by up to two distinct SKOS terminological references. The resource also has a note of type *String*.

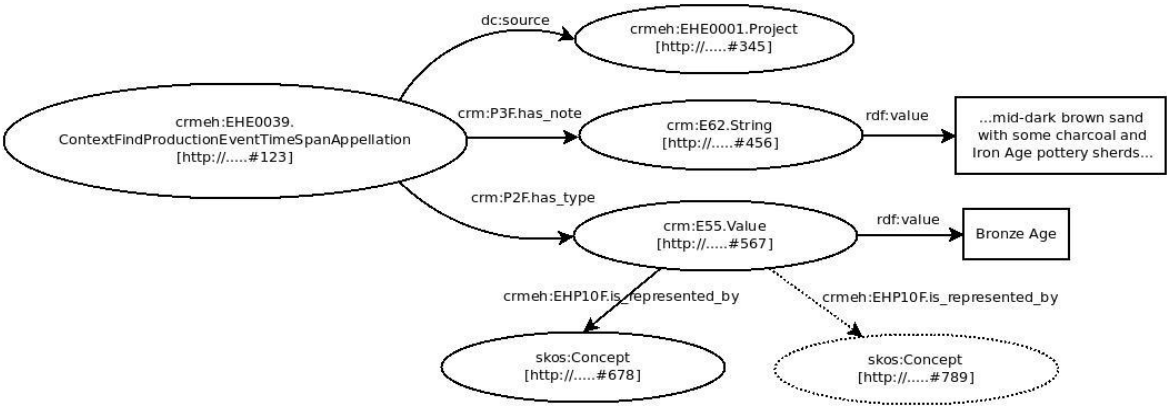


Figure 7.5: EHE0039. ContextFindProductionEventTimeSpanAppellation graph

The above figure produces the following RDF code.

```
<crmeh:EHE0039.ContextFindProductionEventTimeSpanAppellation
      rdf:about="http://tempuri/star/base#suffolkc1-6115.286969">
      <dc:source rdf:resource="http://tempuri/star/base#suffolkc1-
      6115" />
      <crm:P2F.has_type>
            <crm:E55.Type>
            <rdf:value>Iron Age</rdf:value>
            <crmeh:EXP10F.is_represented_by
            rdf:resource="http://tempuri/star/concept#134722"/>
            </crm:E55.Type>
      </crm:P2F.has_type>
      <crm:P3F.has_note>
            <crm:E62.String>
            <rdf:value>...base of the ditch and 0019, a mid-dark
            brown sand with some charcoal and Iron Age pottery
            sherds. 0008 was a circular pit, with fairly steep
            sides and...</rdf:value>
            </crm:E62.String>
      </crm:P3F.has_note>
</crmeh:EHE0039.ContextFindProductionEventTimeSpanAppellation>
```

7.2.2.6 EHE1001.ContextEvent

The graph describes the triples of *EHE1001.ContextEvent* resource. The resource connects an EHE0007 and an EHE0026 resource while it links to a *EHE0001.Project* entity. The resource has a note of type *String* holding the phrase instance in which the two constituent entities are found. The resource does not have a particular value other than the note of the phrase and thus it is not represented by a SKOS terminological reference.

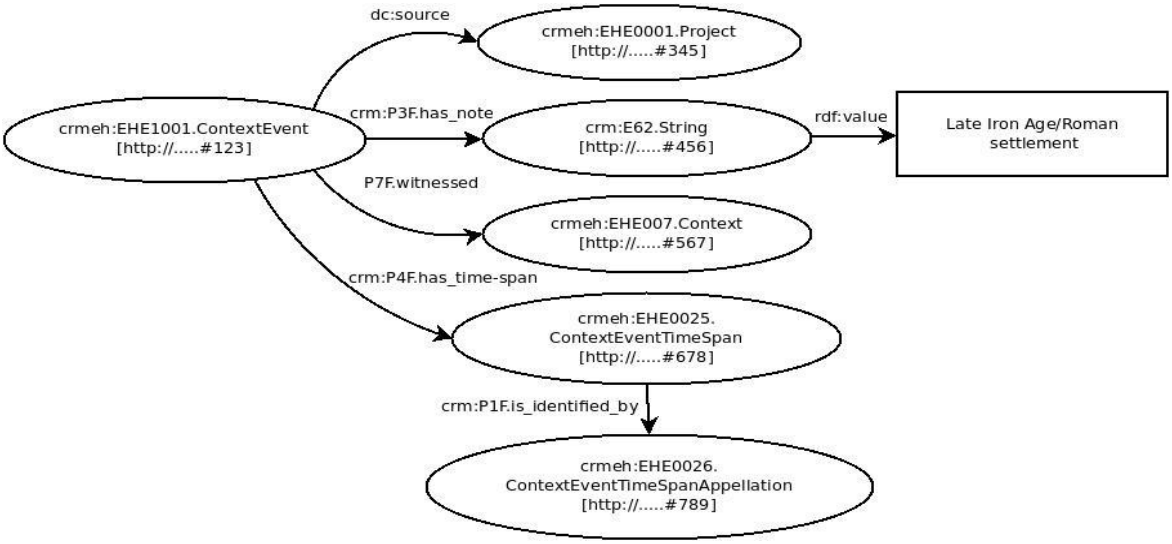


Figure 7.6: EHE1001.ContextEvent graph

The above figure produces the following RDF code. The example shows how linking is achieved between the resources EHE0007 (unique id: suffolkc1-6115.286732) and EHE0026 (unique id: suffolkc1-6115.286956)

```
<crmeh:EHE1001.ContextEvent
      rdf:about="http://tempuri/star/base#suffolkc1-6115.286597">
      <dc:source rdf:resource="http://tempuri/star/base#suffolkc1-
      6115" />
      <crm:P3F.has_note>
            <crm:E62.String>
            <rdf:value>Late Iron Age/Roman settlement</rdf:value>
            </crm:E62.String>
      </crm:P3F.has_note>
      <crm:P4F.has_time-span>
            <crmeh:EHE0025.ContextEventTimespan
            rdf:about="http://tempuri/star/base#suffolkc1-
            6115.286597.286956">
           <crm:P1F.is_identified_by
            rdf:resource="http://tempuri/star/base#suffolkc1-
            6115.286956" />
           </crmeh:EHE0025.ContextEventTimespan>
      </crm:P4F.has_time-span>
      <crm:P7F.witnessed rdf:resource=
      "http://tempuri/star/base#suffolkc1-6115.286732" />
</crmeh:EHE1001.ContextEvent>
```

7.2.2.7 EHE1002.ContextFindProductionEvent

The graph describes the various triples of a EHE1002.ContextFindProductionEvent resource. The resource connects an EHE0009 and an EHE0039 resource while it links to a *EHE0001.Project* entity. The resource has a note of type String holding the phrase instance in which the two constitution entities are found. The resource does not have a particular value other than the note of the phrase and thus it is not represented by a SKOS terminological reference. Note the figure below follows closely the CRM-EH ontological arrangements and includes the class EHE0038, which carries a fine ontological distinction from EHE0039 not applicable to information extraction. Thus, a mock resource is created that has a compound unique id based on document name, EHE0009 id and EHE0039 id.

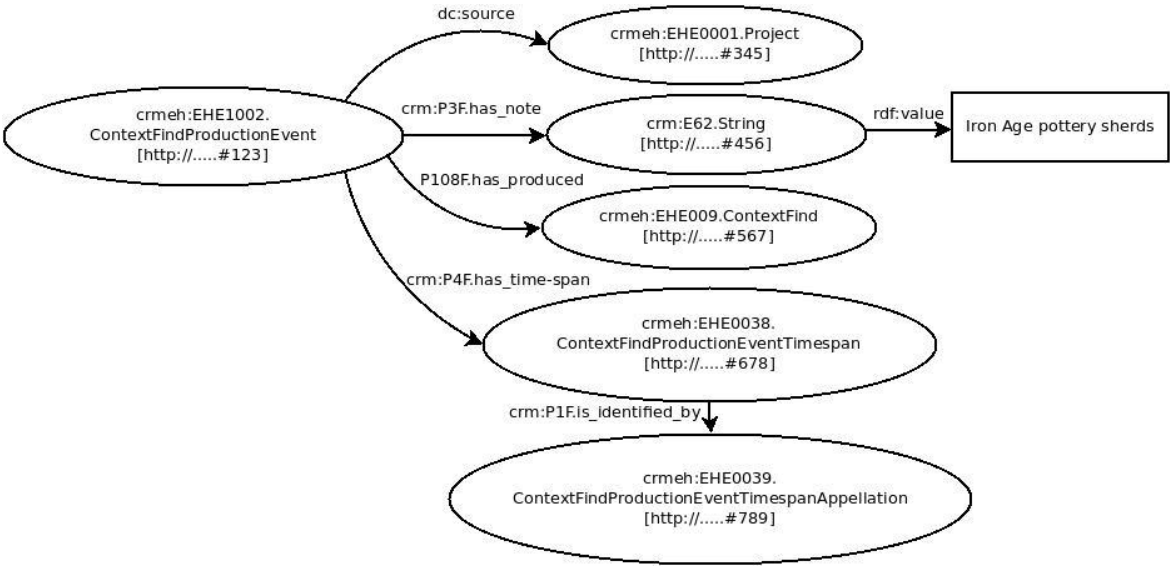


Figure 7.7: EHE1002.ContextFindProductionEvent graph

The above figure produces the following RDF code.

```
<crmeh:EHE1002.ContextFindProductionEvent
      rdf:about="http://tempuri/star/base#suffolkc1-6115.286616">
      <dc:source rdf:resource="http://tempuri/star/base#suffolkc1-
      6115" />
      <crm:P3F.has_note>
            <crm:E62.String>
            <rdf:value>Iron Age pottery sherds</rdf:value>
       </crm:E62.String>
       </crm:P3F.has_note>
      <crm:P4F.has_time-span>
      <crmeh:EHE0038.ContextFindProductionEventTimespan
      rdf:about="http://#suffolkc1-6115.286616.286969">
          …
       <crm:P108F.has_produced rdf:resource=
      "http://tempuri/star/base#suffolkc1-  6115.286777"/>
</crmeh:EHE1002.ContextFindProductionEvent>
```

7.2.2.8 EHE1004.ContextFindDepositionEvent

The graph describes the various triples of a EHE1004.ContextFindDepositionEvent resource. The resource connects an EHE0007 and an EHE0009 resource while it links to a *EHE0001.Project* entity. The resource has a note of type String holding the phrase instance in which the two constitution entities are found. The resource does not have a particular value other than the note of the phrase thus it is not represented by a SKOS terminological reference.

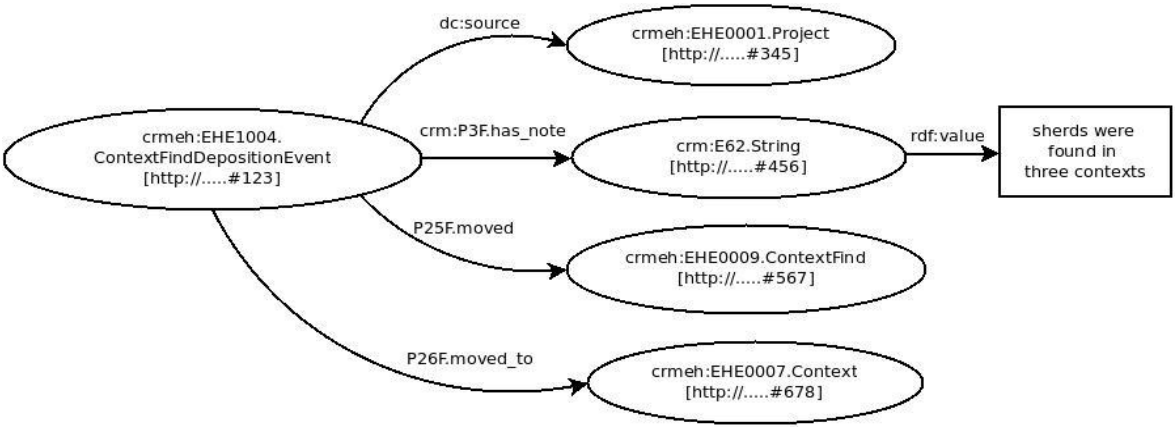


Figure 7.8: EHE1004.ContextFindDepositionEvent graph

The above figure produces the following RDF code. The example shows how linking is achieved between the resources EHE0007 and EHE0009

```
<crmeh:EHE1004.ContextFindDepositionEvent
      rdf:about="http://tempuri/star/base#suffolkc1-6115.286645">
      <dc:source rdf:resource="http://tempuri/star/base#suffolkc1-
      6115" />
      <crm:P3F.has_note>
            <crm:E62.String>
            <rdf:value>sherds were found in three
            contexts</rdf:value>
           </crm:E62.String>
      </crm:P3F.has_note>
      <crm:P26F.moved_to rdf:resource=
      "http://tempuri/star/base#suffolkc1-6115.286831" />
      <crm:P25F.moved rdf:resource=
      "http://tempuri/star/base#suffolkc1-6115.286830" />
</crmeh:EHE1004.ContextFindDepositionEvent>
```

7.3 The Andronikos Web Portal

The purpose of the web portal Andronikos (http://www.andronikos.co.uk) has been discussed in section 3.5.The name of the portal is inspired by Manolis Andronikos (1919 - 1992) a famous Greek archaeologist from the Aristotle University of Thessaloniki, whose excavation efforts brought to light the palace of Vergina (1977) and what is believed to be the tomb of Philip II of Macedon.

The portal is an Apache 2, MySQL, PHP development which provides access to the semantic indices of 2460 OASIS grey literature documents. It makes available a range of HTML pages that enable browsing and human inspection of the indexing results and support retrieval strategies that utilise the semantically enriched document views and metadata. The document browsing facility is driven by server-side scripting (PHP) and MySQL database employing GUI elements (i.e. drop-down boxes) for accessing and filtering pages.

The final evaluation results are available via the portal along with the source texts (see Sample Documents or access all reports via Home). The pages support 3 distinct views of the documents that correspond to the 3 different sets of semantic annotation sets, namely Pre-process, NER (CRM) and RE (CRM-EH). The portal makes available the (existing) author-based OASIS metadata, together with links to the XML and RDF versions of the semantic indices. The indices can be either inspected within the web browser environment or can be downloaded for further use and manipulation by third party applications.

The Term Overlap application, discussed in section 4.3.2 and various knowledge resources have access currently restricted to members of the STAR project, since the knowledge resources are owned by other organisations.

7.3.1 OASIS Metadata View

The OASIS metadata view (Figure 7.9) presents a range of author-based metadata that accompany the documents. The metadata include information about authors and location of the excavation such as parish, county, place and grid references. Also the metadata capture some basic information about finds and monuments which are associated with the document. Such metadata descriptions are very basic and abstract on a document level and cannot be compared with the richness of the semantic annotation metadata that reflect information snippets at a contextual level. In addition, not all documents contain metadata about document finds and monuments.

The following table 7.9 presents an example of existing OASIS metadata for an individual document. The metadata set contains information about the document author (database key to the author table), the location and grid reference of the excavation. Also the metadata table contains information on find and monument types relating to the document. The list includes "ditch", "ridge and furrow" and "cremation" monument types and "human remains" find type. Compared with the list of six top semantic annotations metadata for the same document as seen below (Figure 7.9), the list of OASIS metadata for finds and monument types is rather limited. The semantic annotations contain a richer set that describes terminological (SKOS) references and term frequencies while both OASIS and semantic annotation metadata agree on the monument types that are relevant to the document. The OASIS metadata offer an abstract view of the document via author-based entries which can be used to satisfy retrieval on a generic document level while the CRM semantic annotations are richer and more extensive supporting finer information retrieval needs as discussed below (section 7.4.1) (Falkingham 2005).

Land west of Orchard House, St Ives, Cambridgeshire: Evaluation

Wessex Archaeology - 2007

Authors	Location	Grid Reference	Finds - Context
2174 - AUTHS 2174 - OBIB	**Site:** 2174 **County:** Land west of Houghton Road **District:** CAMBRIDGESHIRE **Parish:** HUNTINGDONSHIRE **Place:** SAINT IVES	**Type:** 72365 **Mcode:** 2174 **Easting:** TL **Northing:** 29925	**Reference:** 2174 **Desrciption:** FINDS **Type:** HUMAN REMAINS **Reference:** 2174 **Desrciption:** MONUS **Type:** DITCH **Reference:** 2174 **Desrciption:** MONUS **Type:** RIDGE AND FURROW **Reference:** 2174 **Desrciption:** MONUS **Type:** CREMATION

E53.Place

TERM	SKOS	Count
pit	ehg003.55	10
cremation burial	70018	7
ridge and furrow	68628	7
ditch	ehg003.20	6
context	ehg003.137	5
cemetery	70053	3

Figure 7.9: OASIS Metadata table of document (Authors, Location , Grid Reference, Finds-Context) compared with CRM semantic annotation metadata

7.3.2 Pre-processing View

The Pre-processing view makes available information which was extracted during the first phase of the pipeline (section 3.3.1). The Pre-processing phase produced a range of annotations, such as tokens, noun-phrases and verb-phrases which are used by the succeeding NER and RE phases of the pipeline. The phase also aims at detecting document heading and table-of-content (TOC) spans as well as document summary sections. The aims and objectives of the Pre-processing phase are discussed in chapter three, section 3.3.1.

As discussed earlier (chapters 5 and 6), the main focus of the OPTIMA pipeline are the tasks of NER and RE with regards to CRM and CRM-EH ontologies respectively. However, the annotation results of the Pre-processing phase are also incorporated into the portal structure since they constitute a useful resource which provides an abstract view of the documents, revealing summary section and document headings. The view enables users to scan through document elements such as headings, TOC and summary sections, in order to obtain a quick view on the document structure and areas of discussion. The Pre-processing annotations of span and sections types are treated as a "bonus" and included in the portal, although not always accurate as discussed by the examples below.

An example of table-of-contents annotation is shown on figure 7.10 which presents the grouping of different spans. Grouping of spans is displayed by table cells. As observed, there are more than one table cells per TOC. This particular graphical arrangement is the result of the discontinuity of the extracted spans as shown in right part of figure 7.10. The extraction rule fails to recognise that the span "Bronze Age)" is part of the span "3.2 Phase I (Late Neolithic/Early Bronze Age)" due to the additional carriage return (new line), which is introduced after the forward-slash character. Therefore, the pipeline may deliver more than one TOC annotation per document although all annotations of that kind originate from the same document section. This particular pipeline behaviour does not limit the capability of the system to extract TOC spans with some accuracy. Hence, TOC annotations are included in the portal since they can be used to support search strategies that may require information with regards to the document structure.

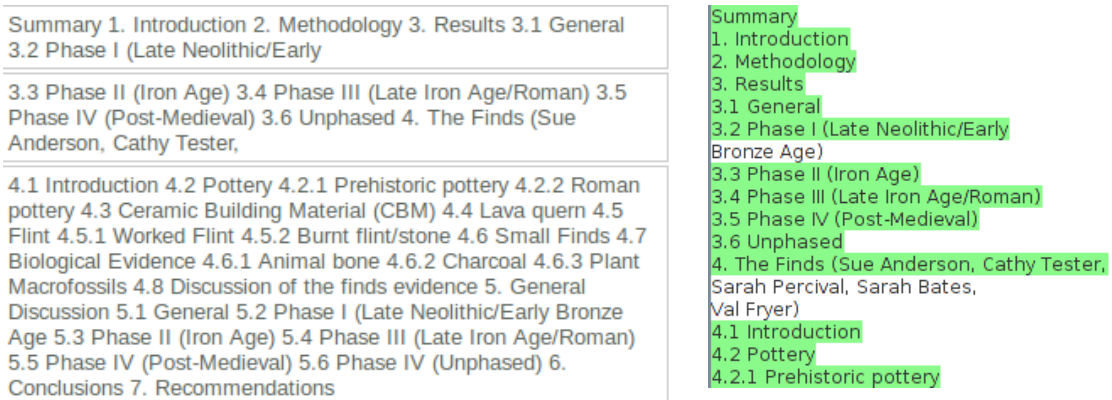


Figure 7.10: View of TOC annotation type. On the left, annotations as presented in the web portal Andronikos. On the right, annotations as appear in the GATE environment.

Similar to the annotation of TOC spans, the Pre-processing pipeline detects and annotates document heading spans. The initial intention, as discussed on chapter 3 section 3.3, was to detect such spans in order to exclude them from the NER process since they might make use of CRM concepts in an abstract and context independent fashion. On the

other hand, such heading annotations can be used to support retrieval strategies which are informed by document structure. Used as an "X-ray" of the document content, heading annotations can be used to inform users about the structure and sections contained in a document. Thus, such annotation types are regarded as useful and included in the web portal.

Figure 7.11 presents heading annotations as they are delivered by the web-portal and as they are detected by the pipeline in the GATE environment. In the particular example below, the span "Sue Anderson" (a report section author) does not correspond to a valid heading but is detected as such because it resembles a heading like structure, due to its upper-case usage and location between two valid heading spans. Although the heading detection rules are not always accurate, the inclusion of the heading annotations by the web-portal is considered to be potentially useful and supportive to document-centred retrieval strategies

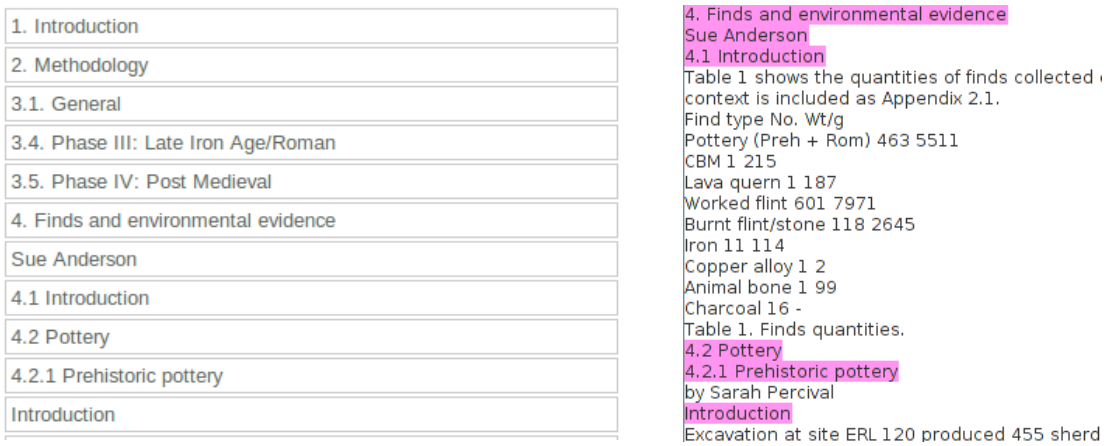


Figure 7.11: View of heading annotations. On the left, annotations as presented in the web portal Andronikos. On the right, annotations as appear in the GATE environment.

The Pre-processing phase also detects summary sections of documents as discussed in section 3.3.1.2 .Such document summary section are strongly representative of a grey literature report, revealing core and significant information. Detection and annotation of such sections can be used to support retrieval strategies that prioritise particular document sections, i.e. weighting matches from summary sections as more important than other matches. The following example discusses the major Phases of an excavation that reflect major periods covered by site interpretation. The summary reveals the activity of the site that spans from the Late Neolithic to Post Mediaeval period and the major excavation finds that are associated with the phases, such as Beaker pottery, deposits of charcoal and metallic objects. The web-portal makes available the summary sections of documents that are successfully extracted, supporting a quick overview of a grey literature document.

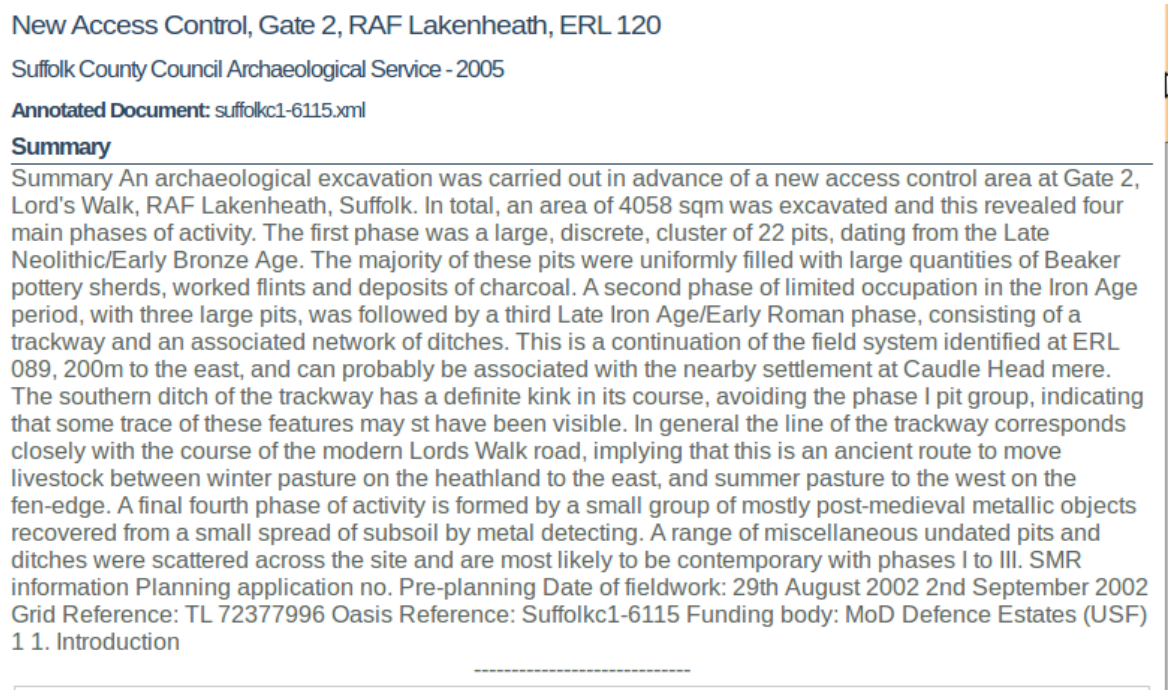


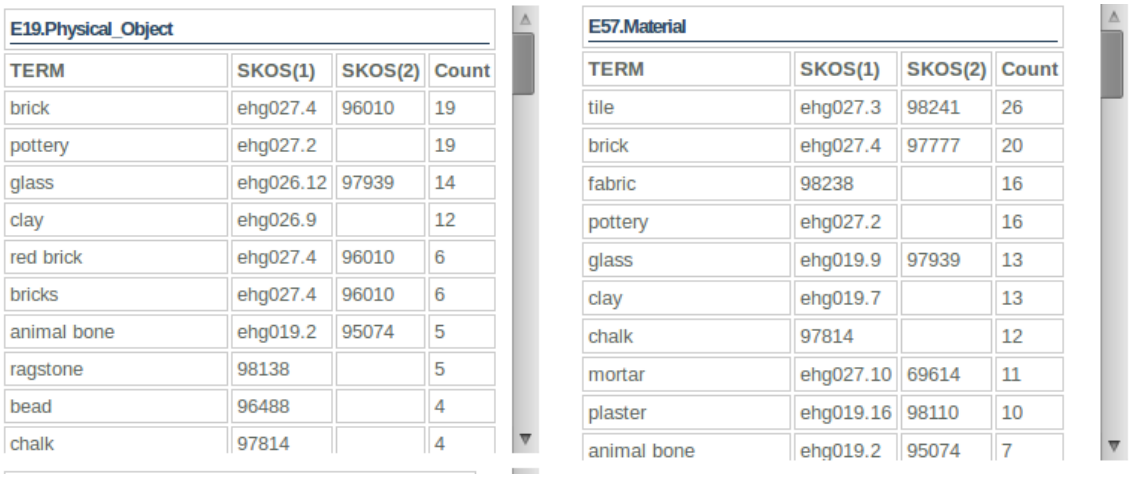
Figure 7.12: Pre-processing view of a summary section in Andronikos web-portal.

7.3.3 NER CRM View

The NER view makes available semantic annotations of the four CRM entity types (*E19.Physical Object, E49.Time Appellation, E53.Place,* and *E57.Material)* that are targeted by the OPTIMA pipeline, plus the negated phrases delivered by the Negation Detection phase. The web-portal pages provide access to two distinct views of CRM document annotations. The HTML view (Figure 7.13) presents the four entity types in tabular format decoupled from content. The view arranges the CRM annotations of each document by number of occurrences, with the most frequent mentions shown first. Also the tabular view presents the text string and the SKOS terminological reference(s) associated with each annotation. The tables offer a compiled view of the CRM annotations enabling quick but also thorough inspection of document annotations. Thus the view can be used to support semantic indexing at the document level based on a threshold selection of the most frequent annotations.

The second view, XML (Figure 7.14), couples the semantic annotations of the four CRM entities with content. The coupled XML files were produced by the Flexible Exporter GATE module and were post-processed by a bespoke PHP script (section 7.2) for adding XML name-space information, root tag definition and link to presentation file. Attaching a bespoke Cascading Style Sheet (CSS) presentation file to the XML files enables inspection of annotations in context using highlight colours. The coupled view of semantic annotations and contexts is useful for directing users' attention to particular

document sections that present higher concentrations of annotations. Using colours to highlight textual instances of CRM entities helps users identify particular types of semantic annotations in context. In addition, all XML documents made available by the portal are downloadable for further manipulation by advanced users. XML manipulation can be directed towards database population, generation of HTML views, or other forms of abstraction that may be useful to user needs.



E19.Physical_Object

TERM	SKOS(1)	SKOS(2)	Count
brick	ehg027.4	96010	19
pottery	ehg027.2		19
glass	ehg026.12	97939	14
clay	ehg026.9		12
red brick	ehg027.4	96010	6
bricks	ehg027.4	96010	6
animal bone	ehg019.2	95074	5
ragstone	98138		5
bead	96488		4
chalk	97814		4

E57.Material

TERM	SKOS(1)	SKOS(2)	Count
tile	ehg027.3	98241	26
brick	ehg027.4	97777	20
fabric	98238		16
pottery	ehg027.2		16
glass	ehg019.9	97939	13
clay	ehg019.7		13
chalk	97814		12
mortar	ehg027.10	69614	11
plaster	ehg019.16	98110	10
animal bone	ehg019.2	95074	7

Figure 7.13: HTML table view of semantic annotations of a single document for the CRM entities E19.Physical_Object and E57.Material

The above figure 7.13 presents the table view of semantic annotations of a single document with respect to the CRM entities E19.Physical_Object and E57.Material. As discussed in chapter 4 section 4.3.3, there is significant overlap between the terminological resources associated with these particular CRM entities. The level of overlap is reflected by the above tables which contain overlapping terms such as brick, glass, pottery, etc.

The tables contain 4 different field types, namely “Term”, “SKOS(1)”, “SKOS(2)” and “Count”. Under the “Term” field the tables hold the textual values of the semantic annotations; thus “brick”, “bricks” and “red brick” are all different table entries. However, all E19.Physical_Object table entries share the same SKOS(1) and SKOS (2) references but a different terminological reference when in table E57.Material. The fields SKOS(1) and SKOS(2) correspond to the two terminological references that each term can have; one reference originating from a glossary commencing with “ehg” and another originating from a thesaurus. Therefore, based on the mapping between terminological resources and ontological classes as discussed in section 4.3.2 (Table 4.1), the material sense of “brick” is associated with the pair of references “ehg27.4 - 97777” and the physical object sense with the pair “ehg27.4 – 96010”.

Figure 7.14 below shows a document section in which the disambiguation of material-physical object sense has been achieved for the concepts of "brick" and "pottery". The example shows that the terms "brick" and "pottery" can have a material or a physical object sense depending on contextual arrangements. Thus, in the cases "red brick wall" and "brick wall" the sense is material, while in the case "The brick from the floor was dated to the late 18th century", the sense is physical object. Similarly, in the case "Beaker pottery sherds", "pottery" has a material sense while in the case "pits contained Early Bronze Age pottery" , the sense is physical object. Different senses (i.e. CRM entities) are highlighted in different colours; orange is used for physical object and purple for materials.

The examples also show 3 different samples where, place type entities highlighted in green and time appellation entities highlighted in blue. Place entities contain both individual archaeological contexts, such as "fills" and "pits" and larger grouping of contexts, such as "walls" and "floor". Time appellations can be single, i.e. containing only one concept, such as "Early Bronze Age", or compound i.e. containing two concepts, such as "Late Neolithic/Early Bronze Age". In both cases time appellation concepts may contain moderators, such as "later", "early", etc.

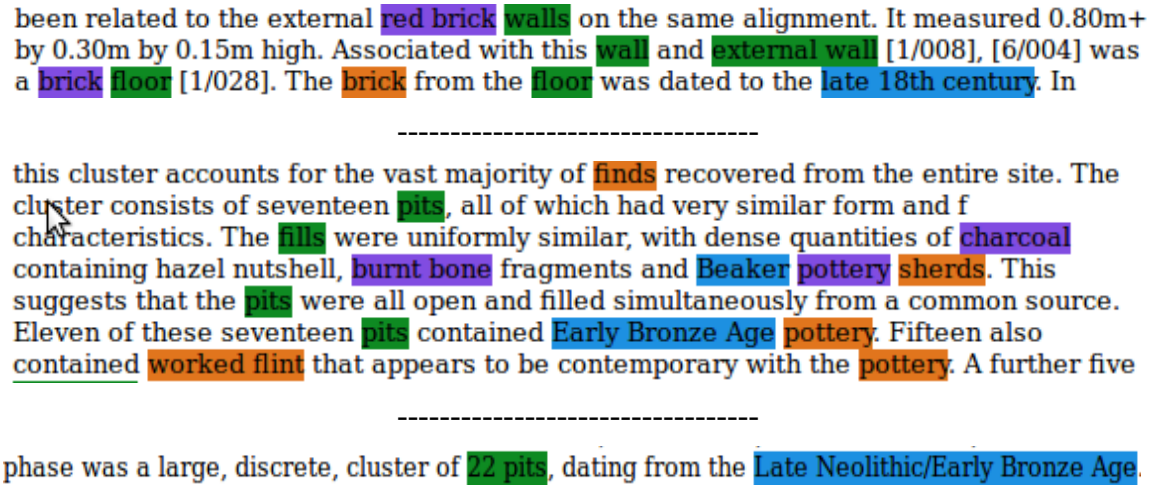


Figure 7.14: CRM annotations in context, XML file attached to presentation CSS file for highlighting different annotation types from three different samples.(Keys; Orange: E19.Physical_Object, Blue:E49.Time Appellation, Green:E53.Place, Purple:E57.Material)

The figure 7.15 below presents the table view of semantic annotations of a single document with respect to the CRM entities E49.Time_Appellation and E53.Place. The tables show the dual terminological reference scheme that is applied to terms when necessary. For example, the time appellation case "late Iron Age / early Roman" is assigned two distinct terminological references corresponding to the two concepts that are compound in the phrase. The dual assignment resolves the absence of any intermediate

concept of this kind in available terminological resources. On the other hand, in the case of place entities such as "pit" and "ditch", the dual terminological reference corresponds to the appearance of the concept in two different resources, i.e. glossary and thesaurus. As discussed below (section 7.4.1.3), a retrieval application can exploit either terminological reference to produces matches.

The place table also contains both individual archaeological contexts, such as "fill", and larger interpretive groupings, such as "trackway". The use of moderated terms is also apparent, such as "circular pit". This term enjoys the same terminological reference as "pit", since the moderator has been detected by the NLP adjectival expansion module. On the other hand, "early Bronze Age" and "Bronze Age" enjoy two distinct terminological references because the two are well defined concepts, which have been Skosified at the level of gazetteers.

E49.Time Appellation

TERM	SKOS(1)	SKOS(2)	Count
beaker	134731		45
iron age	134722		17
roman	134738		14
early bronze age	134732		13
bronze age	134723		11
prehistoric	134718		9
late iron age/early roman	134738	134737	8
post-medieval	134746		7
later neolithic	134721		5

E53.Place

TERM	SKOS(1)	SKOS(2)	Count
fill	ehg003.142		76
pit	ehg003.55	70398	45
ditch	ehg003.20	70351	42
pits	ehg003.55	70398	36
trackway	70270		22
circular pit	ehg003.55	70398	18
ditches	ehg003.20	70351	16
surface	ehg003.82		8
posthole	ehg003.62		8
context	ehg003.137		8

Figure 7.15: HTML table view of semantic annotations of a single document for the CRM entities E49.Time_Appellation and E53.Place

The CRM view also delivers the results of negation detection. Figure 7.16 presents the set of negated phrases detected from a single document. The example shows the kind of phrases which are extracted by the negation module. Such phrases are excluded from producing any CRM annotations. Thus, in the example "no traces of a Roman settlement", neither "Roman" nor "settlement" are annotated as CRM entities.

Negation

TERM	Count
no dating evidence within the deposit as the tile	1
no dating evidence was recovered from this layer	1
no evidence of burning to indicate the orientation of the hearth	1
no wall	1
no dating evidence for these deposits	1
no traces of a roman settlement	1

Figure 7.16: HTML table view of negated phrases of a single document

7.3.4 CRM-EH Relation Extraction View

The CRM-EH view presents, similarly to the CRM view, an HTML tabular view of document annotations and an XML version of annotations in context. The portal pages make available the CRM-EH annotations delivered by the CRM-EH Relation Extraction phase of the pipeline. Thus, the pages of this view include tabular and contextual views of the specialised CRM-EH entities but most importantly the pages present the results of the Relation Extraction phase of the pipeline.

In detail, the views present the CRM-EH entities; *EHE0007.Context*, *EHE0009.ContextFind*, *EHE0026.ContextEventTimeSpanAppellation*, *EHE0030.ContextFindMaterial*, *EHE0039.ContextFindProductionEventTimeSpanAppellation* and the CRM-EH event entities; *EHE1001.ContextEvent*, *EHE1002.ContextFindProductionEvent*, *EHE1004.ContextFindDepositionEvent* and the property *P45.consists_of.* In addition, the portal provides links to the XML and RDF versions of the CRM-EH semantic indices, which are downloadable for further use and manipulation, as discussed in section 9.1.1.

The delivery of XML and RDF versions of annotations is discussed above in section 7.2. The transformation of semantic annotations to RDF triples serves the purpose of semantic indexing of grey literature documents aimed at document retrieval and cross-searching with respect to the CRM-EH ontological definitions. The role and usage of semantic indices by an information retrieval application, in this case the STAR demonstrator, is discussed below in section 7.4.

The HTML view of CRM-EH annotations makes available in a tabular format the total number of relations that have been identified from each document, either in the form of CRM-EH event entities or as CRM properties.

The table provides an abstract view of the total number and type of entities identified in a document, which allows users to obtain basic information about the relation extraction result. Figure 7.17 presents the figures for a particular document (suffolkc1-6115). Overall, 114 relation phrases were identified in the document containing 14000 words. The CRM-EH view arranges the extracted phrases under CRM-EH event and properties categories. Figure 7.18 below presents examples of the CRM-EH event

Annotation	Count
EHE1001.ContextEvent	30
EHE1002.ContextFindProductionEvent	24
EHE1004.ContextFindDepositionEvent	41
P45.consists_of	19

Figure 7.17: Frequency and type of relation extraction phrases identified in a document

entity *EHE1001.ContextEvent*. This particular event entity is used to model phrases that relate archaeological context with time appellation. Such phrases can be complex, containing more than two argument entities, or they can be simple, consisting of two argument entities in a sequence. For example, the simple case "settlement at Houghton, which dates from the early medieval period", relates a settlement with the Early Medieval period. On the other hand, the phrase "Early Romano-British burials and possible settlement" is short but complex and relates two archaeological contexts with a time appellation. The use of the adjectival moderator ("possible") adds uncertainty. This is characteristic of natural language but less common in data entry with exception the free text database fields.

The tabular view of the portal makes available the extracted phrase and the argument entities. The argument entities of the relation enjoy an ontological reference (e.g. *EHE0007.Context)*, a terminological reference (e.g. 68977) and a string value (e.g. "settlement"). The contextual view highlights the annotations in context using colour. The light green colour is used to highlight instances of the entity EHE0007.Context, the light blue is used to highlight *EHE0026.ContextEventTimeSpanAppellation* instances, while the dark green is used to highlight the whole phrase that relates the above two entities. Due to the use of CSS version 2, transparency is not possible, hence the dark green colouring is overlapped by the entity colours. In some cases, where the phrase is very short, the dark green colour appears only between the two argument entities, as for example in the case of "Anglo-Saxon cemeteries".

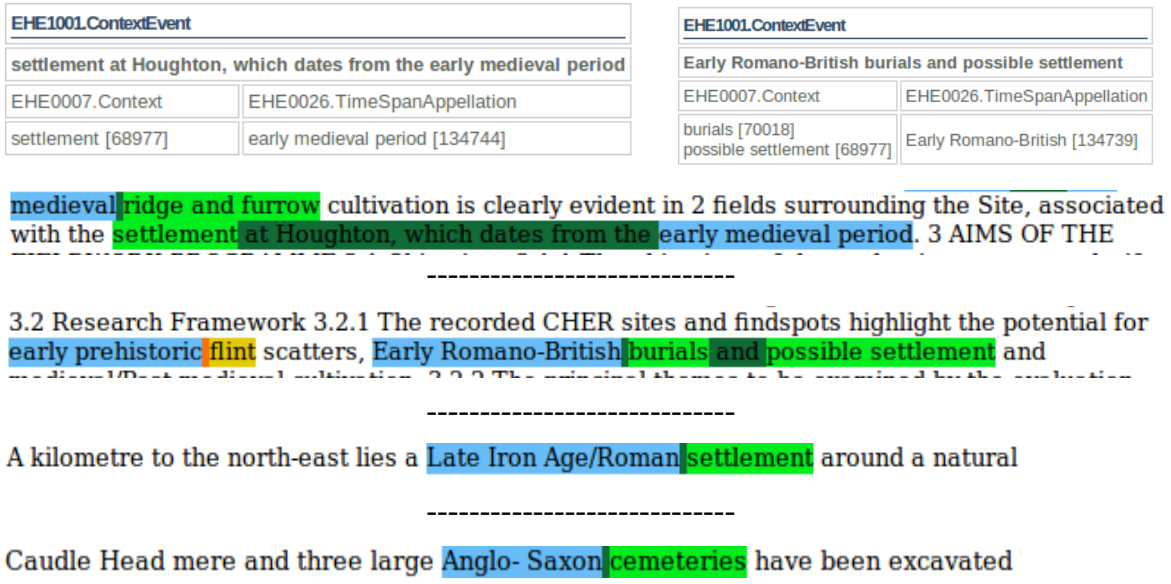


Figure 7.18: Tabular format and contextual view of examples of relation extraction phrases with respect to the CRM-EH event entity EHE1001. ContextEvent

The *EHE1002.ContextFindProductionEvent* entity is used to model phrases that relate archaeological finds with time appellations. As previously mentioned, phrases can be complex; they may contain more than two argument entities or can they can be simple, consisting of two argument entities in a sequence. For example, the phrase "brick was a mixture of type dating to the late 16th to 17th century", relates the context find "brick" with a compound time appellation span containing two distinct appellations (Figure 7.19). The phrase also reveals the role of the word sense disambiguation module, where in this case the physical object sense of brick is correctly resolved.

The phrase "arrowhead, also of later Neolithic date" does not make use of a verb but the syntactical evidence is enough to relate the context find "arrowhead" with the time appellation "later Neolithic". Simple cases of *EHE1002* instances may relate two argument entities in a sequence, as for example the phrase "Roman Finds" where the dating information is implicitly declared. The portal, as discussed above, uses tabular elements to present the extracted phrases and entity details and contextual highlighting of annotations to present annotations in context. The highlight colours used for these event types are; light brown for finds, light blue for time appellations and orange for the whole event phrase.

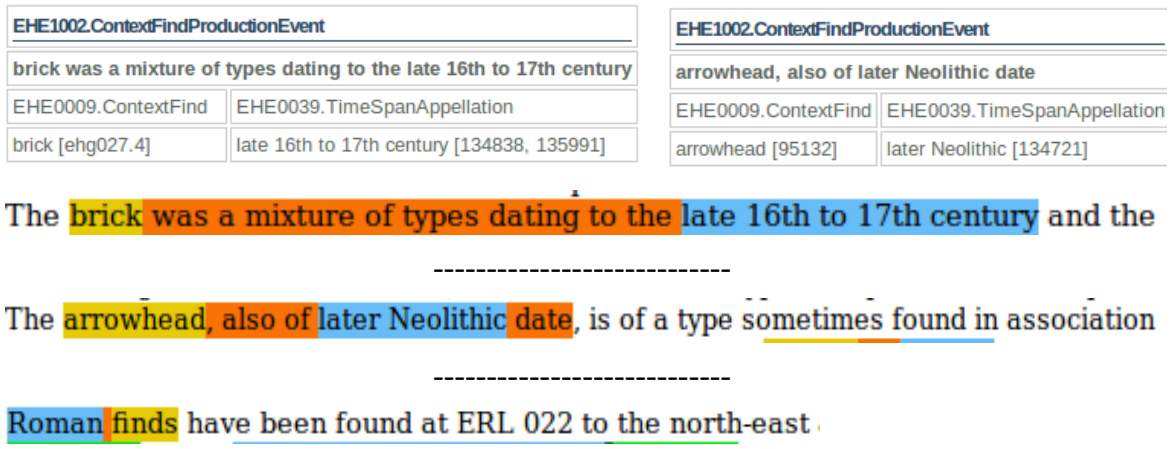


Figure 7.19: Tabular format and contextual view of examples of relation extraction phrases with respect to the CRM-EH event entity EHE1002. ContextFindProductionEvent

The *EHE1004.ContextFindDepositionEvent*entity is used to model phrases that relate archaeological finds with archaeological contexts. Such phrases are usually more complex and harder to extract than the production and context event cases above. In most cases, the phrases contain verb structures connecting the event arguments. Simple cases might connect event argument via prepositions. For example, the phase "animal bone in pit" (Figure 7.20) relates the find "animal bone" with the archaeological context "pit" via the preposition "in". In other more complex phrases, such as "pits contained Early Bronze Age

pottery", the event arguments (pits – pottery) are connected via a verb ("contained").

In the above example, "pottery" is participating in two events, a deposition event with "pits" and a production event with "Early Bronze Age". Hence, "pottery" can be used as a semantic hub in order to connect a time appellation to the archaeological context ("pit"). Generally in OASIS reports (i.e. reports following analysis of all the finds) when a date is mentioned for a find, there is an assumption that the find's date has been taken as diagnostic of the context in which it was found. Moreover, phrases can be long as in the case "copper alloy artefacts were recovered from various contexts". The phrase relates "artefacts" with "contexts" in their plural form not at an individual (unique item) level.

The portal, as discussed above, uses tabular elements to present the extracted phrases and entity details and contextual highlighting of annotations to present annotations in context. The highlight colours used for this event type are; light brown for finds, light green for archaeological contexts and olive for the whole event phrase.

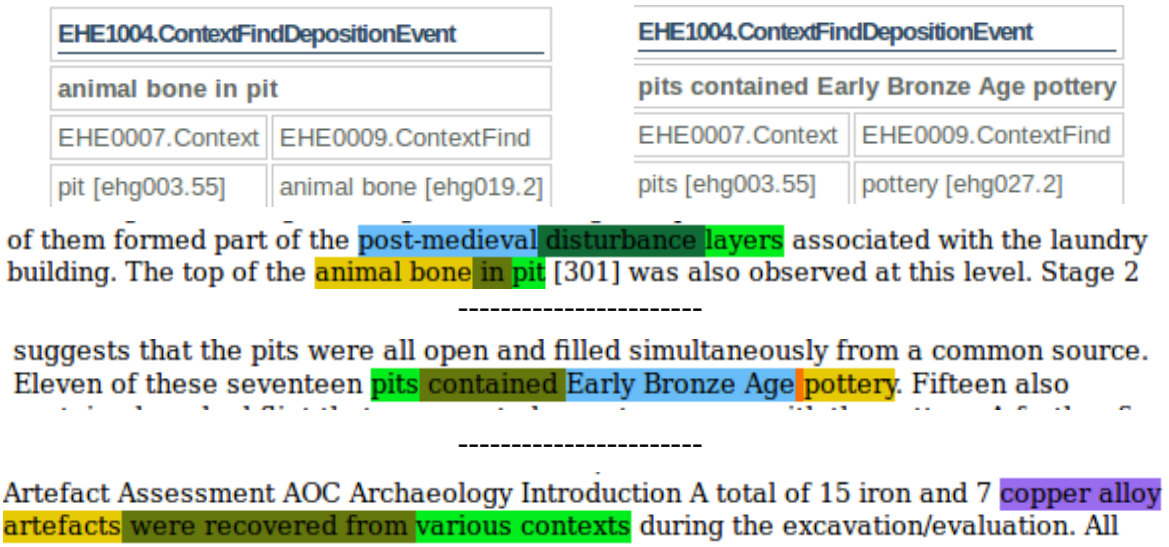


Figure 7.20: Tabular format and contextual view of examples of relation extraction phrases with respect to the CRM-EH event entity EHE1004. ContextFindDepositionEvent

The last kind of relation extraction phrase which is made available in the portal concerns the CRM property *P45.consists_of* (Figure 7.21). Phrases of this CRM property relate a physical object with material and particularly in the case of the CRM-EH ontology, relate an archaeological find with a material. Such phrases are less complex than the event phrases discussed so far. Usually the *consists_of* relation contain two entity arguments, which in most cases are found in sequence or connected via a proposition. For example the phrase "iron slag" relates implicitly the physical object "slag" with the material "iron". Similar examples are the phrases "pottery sherds" and "flint flakes".

More complex phrases usually connect the relation arguments via a proposition, as for

example the phrase "single sherd of Roman pottery" where "single sherd" and "pottery" are connected via the proposition "of". Use of verb-phrases is not frequent, possibly a verb-phrase such as "made of" might be used to relate arguments of a *consists_of* relation but usually the verb "made" is unnecessary, as in the above example. The complex phrase also reveals the use of the adjectival expansion module (section 5.7) which has expanded "sherd" to "single sherd" and the occurrence of overlapped event/property phrases. In this case, "single sherd" is related to the time appellation "Roman" via a production event and with the material "pottery" via a *consists_of* property. The portal presents the results of the *consists_of* phrases in a tabular format, similarly to the other types of events, with contextual annotations highlighted in colour, in this case dark purple for material and light purple for the whole phrase.

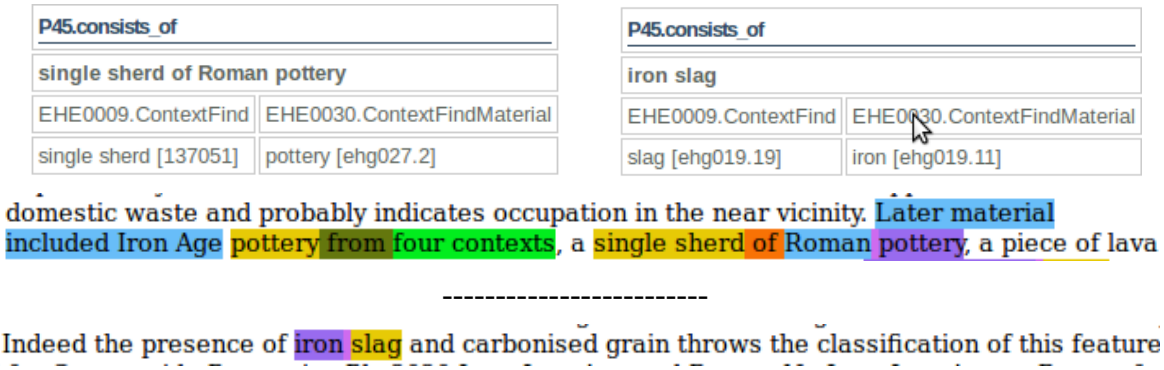


Figure 7.21: Tabular format and contextual view of examples of relation extraction phrases with respect to the CRM property entity P45.consists_of

The discussion so far has revealed the main three levels of document abstraction (Pre-processing, NER, CRM-EH RE) made available by the Andronikos web portal. Each abstraction delivers a set of annotation types in an interoperable format. The format is XML that couples annotations and context. Decoupled RDF graphs of annotations are also produced for annotation types produced during the CRM-EH Relation Extraction phase of the pipeline. Such RDF document abstractions can be used as semantic indices supporting information retrieval. The following section discusses the use of such RDF semantic indices of grey literature documents by the STAR demonstrator for enabling retrieval via semantic queries with respect to the CRM-EH ontology.

7.4 The STAR Demonstrator

The STAR demonstrator is a web application, outcome of the STAR project, aiming to support cross-searching scenarios between a range of disparate archaeological datasets and grey literature documents. The semantic indices of grey literature produced by the above stage are integrated into the demonstrator's architecture enabling document retrieval and cross search with archaeological datasets. Grey literature indices can be exploited by the current demonstrator to support queries specified as archaeological contexts, finds and materials. Time periods are currently supported. The search mechanism is driven by a SPARQL engine and automatically builds complex semantic queries, which correspond to the user interaction with the interface controls. For example, the user builds a query for finding archaeological contexts of a particular type, i.e. "hearth", which contains a find of a particular kind, i.e. "coin". The underlying mechanism translates the user selection to a complex SPARQL query that conforms to the ontological model.

The semantic search is based on controlled (URI) identifiers of the vocabulary which support cross-searching between the datasets and grey literature. As users enter query concepts into the type fields, controlled types are automatically suggested for selection. In addition, queries can be targeted to the ontological classes without invoking a controlled vocabulary input. For example search for any type of archaeological context that contains any type of find. By default all datasets and grey literature are included by the search mechanism, however the interface enables users to dynamically specify searches on particular datasets or grey literature. The demonstrator is browser agnostic and it has been tested on a range of commercial web browsers.

7.4.1 STAR Demonstrator Cross-Search Scenarios

The following example scenarios present useful and challenging semantic searches enabled by semantic indices of grey literature documents. The scenarios demonstrate the semantic capabilities of the indices to support CRM-EH oriented queries that answer polysemous ambiguity, orthographic-synonymy, topicality and retrieval with respect to ontological relationships. In addition, the following examples present cross-searching results between grey literature and datasets accessible by the STAR demonstrator (http://hypermedia.research.glam.ac.uk/resources/star-demonstrator/). Demonstrator results originating from grey-literature commence with a "#" followed by the unique identifier of document while the dataset results are numerical.

7.4.1.1 Polysemous Ambiguity

Polysemous is a word that potentially has more than one sense. Semantic searches that deal with polysemous words are capable of delivering results that correspond to the intended word sense. The following example presents the results of a query on retrieving documents containing the concept of "cut" in the sense of archaeological context (EHE0007.Context). The result of the *#lparchae1-20549_1* document (Figure 7.22) corresponds to an archaeological context (linear cut) found in the phrase **"***two pits, a posthole and a linear cut, which are broadly dated from the Neolithic period to the Late Bronze Age*"

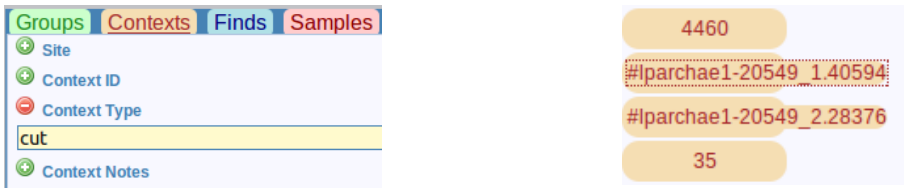


Figure 7.22: Partial list of search results dealing with polysemous ambiguity of concept "cut"

In another similar query example (Figure 7.23), the demonstrator retrieves the document #heritage1-4830 for the concept of "well" in the phrase; "A brick-lined well [2025] lay at the eastern edge and a crescent-shaped area of red brick"

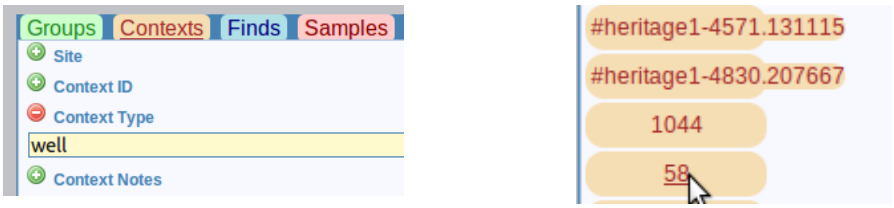


Figure 7.23: Partial list of search results dealing with polysemous ambiguity of concept "well"

7.4.1.2 Orthographic-Synonym Definition

Synonyms are different words that share the same meaning. Orthography is concerned with the spelling of words, for example variation between singular and plural forms. Semantic searches that deal with orthographic and synonym variations are capable of retrieving results at the level of concepts, independent of spelling variations. The following example presents (Figure 7.24) the results of a query on retrieving documents containing the concept of "human remains" in the sense of an archaeological find (EHE0007).

The result from the #norfolka1-22647_1 document corresponds to a "human bone" recovered from a ground fill during excavation. The result originates from the phrase "A quantity of human bone including three skulls was recovered from its fill" and demonstrates the ability of document indexing to support retrieval via synonyms.

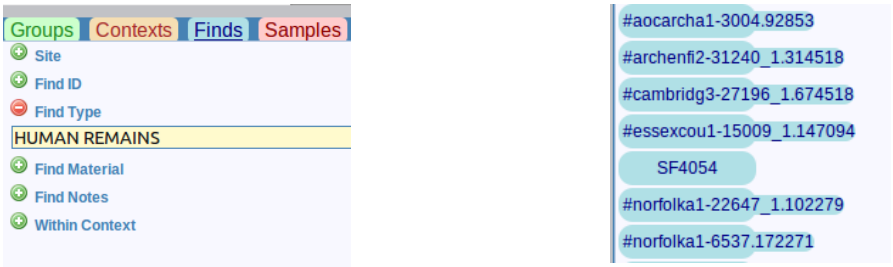


Figure 7.24: Partial list of search results of retrieval via synonyms for the concept "human remains"

7.4.1.3 Topicality

Topicality is a specific form of polysemy, previously discussed as ontological polysemy (section 5.6.2) where the sense of a word is influenced by the context in which the word is used. For example, the word "brick" can be either a physical object or a material, depending on its contextual use. Semantic searches that deal with the issue of topicality are capable of retrieving results that correspond to the correct sense. The following example presents (Figure 7.25) the results of a query on retrieving documents containing the concept of "brick" in the sense of an archaeological find (EHE0007.ContextFind).

The result of the #albionar1-15196_1 document corresponds to "bricks" of the Roman period, found in the phrase "Ceramic artefacts included pottery sherds, roof tiles and bricks all dated to the Roman period". A similar result originating from document #aocarcha1-17523_17, phrase "Post Medieval period; the stones location, in context with post medieval bricks, suggests that it originated as a piece of masonry during the latter period" corresponds to "Post Medieval bricks" that are described as pieces of masonry.

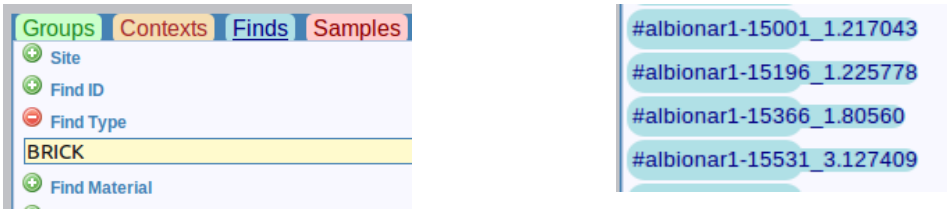


Figure 7.25: Partial list of search results of correct sense retrieval for the concept "brick" in the sense of archaeological find (EHE0009.ContextFind)

On the other hand, when the query is targeted at the concept of "brick" in the sense of the material of an archaeological find (Figure 7.26), the system retrieves documents (#northpen3-21389_1 and #suffolkc1-17649_1) containing phrases like "*A layer of small brick tiles forming the street paving was removed*" and " *comprising loose mortar, brick flint fragments with common charcoal inclusions*". In both cases, the concept of brick is used in the sense of the material of an object, as in the case of "small brick tiles" .

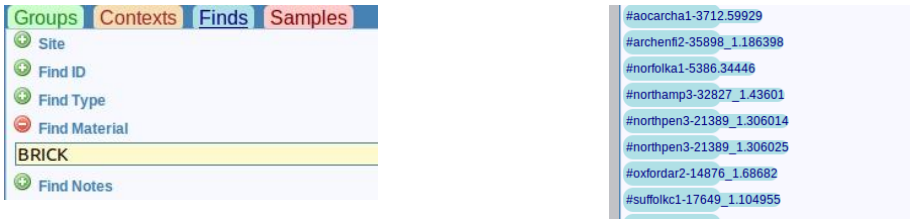


Figure 7.26: Partial list of search results of correct sense retrieval for the concept "brick" in the sense of the material of an archaeological find (EHE0030.ContextFindMaterial)

7.4.1.4 Ontological Relationships

The following search scenarios demonstrate the capability of semantic indices to support document retrieval with respect to ontology entity relationships. The demonstrator's interface enables users to specify the type of entities and the kind of relationship participating in a query expression. The results from grey literature are based on the semantic indexing of documents with respect to CRM-EH events and properties as discussed above (section 6.2.2). The demonstrator enables users to query relationships between a sub-set of CRM-EH entities. In terms of grey literature, the demonstrator supports queries of relationships between archaeological contexts, finds and materials.

The following example (Figure 7.27) presents the results of a query to retrieve documents containing a relationship between an archaeological context and an archaeological find. In particular, the query concerns the concept of "hearth" (sense of archaeological context EHE0007.Context) containing a "coin" in the sense of an archaeological find (EHE0009.ContextFind). The result of the *#archaeol8-6428* document corresponds to a "hearth" containing a "coin", as shown in the phrase "*It differs from the other coin finds, however, in that it was associated with a hearth uncovered*".

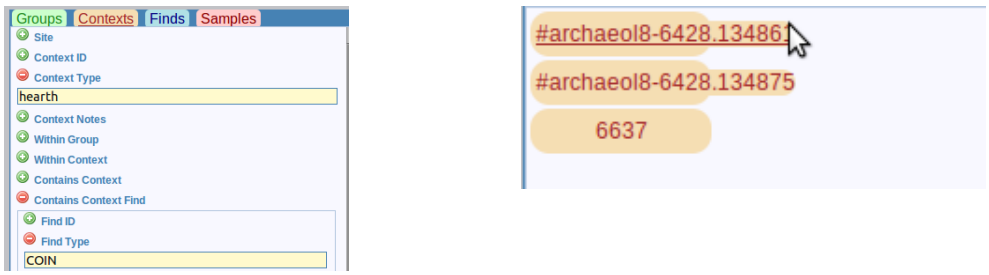


Figure 7.27: List of search results for the query; context (hearth) containing a find (coin)

The demonstrator also enables specification of the above relationship in an inverted expression where users query for an archaeological find that is found in an archaeological context, since relationships in CRM-EH as expressed by event entities are bidirectional. The following example (Figure 7.28) presents the results of the query "animal remain", in the sense of archaeological find (EHE0009.ContextFind), found in a "pit" (sense of archaeological context EHE0007.Context). The result of the *#cambridg1-24504_1* document corresponds to an "animal bone" that is found in a "pit", as shown in the phrase "*the test pit produced a range of artefactual material which included animal bone (medium/large ungulate) a fragment of which*".

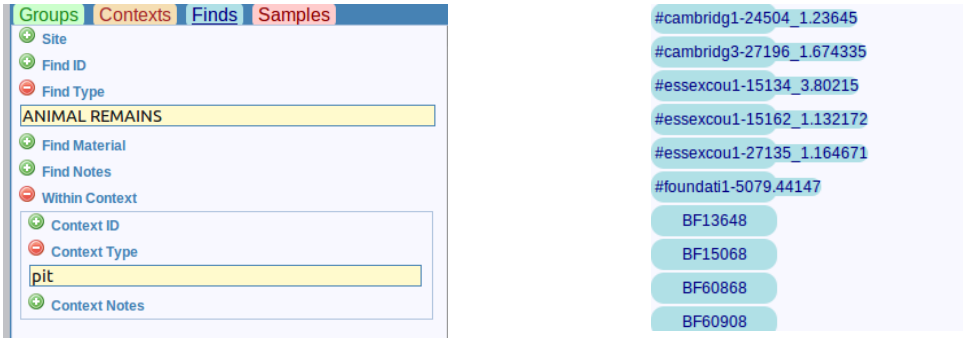


Figure 7.28: List of search results for the query; archaeological find (animal remains) found in archaeological context (pit)

Another form of relation query supported by the demonstrator is between archaeological finds and their materials. The following example (Figure 7.29) presents the results of the query "slag", in the sense of archaeological find (EHE0009.ContextFind), consisting of "iron", in the sense of archaeological find material (EHE0030.ContextFindMaterial). The result of the *#norfolka-6119* document corresponds to a "slag" consisting of "iron", as shown in the phrase "*a scatter of pottery sherds along with some iron slag which might indicate a metal working site*"

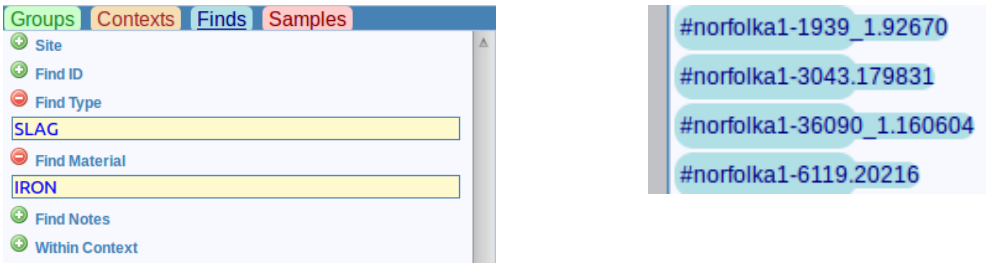


Figure 7.29: List of search results for the query; find (slag) consisting of material (iron)

7.4.2 False Positives

The search scenarios discussed so far have presented examples where the semantic indices delivered results which correctly corresponded to the query entities. However, there are cases where the indices deliver results that do not correspond to a correct ontological entity. This is caused by false semantic annotations (index entries) produced during the NER or the RE phases of the pipeline. Such annotations are known as *false positives* and concern cases where the automatic association of textual instances to the ontological classes is not correct. The IE mechanism invoked several NLP techniques to reduce false positives, such as Part of Speech tagging (POS), Noun-phrases, Verb-phrases, negation detection and disambiguation grammars. The next chapter discusses in detail the evaluation phase and strategy followed for benchmarking the performance of the OPTIMA pipeline. The following paragraphs of this section reveal the most common cases of false positives and discuss the basic factors which cause their delivery.

Resolving polysemous cases is a challenging task for any IE system. When disambiguation of polysemy relies on the input of Part of Speech taggers then inevitably any tagger mis-annotation is passed on to the disambiguation grammar. For example, the phrase "Well stratified pottery from Weobley has scarcely ever been found" delivered "Well" as a EHE0007.Context (#archenfi2-31380_1.135184). The term in this case should not have been annotated as a noun, however being at the beginning of the sentence and capitalised, it was annotated from the POS as noun and consequently delivered as a false positive. There are other cases where the POS is correct but the logic supporting the grammar fails to deliver correct results. For example the phrase "four of the skeletons had cut marks on the cervical vertebrae, giving rise to the interpretation" delivers a EHE0007.Context (#aocarcha1-40344_1.55964) when clearly here "cut" refers to the marks, rather than a "cut" into the earth

Similar to the above cases are the phrases "A brick built structure ([104]) was recorded in the central northern part" and "It comprised brick-like tiles. The water main crossed it east-west". The first case delivers "A brick" as a EHE0009.ContextFind (#albionar1-16331_1.56751) while the second annotates "brick-like" as a EHE0030. ContextFind (Material).

In both cases the logic that drives the JAPE grammars fails to deliver correct results. In the first case, the use of the "A" determiner is apparent but this evidence should not be enough to classify brick as an object, while in the second case the use of hyphen alters the meaning of the term. In this particular example it is not clear if the hyphenation is used to describe a tile, which has a brick-like shape, or a tile, which has a brick-like material. Based on the second reading the annotation is (partly) right. However the example reveals the true challenges in distinguishing the ontological sense between material from physical object and how hard it can be sometimes, even for humans, to distinguish the two senses. This point is discussed further in the evaluation.

Correct annotation of CRM-EH events is also a very challenging task. There are cases where the grammars may "stretch" the ontological meaning of the events, as for example in the phrase "*was a rubble layer of red brick and lime mortar fragments*" where "red brick" is annotated as a EHE0009.ContextFind deposited (#molas1-9721_1) in a EHE0007.Context "rubble layer". Clearly fragments of red brick and lime mortar were deposited in a layer but the deposition event is only implicitly stated in the sentence. Although, the deposition event is not mentioned in the phrase (only assumed implicitly) still the system delivers a deposition event annotation. In particular the brick and mortar are the layer i.e. aggregate from the layer. CRM-EH distinguishes discrete context finds from the physical staff of a context but this fine detail was considered not of importance for the STAR purposes.

In some other cases the grammars deliver false positive matches, which do not correspond to the ontological definition, as for example the phrase "The brick fabric of the oven was constructed upon layers of pre-fired material", which is incorrectly annotated as a deposition event (#stokeont2-33662_2) , since the phrase uses layers to refer to the construction technique not a context of an archaeological excavation.

7.5 Summary

The chapter has concluded the phase of semantic indices delivery. Two separate indexing files have been produced for each of the 2460 files that were processed by the OPTIMA pipeline. The files enabled retrieval and inspection of the documents by two separate web-based applications. The Andronikos web-portal utilised the coupled form of XML indices for producing HTML documents of semantic annotations aiding inspection and fact finding. The STAR demonstrator, on the other hand, utilised the (decoupled from content) RDF triples of semantic annotation for supporting information retrieval and cross searching with archaeological datasets.

The use of the semantic indices by third party web-based applications and their capability to support further use and manipulation supports the initial assumption of the capability of semantic annotation with respect to CRM and CRM-EH ontologies to support semantic-enabled information retrieval and document inspection. The chapter has contributed the process of transformation from GATE semantic annotations to interoperable formats (XML and RDF) with respect to the given ontologies and discussed their use by third party applications. In addition, the discussion revealed problematic areas and issues regarding delivery of *false positive* matches that distort the semantic accuracy of the indices. The following chapter presents the evaluation strategy, analysis and results of the OPTIMA pipeline performance and discusses further the issues which affect its accuracy as reflected by the standard metrics of *Recall* and *Precision*.

Chapter 8

Evaluation

8.1 Introduction

The chapter discusses the evaluation of the semantic annotation work. The IE performance of the OPTIMA pipeline is measured following an established evaluation methodology drawn from literature. Evaluation results are discussed in detail and conclusions regarding the system's performance and achievement are revealed. The discussion commences with an introduction to the evaluation aims and objectives. A brief literature review with regards to the evaluation methods of semantic annotation applications is presented before the discussion revealing the evaluation method. A detailed discussion of the evaluation results follows which presents the system performance against a range of test-bed configurations. The chapter finishes with conclusions and observations.

8.2 Evaluation Aims and Objectives

The evaluation aims to address the performance of the IE system (OPTIMA) in terms of its capacity and accuracy to provide semantic annotation with respect to the domain ontologies, CIDOC CRM and CRM-EH. The evaluation employs standard methods and procedures for benchmarking IE systems, which drive the evaluation strategy and support the analysis of the evaluation results.

In terms of system performance, the evaluation is focused on the capability of the system to support the IE tasks of Named Entity Recognition (NER) and CRM-EH Relation Extraction, with respect to the given domain ontologies. It also aims to evaluate the system's performance with respect to five different semantic expansion modes (Only Glossary, Synonym, Hyponym, Hypernym, All Available), in order to evaluate the capability of the contributing terminological resources to support the NER outcome of the system. The performance of the contributing IE modules of Negation Detection, Disambiguation and Noun Phrase Validation is also evaluated and results are discussed. However, NER is the primary focus of the evaluation and the contribution of the additional modules is evaluated via the main NER evaluation exercise.

The appropriateness and validity of the CRM-EH entity specialisation technique is also evaluated. As discussed in section 6.2.2.2, the OPTIMA pipeline provides the required CRM-EH specialisation based on the assumption that CRM entities which are identified by the NER phase can qualify as CRM-EH specialised entities only when found in a 'rich' phrase identified by the Relation Extraction phase. The above assumption is put to test by benchmarking the outcome of two different system configurations, which provide CRM-EH specialisation with and without using the 'rich' phrases setting. In addition, the contribution of the relation extraction patterns, as finalised after the corpus analysis study (section 6.3.3), is also evaluated. An early system configuration which did not use the syntactic patterns is invoked and results are compared and contrasted with the performance of the final system version.

The task also aims to deliver the test data for the evaluation, known as the "Gold Standard" (GS). The iterative process of the GS definition is discussed and the process of manual semantic annotation of selected passages by experts for the definition and annotation of the evaluation corpus is also revealed.

The pipeline, as discussed in section 4.7 makes use of some ANNIE modules such as Tokenizer, Sentence Splitter, Noun Phraser, etc. The performance of such modules is not evaluated, since their level of contribution and their configuration remains unchangeable, affecting in the same way all other evaluation configurations that are put on test. In addition, the contribution of the Pre-processing phase in terms of individual document sections is not evaluated. As discussed below in section 8.4.2.1, the gold standard corpus is based on the manual selection of summary sections of archaeological reports.

8.3 Evaluation for Semantic Annotation

The evaluation of IE systems, as discussed in section 2.3.1, was established by the Machine Understanding Conference, MUC 2. Two primary measurements were adopted by the conference, *Precision* and *Recall*, originating from the domain of Information Retrieval but adjusted for the task of IE (template filling). According to the MUC definition, when the answer key is N_{key} and the system delivers $N_{correct}$ responses correctly and $N_{incorrect}$ incorrectly then $Recall = \frac{N_{correct}}{N_{key}}$ and $Precision = \frac{N_{correct}}{N_{correct} + N_{incorrect}}$.

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

$$Recall = \frac{N_{correct} + (0.5 * Partial\_matches)}{N_{key}}, \quad Precision = \frac{N_{correct} + (0.5 * Partial\_matches)}{N_{correct} + N_{incorrect}}.$$

Partial matches are weighted as "half" matches. The value of the weight can change if partial matches seem more or less important. When partial matches are treated as correct matches then 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. The thesis discusses the evaluation results for *Average* and *Lenient* scalar modes.

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

The definition of the test data (Gold Standard) is a very critical stage in the evaluation process. An erroneous definition could distort the results of the evaluation and lead to false conclusions. It is a safe choice to use an available Gold Standard corpus, built by experts to support the evaluation of IE systems. However, the availability of such corpora is not always high. In the case of archaeological reports, there was no available Gold Standard of semantically annotated documents with respect to the CRM ontology. Therefore, the evaluation stage pursued the definition of a Gold Standard corpus tailored to the purposes of the evaluation task.

The definition is based on a set of annotation principles, which were employed by the SEKT project (Peters et al. 2003). SEKT followed the MUC and ACE criteria for NER and adopted a range of principles for semantically annotating a corpus of 292 news articles.

Such criteria concern:

- *Orthography* of annotations, where for example possessive ending and plural endings are treated as an integral part of the same entity.
- *Topicality* of annotations, where an entity is considered as valid only when relevant to the overall discourse of the text.
- *Phrasal Annotation,* to cover phrasal structures excluding determiners and quantifiers. The principle implies the use of embedded annotations which maintain the syntactic dependencies between entities.
- *Negated Entities* or non-existing entities are not annotated
- *Generic* annotations following the ACE guidelines are annotated. Generic is an entity which is not a particular, uniquely referent entity but a kind or type of entity. As opposed to SEKT, Generic annotations were included in the Gold Standard definition.

The above principles were adopted in the definition of the Gold Standard and were followed during the iterative process of defining the annotation instructions.

The final definition of the Gold Standard may be set by one or more experts. In the range of MUC and ACE conferences, test data were prepared by a committee. However, it is not always certain that experts will agree on semantic annotation, especially when a semantic annotation is capable of carrying fine ontological distinctions. In the case of adaptive Machine Learning (ML) Information Extraction, manual semantic annotation is not only used for defining the Gold Standard but also for delivering the ML training set. The training data is used to train the ML algorithm, thus it is critically important such data be adequate and non-erroneous.

ML practice has addressed the issue of problematic manual annotations by enabling multiple annotations per document. The technique allows more than one person to annotate the same text in order to address discrepancies between different annotators. To calculate the agreement level between annotators the technique employs the Inter Annotator Agreement (IAA) metric (Maynard, Peters and Li 2006). The metric shows the level of agreement between different manual annotation sets, either for particular entities or overall. The best IAA score is 1 or 100% which shows a total agreement between different annotators and the worst is 0 where there is absolutely no agreement. A "Super Annotator" can act as a referee to conciliate differences of manual annotations (Savary, Waszczuk, and Przepiórkowski 2010). However, a low IAA score may sometimes be caused by insufficient or unclear manual annotation guidelines. When improvement of the guidelines

does not bring any significant improvement in the overall agreement between individual annotators, then most probably the task is very challenging due to inherited ambiguities.

Zhang, Chapman and Ciravegna (2010) agree with Byrne (2007) that manual document annotation in archaeology is a challenging task. Domain specific issues such as complexity of language, uncertainties, composite terms, acronyms and so on describe a challenging task, where the overall IAA score is normally around 60%. Due to differences in knowledge and experience, manual annotators may produce discrepant annotations. A pilot evaluation is critical for revealing such discrepancies. Liakata et al. (2006) and Zhang, Chapman and Ciravegna (2010) highlight the importance of pilot evaluation for refining manual annotation guidelines and for identifying problematic annotation at early stages, before committing to large scale manual annotation.

Zhang, Chapman and Ciravegna (2010) produced a training set of manual annotations of archaeological documents via an iterative process of three main phases. The first phase aimed to identify as many discrepancies as possible at low cost, by annotating a small corpus of 5 documents (5 to 30 pages long) using two different annotators. During the second phase, 5 annotators used refined and enriched guidelines, for annotating 25 documents. The pilot corpus produced during this phase was used to evaluate the capacity of the manual annotations for machine learning, i.e. how well a machine could learn from such annotations. In addition, during the second phase individual discrepancies and inconsistencies of annotators were revealed regarding annotation of different entity types. The third phase delivered the final version of manual annotations. Based on the observations of the second phase, the third phase allocated a selection of entities per annotator. Unlike the other two phases, the third phase did not produce duplicate annotations per document. Instead, each annotator was assigned to a particular set of entities for which she/he had the most consistent performance. This approach claims to deliver a cost-effective and less tedious manual annotation technique.

The SAPIENT tool (Liakata and Soldatova 2009) was employed by the ART project (Soldatova et al. 2009) to deliver the task of sentence-based annotation of scientific papers. The tool enables experts to manually annotate full papers with respect to 11 scientific concepts, such as Background, Conclusion, Experiment, Hypothesis, etc. Users of the tool can also link related sentences together to form spans of interesting regions, which can be utilised further by text mining applications. The process of semantic annotation used 16 chemistry experts for annotating 42 journal papers. A pilot annotation phase split annotators in five groups of three, assigning to each group eight different papers, plus two

common papers across groups. The common documents were used to identify the most deviant annotators by revealing major discrepancies between manual annotation sets. The final (third) phase of the annotation process employed eight annotators which showed higher IAA scores.

The cost of the manual annotation of 25 journal papers was £1000, where each paper annotation cost £40 and took approximately 2 hours to complete. Zhang, Chapman and Ciravegna (2010) on the other hand, estimated that the annotation of 25 documents, containing overall 471001 words, took between 10 to 15 person-days to complete. Although, they do not give an exact figure on the actual cost, their estimation appears comparable to the cost of the ART project based on average pay of skilled personnel.

The Gold Standard (GS) as a standard method for evaluating semantic annotation requires the definition of a final version of manually annotated documents. The evaluation task discussed in this chapter adopted an iterative process for deriving the final version of GS, involving a pilot evaluation and IAA analysis which led to the final evaluation. This drew on the techniques and principles discussed above but central to the process was the end-user relevance of annotations. The GS aimed to represent the desirable result of semantic annotation of archaeological documents with respect to the end-users of such documents. Therefore, the definition task was assigned to archaeological experts, with various levels of specialisation, following standard methods of annotation. The annotators focused on the value of such annotation for study and research rather than annotating for a ML training set.

8.4 Evaluation Methodology

The evaluation method is based on an iterative process of Gold Standard definition via a pilot evaluation. Upon successful definition of the final Gold Standard version, the IE system is prepared and the evaluation experiment is executed. Evaluation data are then collected and analysed and findings are discussed.

The contribution of the Inter-Annotator analysis is significant, both during pilot and main evaluation phases, for revealing any major discrepancies and deviant annotation results. The methodology adopts an "end-user" focus, where annotators are expected to exercise judgement as competent users. The instructions for evaluators are intended to be relevant to future cross search and hence neither the scope of the ontology elements nor precise vocabulary was 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. In addition, the role of the Super Annotator, as a normalising factor for conciliating annotators' disagreement, is critical for deriving a single and final definition of the Gold Standard.

8.4.1 Pilot Evaluation

The aim of the pilot evaluation was to examine the effectiveness and comprehensiveness of the manual annotation instructions for supporting the annotation task, as well as to explore the capacity of GATE modules for supporting the evaluation and the IAA analysis tasks. An initial evaluation which was conducted during the prototype study (chapter 3), examined the functionality of GATE for the provision of evaluation results in terms of Precision, Recall and F-measure metrics. To avoid confusion, the evaluation of the earlier prototype study will be referred to as e*arly evaluation*, while the term *pilot evaluation* is used to describe the first stage of the final evaluation phase and the term *main evaluation* to describe the second (main) phase of the final evaluation. The early evaluation helped to gain experience with the manual annotation (OAT) and evaluation tools (Corpus Benchmark Tool and Annotation Diff) of GATE. The pilot evaluation introduced the IAA analysis in order to reveal differences between annotators and to refine the instructions of manual annotation before proceeding to the main evaluation phase.

The pilot evaluation used the same small corpus of 10 summary extracts which was used by the early evaluation phase. The manual annotation task relied on volunteers since there was no budget available for paying annotators. Considering the logistic constraints of the manual annotation task and the end-user focus of the evaluation task, summary extracts, originating from archaeological excavation and evaluation reports were understood to be the best available option. Such passages contain rich discussion which highlights the major findings of archaeological excavation and evaluation reports while their size is not large, allowing annotators to complete the annotation task within hours rather than days.

The manual annotation instructions were written to reflect the end-user aims of the evaluation (supporting retrieval and research of archaeological reports), hiding complex and sometimes unnecessary ontological details. Annotators were instructed to annotate at the level of archaeological concepts rather than identifying more abstract ontological entities in context. The instructions in effect directed annotators to adopt the principles of orthography, topicality, phrasal annotation and negation detection as discussed above. In detail, the instruction directed the task of manual annotation at the concepts of

archaeological place, archaeological find, material of archaeological finds and time appellation, thus annotating textual instances that have a value from an archaeological point of view. In addition, the instruction asked for annotation of 'rich' phrases which involved two or more of the targeted concepts. Such 'rich' phrases were used in the evaluation of the CRM-EH relation extraction phase. However, as with the single concept annotation, the ontological details of such events were not included in the instructions, in order to avoid over complicating the annotation task and diverting from the end-user focus. Neither vocabulary was specified, apart from a few examples included in the instructions, annotators had to identify relevant strings and assign ontological entities. The instructions finalised after the pilot evaluation refinements can be found in [Appendix D3].

In total, three annotators were employed by the pilot evaluation task, a senior archaeologist, a commercial archaeologist and a research student of archaeology. The evaluation corpus of the ten summary extracts (5 excavation and 5 evaluation reports containing in total 2898 words) was made available in MS Word format. The number of participants and the volume of the evaluation corpus were considered adequate to support the aims of a small scale pilot evaluation study targeted at informing and revealing early problems.

The annotators were instructed to use particular highlight colours and underline tools in order to produce their annotations. Although the annotation task could have applied the GATE OAT tool, the use of MS Word was preferred, since manual annotators had no previous experience working with GATE but were fluent in Word. Exposing annotators to GATE would have required additional training and since the task of annotation was undertaken by volunteers, it was decided to use their time effectively and focus on the annotation task itself.

The resulting manual annotations sets were transferred to GATE as CRM and CRM-EH oriented annotations by the *annotation-editor* using the OAT tool. The role of the annotation-editor was taken by the system developer (Vlachidis), who acted as an intermediary between Word and GATE annotations, transferring the results of manual annotation into GATE without interfering with the annotation outcome. Since annotators were asked to select 'rich' phrases, some minor normalisation was required during transfer of annotation from Word to GATE. Such alternations did not distort or alienate the transferred from original annotations but aimed to normalise the manual input. For example, some annotators used a single highlighted span to cover the range of comma separated terms, while others used as many spans as the individual conjunct entities. This

was normalised by the editor, so their conjunct terms produced individual annotations. The full list of normalisation factors and principles followed by the annotation-editor can be found in [Appendix D4]

Upon completion of annotation transfer, the differences between the individual manual annotation sets were analysed using the IAA module of GATE. The module was configured to report the inter-annotator agreement score in terms of Precision, Recall and F-measure metrics. The metrics were reported on both *Average* and *Lenient* mode. The *Average* mode treats partial matches as half matches as shown by the Precision and Recall formulas (section 8.3). On the other hand, the *Lenient* mode treats partial matches as full matches figuring Precision and Recall as below.

$$Recall = \frac{N_{correct} + Partial\_matches}{N_{key}}, \; Precision = \frac{N_{correct} + Partial\_matches}{N_{correct} + N_{incorrect}}$$

The overall IAA F-measure score for the three annotation sets with the *Average* mode of reporting was 58%, while with the *Lenient* mode the score increased to 68%. The results agree with Byrne (2007) and Zhang, Chapman and Ciravegna (2010) for low IAA score in manual annotation of archaeological documents, due to domain language characteristics but also affected by the borders of annotations. In the *Lenient* mode of reporting, individual differences in terms of annotation borders are not encountered since partial matches are considered full matches. An increment of 10% from *Average* to *Lenient* mode is evidence of the disagreement between annotators on annotation boundaries, such as disagreement on including adjectives or other moderators e.g. annotate 'large pit' or just 'pit'. The following tables (8.1, 8.2) present the IAA annotation score for the three different annotation sets in terms of overall agreement and agreement on individual entities.

	Precision		**Recall**		**F-measure**	
	Average	*Lenient*	*Average*	*Lenient*	*Average*	*Lenient*
SA-CA-RS	0.52	0.62	0.64	0.76	0.58	0.68
SA-CA	0.60	0.60	0.50	0.50	0.54	0.54
SA-RS	0.62	0.75	0.83	1.00	0.71	0.86
CA-RS	0.38	0.50	0.60	0.80	0.46	0.61

Table 8.1: IAA scores for the three different annotation sets, reported on Average and Lenient mode. SA: Senior Archaeologist, CA: Commercial Archaeologist, RS: Research Student

	Precision		**Recall**		**F-measure**	
	Average	*Lenient*	*Average*	*Lenient*	*Average*	*Lenient*
E19.Physical_Object	0.67	0.89	0.57	0.76	0.61	0.82
E49.TimeAppellation	0.82	0.88	0.77	0.81	0.79	0.84
E53.Place	0.70	0.84	0.68	0.80	0.69	0.82
E57.Material	0.33	0.36	0.65	0.71	0.43	0.47
EHE1001.ContextEvent	0.76	0.82	0.55	0.60	0.63	0.68
EHE1002.ContextFind ProductionEvent	0.48	0.84	0.32	0.56	0.37	0.64
EHE1004.ContextFind DepositionEvent	0.47	0.93	0.18	0.34	0.23	0.45
P45.consists_of	0.53	0.62	0.64	0.77	0.57	0.60

Table 8.2: IAA scores of individual entities reported on Average and Lenient mode. The scores are reported on the IAA score of the three Annotation Sets SA-CA-RS

The data of table 8.2 display a low IAA behaviour, especially when results are reported on the *Average* mode. The *Lenient* mode provides improved scores as expected, since disagreement on the annotation boundaries is not taken into account. The best performing F-measure is between Senior Archaeologist and Research Student, scoring 71% on the *Average* mode and 86% on the *Lenient* mode respectively.

Examining table 8.2 it is made clear that the agreement on *E57.Material* entity is significantly low both on *Average* and on *Lenient* mode of reporting. Also there is a 20% difference between *Average* and *Lenient* mode on the F1 score of the *E19.Physical_Object* entity, indicative of the different annotation boundaries used by annotators. Similarly, the event-based annotations EHE1002 and EHE1004 show a significant difference, around 25%, between *Average* and *Lenient* mode, which is expected since identification of 'rich' phrases can vary in annotation boundary. The IAA score for the event-based entities is rather low, reflecting the embedded ambiguities and challenges of the annotation task.

Based on the results of the pilot evaluation, the manual instruction were refined [Appendix D3] to improve the manual annotation task. In detail, the instructions for the annotation of material entities were re-written to make clear that the annotation task should be focused on materials that have an archaeological interest and are associated with physical objects. Similarly, the instructions for the annotation of place entities were re-written to clarify annotation of both primitive and larger groupings of contexts. Instructions were also refined to highlight that entity annotations spans should contain both

conjunct and adjectival moderators when applicable. In addition, annotation examples were included for each annotation type targeted by the task in order to ease comprehension. Emphasis was given to the instructions for the annotation of the 'rich' phrases, where instruction attempted to clarify the cases of phrasal annotation and how boundaries should be treated. In addition, an introductory paragraph was added to prologue the aims of the annotation task and a set of full examples, showing use of colour coding and 'rich' phrases underlining, was added to help annotators visualise the annotation outcome.

The pilot evaluation revealed a range of terms which were selected by annotators but were not matched by the IE pipeline. Such terms were either not included in the terminological resources or were not matched by the rules. Any non-included terms were by definition not in the EH glossaries. It was decided not to include them in the gazetteer listings in an ad-hoc manner, as the authority of the terminology resources would be undermined (the list of non-included terms can be made available to the EH terminology team for further consideration).

On the other hand, in the case of terms which were included in the thesauri but not in the glossaries, it was possible to improve the system. These terms were not matched by the rules because they were not in the glossaries. The semantic expansion to thesauri resources needs a glossary term to start from. Where a selected by the annotator term exists in the gazetteer but the term or its synonyms/narrower/broader concepts were not included in one of the glossaries, it remained 'invisible' to the semantic expansion matching rules. Such terms were added to the rules of the Hyponym and Hypernym expansion modes since they were available in Thesauri structures though not in the Glossaries. The list of the non-existing and 'extra' terms of rules can be found in [Appendix D5].

8.4.2 Main Evaluation

The main evaluation task was informed by the results and observations of the pilot evaluation stage for delivering a full scale evaluation of the OPTIMA IE system. Taking into account logistic and resource constrains, the task managed to deliver the evaluation aims and objectives with regards to the system's completeness and correctness to provide CRM/CRM-EH oriented semantic annotation. The main evaluation task was divided into four sub-tasks conducted in the following order; 1) Selection of the evaluation corpus 2) Design and execution of manual annotation 3) IAA analysis and 4) Deriving the Gold Standard (Audit by Super Annotator).

The main manual annotation task was based on volunteer labour of archaeologists, considered as representative of the archaeological community generally, who volunteered a couple of hours. The final definition of the Gold Standard was prepared by the annotation editor and audited by the Super Annotator who also volunteered.

8.4.2.1 Selection of the Evaluation Corpus

The selection of the evaluation corpus was influenced by three major criteria. The selection a) had to sufficiently support the evaluation aims and objectives, b) be representative of the OASIS corpus and c) have an appropriate size that could support manual annotation with respect to the available resources. Thus, the selection focused on summary extracts of archaeological excavation and evaluation reports, originating from a range of different commercial archaeology units (to take account of possible differences between units). Compared to other types of OASIS reports, such as watching briefs and observation reports, the evaluation and excavation report types typically contain more information on the findings of archaeological excavations.

The summary extracts present some significant advantages over other document sections. Summaries are brief containing rich discussion which reflects the main document findings. Hence, such sections can support the end-user focus of the evaluation due to their density and richness. On the other hand, summary sections are less labour intensive than the complete reports and are relatively easy to isolate and to extract.

The summaries of the main evaluation corpus were manually extracted based on a random selection of documents (via the RAND() function in Open Office spreadsheet application). In detail, two lists containing the unique document names were defined, each list containing the names of Evaluation and Excavation reports respectively. A random number was assigned to each document name and used to sort the lists in descending order. The selection of documents was based on the criteria of highest number, origin of document (archaeological unit) and size of summary passage. Documents with the highest number which belonged to an archaeological unit not included in the selection yet and having a summary length between 100-300 words were prioritised.

The number of selected documents was influenced by the available resources. Since the manual annotation task was undertaken by volunteers, it was necessary to define a task that could be completed within a couple of hours. Based on the simple estimation that a professional annotator needs on average 50 seconds to annotate a sentence with average 17.5 tokens (Brants 2000), the annotation of 3000 tokens would take approximately 140 minutes. Therefore, each manual annotator was assigned a document to annotate

containing 2500-3000 words maximum. This translates to each composite document containing 10 summaries of 250-300 words.

Overall, 64 summary extracts (32 Excavation and 32 Evaluation) were randomly selected from the two lists. The selected extracts were combined into 7 different documents, where each document contained overall 10 summary extracts; 5 Evaluation and 5 Excavation report summaries. One particular summary extract was included in all seven documents i.e. (7*9)+1=64 summaries altogether. This particular extract acted as a normalisation criterion and was used at a later stage of the analysis for revealing any major discrepancies between annotators. The 64 extracts originated from 19 different commercial archaeological units and were considered to be representative of the OASIS corpus and adequate for supporting the aims of the evaluation task, given the available resource constraints.

8.4.2.2 Conducting the Manual Annotation

The manual annotation task was conducted at the Archaeology Data Service (ADS, York University), with the participation of 12 archaeology practitioners, including ADS staff and post-graduate students. Due to the even number of participants it was decided that each document be annotated by two different annotators. This choice provided good coverage of the evaluation corpus while enabling comparison of the annotation results in pairs, as required by the IAA analysis stage (section 8.4.2.3). Overall, six composite documents containing 55 individual summary passages were annotated by six groups, where each group consisted of two annotators.

Initially, a brief introduction on the purposes of the task and a demonstration annotation of a sample summary was given. The instructions of the task were made available and annotators got the chance to raise questions before engaging with the manual annotation task. It was made clear that task had a user-centred focus asking for their viewpoint as experts. Thus, their annotation should be meaningful from an archaeological research point of view. Each annotator agreed to contribute and was then assigned a particular document. The documents were distributed among annotators ensuring that those sitting next to each other were assigned different documents to avoid collaboration.

The annotators followed the instructions [Appendix D3] and used the MS Word application to deliver annotation by highlighting textual instances using specific colours and underlining rich phrases. The completed manual annotation documents were copied to a USB drive and made available to the annotation editor. The task lasted approximately 2 hours with the first annotator completing the task in 45 minutes and the last 1 hour and 50'.

8.4.2.3 Inter-Annotator Agreement Analysis

The manually annotated documents were made available to the annotation editor (Vlachidis), who transferred the annotations from MS Word files to GATE using the OAT tool. Similar to the pilot study practice, the annotations were transferred "as is", following the same transfer principles [Appendix A4] without altering or amending the annotators input. Again annotators were not supplied with ontology details but annotated specific archaeological concepts and 'rich' phrases connecting them. The annotation editor translated these rich phrases to CRM-EH events, similarly to the pilot study practice (section 3.4). The manual annotations were transferred to GATE on a document basis, with every single document carrying two different annotation sets, each corresponding to an individual annotator's input. The IAA scores (Table 8.3) were calculated using the IAA module of GATE, configured to compute results on F-measure which is a suitable metric for named entity annotations. (The module is also configurable to compute results on Cohen's Kappa which is suitable for text classification tasks).

	Precision		**Recall**		**F-measure**	
	Average	*Lenient*	*Average*	*Lenient*	*Average*	*Lenient*
GROUP 1	0.67	0.80	0.40	0.47	0.50	0.60
GROUP 2	0.53	0.65	0.56	0.68	0.55	0.66
GROUP 3	0.51	0.62	0.56	0.68	0.53	0.65
GROUP 4	0.65	0.82	0.52	0.66	0.58	0.73
GROUP 5	0.65	0.77	0.59	0.70	0.62	0.74
GROUP 6	0.64	0.84	0.40	0.53	0.50	0.65

Table 8.3: IAA scores of the 6 groups participating in the main evaluation phase

The results delivered a moderate agreement score, comparable with the pilot IAA agreement score, indicative of the embedded language ambiguities and challenges in manual annotation of archaeological documents (Byrne 2007 ; Zhang, Chapman and Ciravegna 2010). Refinement of the instructions did not deliver a great improvement in terms of F-measure scores which ranged from 50% to 74%. An improve of the precision rates is evident where the lowest agreement of the pilot evaluation is 50% and the lowest of the main evaluation 62%, both reported on *Lenient* mode. Similarly the highest precision agreement of the pilot evaluation is 62% where the highest of the main evaluation is 84%. However, agreement on recall rates of the main evaluation is not significantly improved compared to the pilot evaluation. The inherited domain ambiguities with regards to vocabulary and language ambiguities as discussed in section 5.6 present

major challenges in reaching a high IAA score.

Agreement between the two annotators in the different pairs, ranged between 50% to 62%, with the highest score delivered by Group 5 (62%) and the lowest by Groups 1 and 6 (50%) which is typical of the domain. The scores are higher when reported in *Lenient* mode, reaching 74% agreement for Group 5. For some groups such as, group 4 and 6 the agreement score improves by about 15% from *Average* to *Lenient* mode, while for the rest the increment is around 10%. The *Lenient* mode of reporting delivers higher scores because it does not consider disagreement on annotation boundaries, which normally cause discrepancy between annotators (Zhang, Chapman and Ciravegna 2010).

The next stage in the IAA analysis focused on revealing the most deviant annotation inputs. The method used a common, across all groups, summary extract, which acted as a benchmarking criterion. The common summary extract was embedded in all (6) documents participating in the evaluation, thus annotated by the 12 annotators. The IAA score in terms of F-measure of the 12 annotation inputs was 0.56 (i.e. 56%). The process of computing the IAA score was repeated 12 times removing every time a different annotator and calculating the score of the remaining 11 inputs. This technique enabled identification of the level of individual annotator discrepancy. The higher the level of discrepancy of an individual annotator, the higher the agreement score of the rest of the annotators when that individual annotator input was removed from the computation. The following table 8.4 presents the levels of individual discrepancy for the 12 manual annotation inputs.

	F-measure	**Discrepancy**
Group 1 – Annotator A	0.54	0.02
Group 1 – Annotator B	0.58	0.04
Group 2 – Annotator A	0.56	0.00
Group 2 – Annotator B	0.56	0.00
Group 3 – Annotator A	0.57	0.01
Group 3 – Annotator B	0.57	0.01
Group 4 – Annotator A	0.56	0.00
Group 4 – Annotator B	0.56	0.00
Group 5 – Annotator A	0.57	0.01
Group 5 – Annotator B	0.57	0.01
Group 6 – Annotator A	0.56	0.00
Group 6 – Annotator B	0.57	0.01

Table 8.4: Level of discrepancy for each individual annotator

The F-measure value corresponds to the IAA score produced by removing each time an individual annotation input and calculating the agreement score for the remaining 11 annotation inputs. The Discrepancy value is calculated by subtracting the F-measure score from 0.56 which is the overall IAA score (F-measure) of the 12 annotation inputs

Based on the above table data, the level of discrepancy for all 12 individual inputs is low ranging from 0 to 4%. The most deviant annotation input is Group 1 – Annotator B which influences the overall agreement by 4%. Although, the overall IAA score for the 12 annotators is not high, this is not caused by one or two individual annotators that are too deviant. All inputs deliver slightly individualised annotations which is indicative of the challenge in manual annotation of archaeological documents. Therefore, all manual annotation inputs are included in the definition of the final Gold Standard version which is discussed in the following section.

8.4.2.4 Deriving the Gold Standard

The determination of a definite and final Gold Standard version is critical for the delivery of summative evaluation results. The gold standard describes the desirable system performance in terms of the annotation result. Therefore, an explicit and unambiguous Gold Standard can be used as a benchmark tool for supporting delivery of conclusive results, which can be used to describe the overall system's achievement.

As discussed above, the IAA score of the manual annotation inputs, as is typical of the domain, is not regarded as high. Therefore, the definition of the final Gold Standard version cannot be derived just by adopting the given manual annotations. Roughly, both manual inputs are half right since annotators agree around 60%. It is necessary for a form of reconciliation to take place, in order to conclude on the final Gold Standard version, which should be drawn from both manual annotation sets. This is done by employing a Super Annotator who acts as a referee between individual annotation sets, reviewing the cases of disagreement and choosing the correct annotation (Savary, Waszczuk and Przepiórkowski 2010). Normally the Super Annotator is a field expert with the experience and knowledge to reconcile individual annotation discrepancies. For the Main Evaluation, the role of the Super Annotator was undertaken by the Senior Archaeologist, who also contributed to the Pilot Evaluation.

The Senior Archaeologist was involved in the development of the CRM-EH ontology and also had been exposed to manual annotation during the pilot evaluation. Thus, he had experience in dealing with manual annotation and good knowledge of the conceptual arrangements of the CIDOC-CRM and CRM-EH ontologies. He was able to dedicate a full

business day to the task. Due to the volume of the evaluation corpus and the amount of discrepancies, it was necessary to prepare a set of Gold Standard proposals for Super Annotator auditing, in order to utilise sufficiently the availability of the Senior Archaeologist.

The preparation of the Gold Standard proposals was conducted by the Annotation Editor. The Editor reviewed the two different manual annotations per document and proposed an integrated version for the Gold Standard by combining the two annotation sets in a complementary fashion. In detail, the Editor first included all common annotations of the two annotation sets which both annotators had agreed. Then he examined the annotations made by a single annotator, including the vast majority but excluding what appeared as awkward and controversial annotations. The excluded annotations were noted and made available to the Super Annotator for particular attention during the auditing process. Also the editor proposed the inclusion of a few annotations that were not identified by annotators but seemed to be valid annotations for consideration by the Gold Standard. Those additional annotations were also noted clearly and shown for particular attention by the Super Annotator.

A graphical representation of the Gold Standard preparation approach is depicted by figure 8.1. The dots represent the volume of annotations per document. The square box contains the annotations delivered by the two annotators and outside the box are the annotations that were missed by annotators. The oval shape contains the common annotations of both sets, around 60% of the box annotations. The irregular shape contains the annotations which are proposed by the Editor for inclusion by the Gold Standard. This contains all common annotations, a large number of non-common annotations and a few non-annotator annotations which are proposed by the editor.

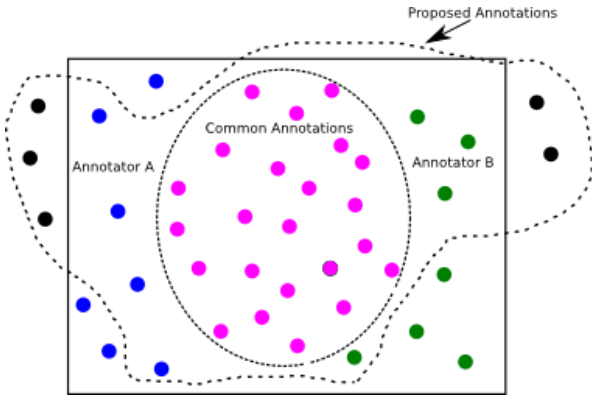


Figure 8.1: A graphical representation of the proposed annotation inclusion to Gold Standard

The proposed Gold Standard was audited by the Super Annotator during an approval session which lasted a business day (including follow up by Super Annotator). Four persons participated in the session; the Super Annotator, the Annotation Editor and two members of the Hypermedia Research Unit (University of Glamorgan). The proposed annotations were projected on a screen and reviewed on a summary per summary basis.

The Editor presented the two manual annotation sets in the GATE environment and the proposed Gold Standard set of each summary extract, highlighting major cases of discrepancy or missed annotations. The Super Annotator checked the proposed set and approved or rejected changes, deciding upon the final definition of the Gold Standard. The members of Hypermedia Research Unit contributed to the discussions during approval by double checking the final results and that all annotation cases were covered.

8.4.2.5 Encountered Problems

The process revealed some important issues regarding the definition of the final Gold Standard version, which might have been identified earlier if the Pilot evaluation (section 3.4) had attempted to deliver its final Gold Standard version. With hindsight, it is important that a pilot evaluation phase attempt completion of all stages which will be later employed by a full scale evaluation, in order to reveal any outstanding issues at an earlier evaluation stage. The revealed issues concern a) the assumptions driving the CRM-EH specialisation of annotations, b) the use of moderators in annotation spans and c) the involvement of frequently occurring annotations that are beyond the scope of CRM-EH entities.

With regards to the CRM-EH specialisation of annotations, the Annotation Editor, based on the pipeline configuration, proposed that singleton entities, i.e. entities that do not participate in rich phrases (CRM-EH Events) shall be annotated only as CRM entities. According to the Editor, only those CRM annotations which participate in 'rich' phrases containing at least two CRM entities should qualify as CRM-EH annotations, from the rationale that the specialisation be supported by the contextual evidence of a surrounding phrase. The Super Annotator (SA) challenged the above approach and suggested that all CRM entities, including singletons, can qualify as CRM-EH entities, since they originate from archaeological documents in any case. The SA suggestion was followed and it was decided that the final Gold Standard should annotate singletons as CRM-EH entities, in order to test the rich phrase qualification assumption.

The inclusion of moderators in annotation spans delivered ambiguities and inconsistencies in the annotation. Annotators were instructed to identify concepts including

their modifiers, supporting in this way the user-centred focus of the annotation process. However, this choice can undermine the performance of the IE system because some system annotations can be detected as only partial matches, purely due to the extended span of a modifier phrase. For example, in case of the manual annotation "*T-shaped small oven*" the system annotates only "*small oven*", because it is built to annotate only the immediate modifier not a whole noun phrase. Similarly, the same phrase can be treated inconsistently between individual annotators, who may disagree on the modifier span. In some other cases, an annotator might not include a moderator that is delivered by the system because the moderator is not important from an archaeological research point of view. For the same case, another annotator might include the moderator, assuming that it has some interest. Agreement on the annotation span is a known problem and use of moderators makes the annotation task even more challenging.

The Super Annotator suggested that the Gold Standard should contain all moderators delivered by annotators without modification. It was agreed to include all moderators and to report the evaluation results both in *Average* and *Lenient* mode, in order to obtain a complete picture of the system performance. The *Lenient* mode computes partial matches as full matches and thus discrepancies on the moderator level are not taken into account by the mode.

The Super Annotator also suggested that the Gold Standard should avoid including frequently occurring concepts that are beyond the scope of CRM-EH entities. A number of different concepts such as *Site*, *Trench*, *Feature*, and *Assemblage* are frequently mentioned in archaeological documents. However, such concepts do not correspond to the scope of the targeted CRM-EH concept EHE0007.Context.

Site is a very specific concept which is modelled by CRM and CRM-EH ontologies differently from the *Place* concept and is not targeted by the OPTIMA system. *Trench* on the other hand, usually refers to areas which supported the excavation work rather than places of archaeological interest that can be modelled as archaeological context. In addition, *Feature* and *Assemblage* are very broad concepts that do not describe particular places or objects but instead might be modelled as properties of certain entities. Thus it was agreed to exclude the above four concepts from the Gold Standard, although some annotators included them in their annotation. A full list of the CRM-EH entities participating in the evaluation exercise can be found in section 6.2.2.2

8.4.2.6 Phases of the Evaluation

The evaluation process, prepared a range of different system configurations which were tested against the Gold Standard. The configurations aimed to test different aspects and achievements of the system in terms of a) terminological resource exploitation via semantic expansion; b) Named Entity Recognition (NER); c) CRM-EH Relation Extraction and Entity specialisation; d) contribution of individual NLP modules, such as Noun Phrase Validation, Word Sense Disambiguation and Negation Detection.

The main evaluation task was conducted in three phases; the first phase addressed the NER performance of the system, the second phase the CRM-EH Relation Extraction and Entity specialisation performance and the third phase evaluated the contribution of individual NLP modules. A fourth phase, not connected with system's configuration sets, compared the NER results of the evaluation corpus with the available OASIS metadata. For conducting the three main phases of the evaluation task, the Gold Standard was expressed in two versions. The first version contained the four main CRM entities (*E19.Physical Object*, E*49.Time Appellation*, *E53.Place* and *E57.Material*) and was used to benchmark the NER performance of the system. The second version expressed the CRM entities, apart from Time Appellation, as equivalent CRM-EH entities (*EHE0007.Context*, *EHE0009.Context Find* and *EHE0030 Context Find Material*) following the Super Annotator suggestion to express all entities, including singletons, in CRM-EH.

The Time Appellation entities were not expressed in CRM-EH since manual annotators were instructed to identify textual instances of time appellation in the broad CRM sense. The CRM-EH specialisations are very particular to the CRM-EH events. For example the entity *EHE0039.Context Find Production Event Timespan Appellation* is a specific specialisation of time appellation that corresponds to the time of a production event. This form of CRM-EH specialisation does not meet the user-centred focus of the evaluation exercise and hence was not included in the second version of the Gold Standard.

8.5 Evaluation Results

The discussion of the results is arranged according to the three phases of the evaluation process and results are reported both in *Average* and *Lenient* mode. The first part discusses the results of the NER phase, the second part discusses the results of the CRM-EH Relation Extraction and Entity specialisation phase, the third phase discusses the contribution of individual NLP modules while there is a fourth part that reveals the comparison between the OASIS author-based metadata and semantic annotations. The discussion concludes with the summative results and lessons learned while conducting the evaluation process.

8.5.1 NER Evaluation Results

The first phase (Figure 8.2) used the CRM version of the Gold Standard to benchmark the NER performance of the system for five different system configurations. The configurations correspond to the five different modes of the semantic expansion over terminological resources namely; *only-glossary*, *synonym*, *hyponym*, *hypernym* and *all-available*. The *Hypernym* system configuration, which delivered the best performance in terms of F-measure score, is used in the second phase of the evaluation which evaluates the CRM-EH Relation Extraction (RE) performance. Although, all five semantic expansion configurations could have been used by the second phase of the evaluation, this would have generated an overwhelming volume of results without adding a significant value to the evaluation.

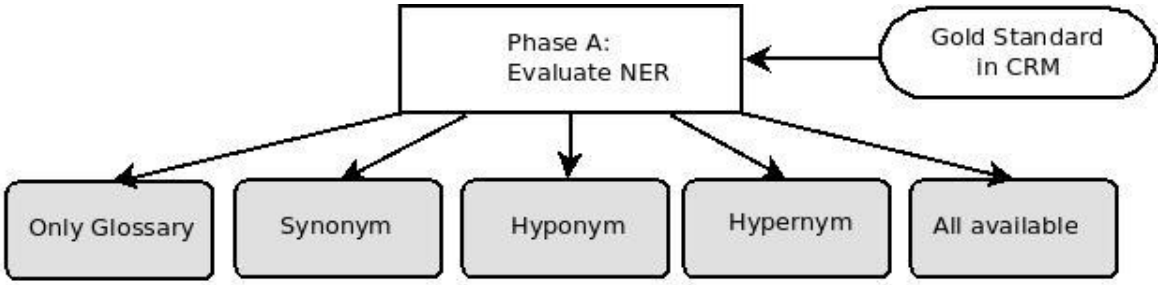


Figure 8.2: Main Evaluation Phase A: Five system configuration modes

The phase was targeted at the NER result of the four CRM entities *E19.Physical Object*, *E49.Time Appellation*, *E53.Place* and *E57.Material*. It executed five different system configurations, which correspond to the five modes of semantic expansion used by the NER pipeline. The results of the five different configurations in terms of Recall, Precision and F-measure are shown below (Table 8.5). The table presents the evaluation metrics in both *Average* and *Lenient* mode of reporting, where *Average* results are around 6% lower than *Lenient.*

The difference between the two modes of reporting is affected by annotation boundaries, mainly because the Gold Standard includes entity moderators. The use of moderators by manual annotators, as discussed above, can be subjective and inconsistent; the system is programmed to include the most immediate moderator which is not always how annotators treat annotation boundaries.

Based on the *Lenient* mode and the F-measure score, the best performing semantic expansion mode is the *Hypernym* scoring 82%. However, inspection of the tabular data makes clear that the *Hypernym* mode does not provide the best Precision score, which is delivered by the *Hyponym* and the *Synonym* modes, both scoring 80%, whereas *Hypernym* Precision score is *78%*. In terms of Recall, the *Hypernym* mode delivers the best score 87%, which is very close to the Recall score delivered by the *All-Available* mode 88%. On the other hand, the *All-Available* mode delivers the lowest Precision score (73%), which is expected since the mode uses all the available terms of thesauri and glossaries, including those which do not relate strongly with the targeted entities.

However, the results of this mode are not dramatically low, probably because the terminological resources relate to the Cultural Heritage domain. Possibly the use of a more general purpose terminological resource would have generated a lower precision score for *All-Available*.

	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. 80	0.88	0.67	0.73	0.72	0.79

Table 8.5: Precision, Recall and F-measure results for the 5 Semantic expansion modes

Figure 8.3 presents the distribution of Precision, Recall and F-measure scores for the five different modes of semantic expansion. The graph displays the behaviour of the three metrics for the five modes starting with the *Only-Glossaries* and ending with the *All-Available* mode. The highest F-measure score (shown in yellow) is delivered by the Hypernym expansion mode. Recall increases until reaching the Hypernym mode and then remains fairly stable. On the other hand, Precision graph from *Only-Glossaries* to *Hyponym* mode shows a slight increase which then turns to a decrease from *Hyponym* to *Hypernym* and to an even steeper decrement from *Hyponym* to *All-available* mode.

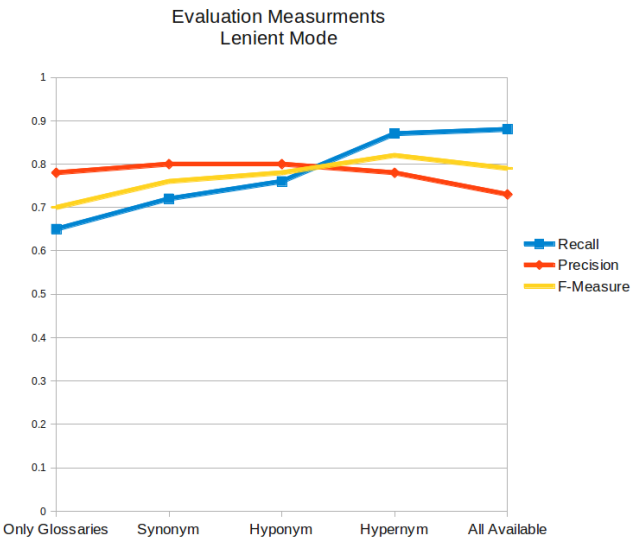


Figure 8.3: Recall, Precision and F-measure evaluation metrics for the Five modes of Semantic Expansion (in lenient mode).

By examining the above graph, a generic observation can be made regarding exploitation of terminological resources. The F-measure score, as the harmonious balance between Recall and Precision, reaches a maximum, when the contribution of terminological resources enables maximum Recall with the least possible affect on Precision. 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 increases including more and more terms from synonyms to narrower and to broader concepts. While Recall increases as the number of contributing concepts increases, up to the point that reaches a plateau, Precision begins to decline at the point where more than the necessary terms are exploited by the NER task.

The Hypernym mode of Semantic Expansion delivers the best F-measure score because it manages to support a high Recall rate without harming too much the Precision result. Although, this particular mode does not provide the best Precision rate, it can be regarded as the optimum choice for supporting an Information Retrieval task which focuses on Recall rather than on Precision. On the other hand, the Hyponym mode delivers better Precision rates and can be employed by an Information Extraction task where Precision is more important than Recall. One of the major aims of the Semantic Annotation task of the

PhD work is to support the STAR project by the Semantic Indexing of grey literature resources, in order to enable Information Retrieval and cross searching. Hence, the thesis adopts the Recall oriented, Hypernym mode of semantic expansion as the optimum system configuration for the Information Retrieval purposes of the STAR project and uses this particular mode for the rest of the analysis in order, to benchmark the various system attributes targeted by the evaluation task.

Figure 8.4 presents the system performance with regards to the four CRM entities; E19.Physical Object, E49.Time Appellation, E53.Place and E57.Material in *Lenient* mode. Table 8.6 presents the full set of results for both *Average* and *Lenient* mode of reporting. The graph presents the F-measure scores of the four entities for the 5 modes of semantic expansion. As discussed above, the Hypernym mode of semantic expansion delivers the best F-measure rates.

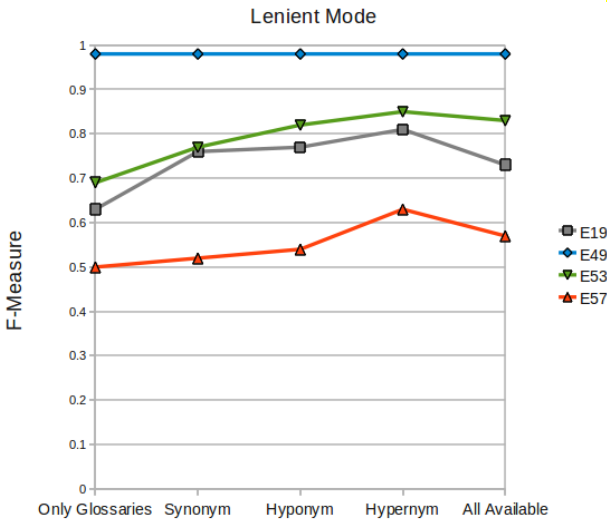


Figure 8.4: F-measure scores of four CRM entity types for the 5 modes of semantic expansion

However, there is a difference in the system performance between the different entity types, reaching for some cases (E49 Time Appellation – E57 Material) almost double scores.

	E19		**E49**		**E53**		**E57**	
	Average	*Lenient*	*Average*	*Lenient*	*Average*	*Lenient*	*Average*	*Lenient*
Only-Glossary	0.58	0.63	0.93	0.98	0.58	0.69	0.46	0.50
Synonym	0.70	0.76	0.93	0.98	0.65	0.77	0.48	0.52
Hyponym	0.72	0.77	0.93	0.98	0.69	0.82	0.49	0.54
Hypernym	0.75	0.81	0.93	0.98	0.73	0.85	0.58	0.63
All-Available	0.67	0.73	0.93	0.98	0.71	0.83	0.51	0.57

Table 8.6: F-measure score of four CRM entities (E19.Physical Object, E49.Time Appellation, E53.Place and E57.Material) for the five modes of semantic expansion

The system performs best (98%) for the Time Appellation entity type (E49). The performance is the same across all 5 modes of semantic expansion because the entity is not affected by the expansion modes. The NER task does not rely on a particular glossary for

the identification of Time Appellations instead, it uses All-Available concepts of the EH Timeline thesaurus. The very good performance of the system is based on the completeness of the Timeline thesaurus to support the task with a sufficient set of non-ambiguous terms. The Timeline thesaurus is the only terminological resource which contributes to the NER that does not have any overlapping terms with other terminological resources. The purity and completeness of the thesaurus resources in addition to their enhancement as "skosified" gazetteer resources (section 4.4), helps the delivery of high Precision and Recall rates as these are reflected by the F-measure score.

The results of Physical Object (E19) and Place (E53) entities range from 63% to 85% depending on semantic expansion mode. Places include archaeological contexts and larger groupings of contexts (but not locations which are not the focus of the semantic annotation). The highest score for both entities is delivered by the Hypernym expansion mode reaching 81% and 85% for the Physical Object and the Place entity respectively.

The system delivers the lowest F-measure score (50%) on the recognition of Material (E57) (table 8.6). The Material entity itself is influenced by ambiguities that are 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. This fine distinction is expressed by contextual arrangements which are challenging to identify and to extract as discussed in sections 5.6 and 5.6.4.

The NER pipeline invoked a specific NLP module (section 5.6) aimed at identifying contextual evidence that could be used for disambiguating the physical object from material sense. However, the results show (table 8.7) that recognition of Material can be problematic and hard to tackle. The system delivers best Recall score 77% (Hypernym mode) and best Precision score 54% (Hypernym mode) which is indicative of the challenge in distinguishing material from physical object entities in archaeological text. The NER of Material entities is supported by terminological resources that contain a large amount of overlapping concepts with different glossaries and thesauri. The extended overlap of such resources has influenced the performance of the system for this particular entity type as is evident from the low Precision scores (table 8.7).

	E19		**E49**		**E53**		**E57**	
	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

Table 8.7: 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.

The complete set of F-measure results (table 8.6) of the four entities for the five modes of semantic expansion, describes the Hypernym mode as the best performing in terms of overall F-measure score (82%). On the other hand, the Hyponym mode is the best performing mode in terms of overall Precision score (80%) that can be employed by Precision focused extraction tasks.

8.5.2 CRM-EH Relation Extraction Evaluation Results

The second phase (figure 8.5) used the CRM-EH version the Gold Standard to evaluate the CRM-EH Relation Extraction (RE) and CRM-EH Named Entity Recognition (NER) specialisation performance of the system. In terms of RE, the evaluation addressed the system performance on identification of 'rich' phrases that relate entities under CRM-EH event or property descriptions. On the other hand, the evaluation of the CRM-EH NER specialisation technique addressed the system's performance in terms of delivering specialised CRM-EH entities using the results of RE, as discussed in sections 6.2.2.2 and 6.4.5. Thus the second phase of the evaluation has executed two main and two complementary configurations addressed at the performance of RE and CRM-EH NER specialisation.

The first main configuration used the RE rules based on the syntactic (part-of-speech) patterns, which resulted from the bottom up analysis (section 6.3.2) while the second configuration used RE rules based on simple offset spans that did not employ any syntactic evidence. The additional complementary configurations were targeted at the specialisation of the CRM-EH entities (EHE0007, EHE0009, and EHE0030). The first complementary configuration required the participations of entities in rich phrases (CRM-

EH events) in order to qualify as CRM-EH, while the second configuration (CRM-EH with Singletons), specialised all CRM entities to CRM-EH without restrictions.

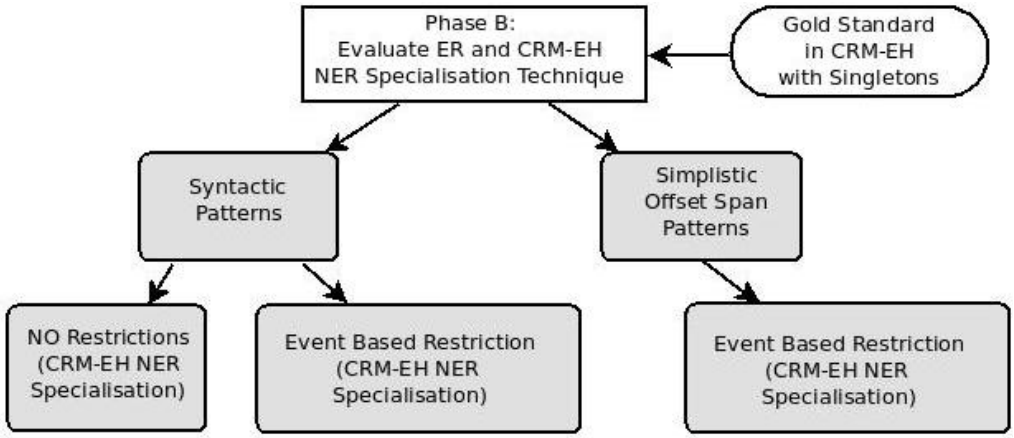


Figure 8.5: Main Evaluation Phase B: Two, plus two system configuration modes

This phase was aimed at evaluating the results for the CRM-EH entities *EH0007.Context*, *EHE0009.ContextFind, EHE0030.ContextFindMaterial* and the CRM-EH events EHE1001.ContextEvent, *EHE1002.ContextFindProductionEvent* and *EHE1004.ContextFindDepositionEvent,* including as well the CRM property *P45.consists_of,* which is used to relate the entities *EHE0030* and *EH0009* under the property (P45) physical object consists of material.

The phase executed three different systems configurations targeted at evaluating the ontological specialisation provided by the pipeline. In detail;

- The first configuration (Singleton) extracted relations between entities using the bottom-up analysis syntactic patterns (section 6.4) while specialising without restrictions any singleton CRM entity, previously identified by the NER pipeline, with its corresponding CRM-EH entity. Hence, entities previously identified as Physical Object, Place and Material were re-annotated without restrictions to CRM-EH Context Find, Context and Context Find Material respectively
- The second configuration (Via Events) used again the bottom-up analysis syntactic patterns for extracting relations between entities but restricted specialisation of CRM entities only to those contained within relation phrases. Hence, singleton CRM entities which did not relate with other CRM entities were not specialised to CRM-EH entities. The configuration is given the name *Via Events* by convention, since three out of the four phrase types extracted by the RE phase are modelled as CRM-EH events. The only exception is the relation between Physical Object and Material which is modelled as CRM property *consists_of.*

- The third configuration used simplistic, offset-based patterns that did not use any syntactic evidence for extracting CRM-EH events, running on restricted specialisation of CRM entities only to those contained within relation phrases. The annotation results of the third configuration were used to compare and contrast with the system configuration that used the bottom-up analysis patterns.
- A fourth configuration which was not executed, could have used the simplistic, offset-based patterns running on unrestricted specialisation (i.e. singleton). However from an evaluation point of view, there would be no significant benefit to examine a non-restricted specialisation of CRM-EH deriving from simplistic offset-based patterns. The main interest of the evaluation was to address which of the two configurations (i.e. syntactic-based or offset-based) delivers best RE and then to examine, based on the best performing RE mode, the case of CRM-EH NER specialisation (i.e. Singleton vs. Via Event). Benchmarking the CRM-EH NER specialisation result of a RE extraction mode that is outperformed by another mode would not deliver any significant findings. Hence it was decided to exclude this fourth configuration from the evaluation phase.

The evaluation results of Relation Extraction for the Syntactic-based and Offset-based system configurations are shown in table 8.7. The table 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. The Offset-based system used simplistic rules in form of *<entity><up to 5 tokens><Verb><up to 5 tokens><entity>* whereas the Syntactic-based system employed sophisticated rules as discussed in section 6.4. Figure 8.6 presents the results of the two different configurations for the range of different CRM-EH events (EHE1001.ContextEvent, EHE1002.ContextFindProductionEvent, EHE1004.ContextFindDepositionEvent) and CRM property (P45.consists_of).

	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

Table 8.8: Precision, Recall and F-measure of relation extraction (CRM-EH event types) between the Offset-based and Bottom-up system configurations.

The results of table 8.8 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. 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 Syntactic-based configuration delivers higher F-measure scores for the range of CRM-EH events and property (Figure 8.6) whereas the Offset-based system delivers better Recall rates but not much higher than the Syntactic-based system. On average, the Offset-based mode delivers 8% higher Recall results than the Syntactic-based system. On the other hand, the Syntactic-based configuration delivers much higher Precision results than the Offset system for the range of CRM-EH events and property. On average, the Syntactic-based configuration delivers 22% higher precision. The significant improvement in the Precision, in combination with the constrained drop in Recall, gives a considerable advantage to the Syntactic-based over the Offset-based configuration. Therefore, the Syntactic-based configuration delivers better RE results and is regarded as a better overall configuration.

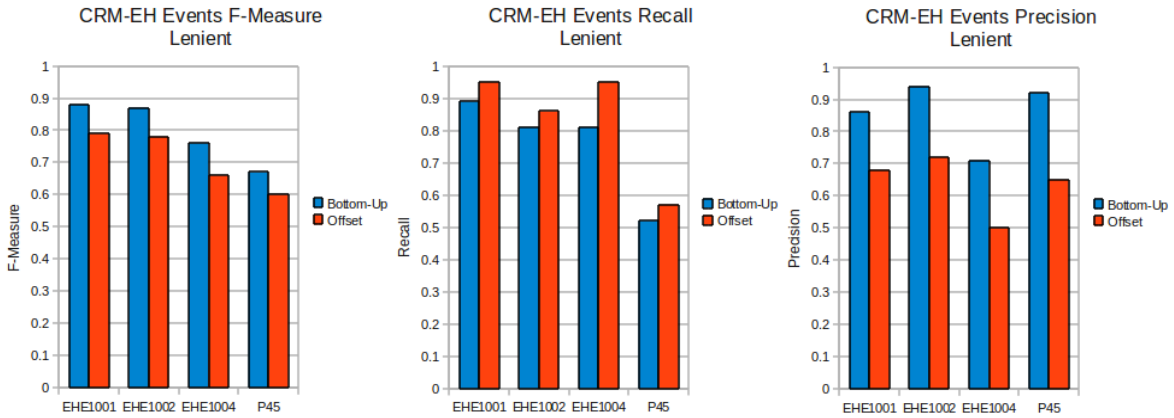


Figure 8.6: Evaluation Metrics (Recall, Precision and F-measure) of the Syntactic-based and Offset-based system configurations for the range of relation extraction phrases (CRM-EH event)

Based the on the above findings with regards to the best configuration, the evaluation proceeds with two alternative system configurations (*Singleton* and *Via Events*) that examine the system's performance with regards to the CRM-EH NER specialisation technique. The evaluation is based on the Syntactic-based configuration on the merit that is the configuration delivering best RE results. Table 8.9 presents the evaluation metrics Recall, Precision and F-measure in both *Average* and *Lenient* mode of reporting, for all CRM-EH entities and event entities participating in the evaluation. Thus, the table presents an amalgamated view of the system's performance for both RE and CRM-EH NER specialisation outcome using the Syntactic-based configuration. The evaluation results of the two system configurations are also presented graphically by figure 8.7.

	Recall		Precision		F-measure	
	Average	*Lenient*	*Average*	*Lenient*	*Average*	*Lenient*
Singleton	0.70	0.78	0.71	0.80	0.70	0.78
Via Events	0.52	0.58	0.81	0.90	0.60	0.66

Table 8.9: Precision, Recall and F-measure results for the two system configurations on CRM-EH extraction phase, including both RE 'rich' phrases (CRM-EH Events) and CRM-EH NER.

Based on the F-measure score of the evaluation results, the Singleton configuration outperforms the Via Events configuration by 12%. In terms of Recall the Singleton system scores 20% higher than the Via Events , which suggests that a great deal of singleton entities can qualify for CRM-EH specialisation without requiring any additional contextual evidence. On the other hand, the Via Events system delivers better Precision scores, outperforming the Singleton configuration by 10%. This supports the initial assumption that contextual evidence can support the specialisation of CRM entities as CRM-EH. However, the overall performance of the Via Events system is lower than the Singleton configuration due to its limited Recall.

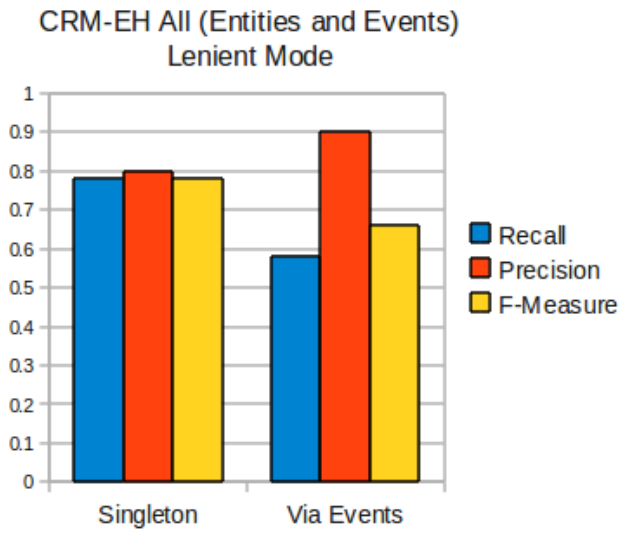


Figure 8.7: Recall, Precision and F-measure evaluation metrics of Singleton and Via Events modes of CRM-EH system configuration for all types (Entity, Event and Property)

A comparison between the *Average* and *Lenient* modes of reporting shows that the *Lenient* results are higher by an average 7.5%. This is slightly higher than the difference between the two modes when reporting the NER results (5%) due to the use of relation extraction phrases (CRM-EH events) in the calculation.

The difference between *Average* and *Lenient* modes on the performance of relation extraction phrases is around 10% (Table 8.10), which is explained when considering that annotation of such events (relations between entities) is based on phrases and differences on annotation boundaries are much more likely to occur than when annotating single entities. The table summarizes the system's performance on the extraction of 'rich' phrases on the Syntactic-based based mode. The overall Relation Extraction F-measure score on the *Lenient* mode of reporting is 80%, slightly higher than including both specialised

CRM-EH entities and relation extraction phrases (78%). Even on the *Average* mode of reporting the system delivers F-measure 70% which is considered encouraging based on the difficulty and complexity of the task.

	Recall		**Precision**		**F-measure**	
	Average	*Lenient*	*Average*	*Lenient*	*Average*	*Lenient*
CRM-EH (Relation Extraction)	0.67	0.75	0.76	0.86	0.70	0.80

Table 8.10: Evaluation results of relation extraction (Event and Property entities EHE1001, EHE1002, EHE1004, P45)

The distinction between the two system configurations (Singleton -Via Events), affects only the performance on the CRM-EH entity types not the relation extraction phrases (CRM-EH events). Both systems deliver the same results with regards to the relation extraction phrases since the specialisation configuration is applied only on CRM-EH entity types. A better approach for comparing the performance of the two configurations, is to present the results only for the CRM-EH entity types excluding the CRM-EH events. Table 8.11 presents the evaluation metrics for the three CRM-EH Entity types (EHE0007.Context, EHE0009.ContextFind and EHE0030.Material) while figure 8.8 presents the results graphically for the *Lenient* mode of reporting.

	Recall		**Precision**		**F-measure**	
	Average	*Lenient*	*Average*	*Lenient*	*Average*	*Lenient*
Singleton	0.74	0.82	0.65	0.72	0.67	0.76
Via Events	0.32	0.34	0.88	0.95	0.45	0.49

Table 8.11: Precision, Recall and F-measure of the two system configurations only on CRM-EH Entities (EHE0007.Context, EHE0009.ContextFind and EHE0030.Material)

The results of table 8.11 show that when removing the CRM-EH Event entities from the calculation of the evaluation metrics, the performance difference between the two system configurations grows significantly. The difference in Recall between the two system configurations increases to 42% and in Precision to 23%. Overall the Singleton configuration, based on the F-measure score, outperforms the Via Events system by 27%.

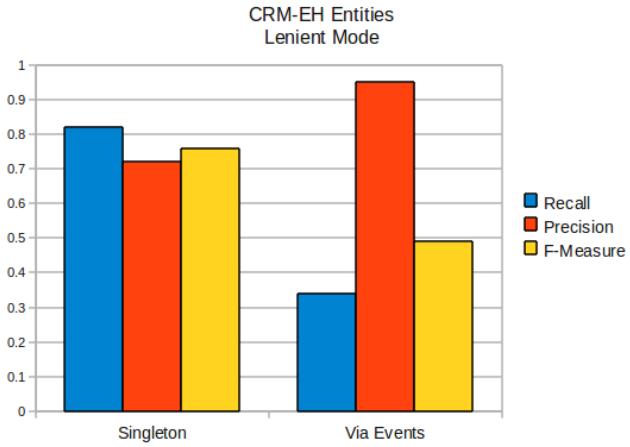


Figure 8.8: Evaluation Metrics of CRM-EH Entities on Singleton and Via Events modes

On the other hand, the Via Events system delivers a very high Precision score that reaches 95%. As mentioned above, this particular behaviour supports the initial assumption that (precision enhancing) CRM-EH entity specialisation can be reached via relation extraction (i.e. via CRM-EH events).

Possibly the good performance of the Singleton configuration derives from the use of archaeological documents by the evaluation corpus. Due to this, many CRM entities that were identified by the NER configuration can qualify as CRM-EH entities without following any contextual requirement because they originate from an archaeological text. On the other hand, the use of a more general purpose text could require a more sophisticated and restricted application of the CRM-EH specialisation. The contextual demanding approach of the Via Events configuration could be employed to provide the required specialisation to CRM-EH for entities that originate from a general purpose text. However, a full scale exercise is required in order to test the validity of this hypothesis.

The evaluation results of the Singleton and Via Events system configurations for the individual CRM-EH entities are presented in figure 8.9. The highest F-measure score of the Singleton configuration is 85% for the EHE0007 entity, followed by 80% for the EHE0009 and 64% for the EHE0030 entity. On the other hand, the highest F-measure score of the Via Events system is 65% for the EHE0009 entity, followed by 46% for the EHE0007 and 35% for the EHE0030 entity. The Singleton configuration outperforms the Via Events configuration for all entity types. However, both configurations have their lowest F-measure score on the EHE0030 entity, indicative, as discussed in NER results, of the challenge in the annotation of material related entities.

The disagreement between the two configurations for the highest scoring entity, suggests that the volume of singleton EHE0007.Context entities is greater than the volume for EHE0009.ContextFind. This is also reflected by the Recall rates, where the two configurations for EHE0007.Context have a massive difference of 60%, with the Singleton configuration scoring 91% and the Via Events just 31%. The difference between Recall rates of the two system configuration is also evident with the other two entity types where the difference on EHE0009.ContextFind is 30% and on EHE0030.ContextFindMaterial is 56%. This significant improvement of the Recall rates by the Singleton configuration, as already discussed above, suggests that singleton entities of archaeological text can be specialised to CRM-EH entities without additional contextual evidence, as for example “pits” in the phrase “There is a large cluster of pits to the east, a network of east-west and north-south”

The Via Events configuration outperforms the Singleton in Precision rates for all CRM-EH entities. Particularly in the case of EHE0030 the Via Events systems delivers 100% Precision, which suggests that all (23) CRM-EH Material entities that are identified by the system are associated with a Context Find entity. The ability of the Via Events system to deliver high Precision results can be exploited in cases where the extraction of CRM-EH entities is required but the origin of the document is not clearly from the archaeological domain.

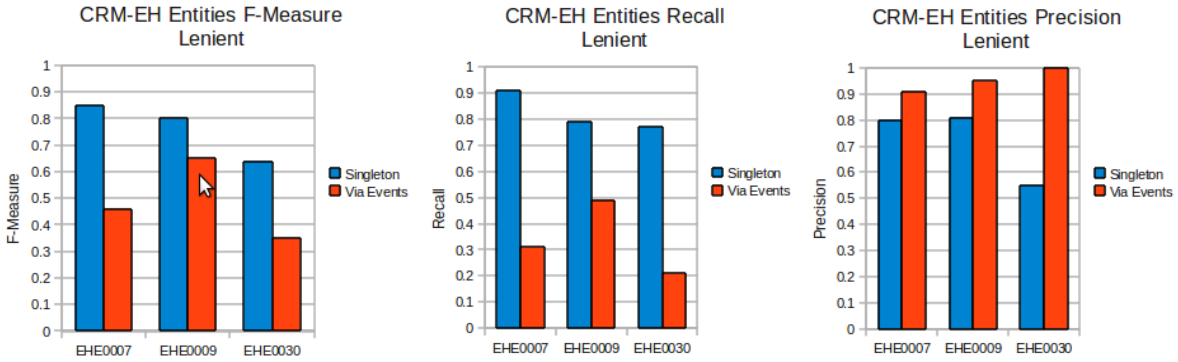


Figure 8.9: Evaluation Metrics of CRM-EH entities (EHE0007.Context, EHE0009.ContextFind and EHE0030.Material) on Singleton and Via Event modes

8.5.3 NLP Modules Evaluation Results

The third phase (Figure 8.10) aimed to evaluate the contribution of the various NLP techniques that contributed to the NER phase. The phase used the CRM version of the Gold Standard and ran five different system configurations which were executed in the *Hypernym* semantic expansion mode. The IE system was stripped of all NLP modules that were used by the NER pipeline to improve accuracy of performance, such as the *Noun Phrase Validation*, *Negation Detection* and *Word Sense Disambiguation* modules. The additional concepts that were added in the matching mechanism by the pilot evaluation were removed also. A basic configuration (*Basic)* was executed and the results were used as indicator of the system performance, without the use of accuracy techniques. The contribution of each individual NLP module was then evaluated by adding the module to the *Basic* configuration and comparing the results. Four different configurations were executed: Basic plus Noun Phrase Validation, Basic plus Negation Detection, Basic plus Disambiguation module and Basic without Pilot Evaluation Added Concepts.

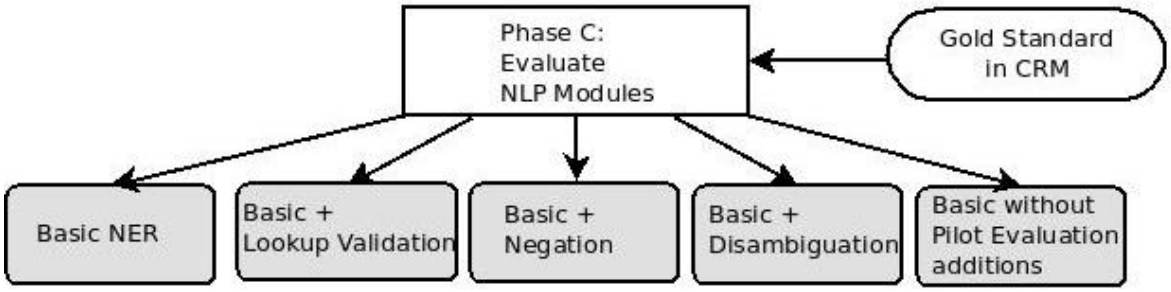


Figure 8.10: Main Evaluation Phase C: Five system configuration modes

This phase of the evaluation aimed at benchmarking the contribution of individual bespoke NLP modules which were invoked by the NER pipeline. In detail, the phase evaluated the contribution of the *Noun Phrase Validation, Negation Detection and Word Sense Disambiguation* modules. In addition, the contribution of the intellectually added concepts [Appendix D7] to the NER system, which resulted from the Pilot Evaluation phase, was also evaluated. The process of evaluation was executed in the *Hypernym* mode of semantic expansion. The selection of this particular mode was based on the performance criteria discussed above, although any mode of semantic expansion could have been employed to evaluate the contribution of the bespoke NLP modules.

The evaluation phase followed the execution of five different system configurations. The NER system initially was stripped from the bespoke NLP modules and from the intellectually added concepts. This particular system configuration was named *No Additions,* which represented the NER system in its simplest form. Then the intellectually added concepts were included in the configuration, which was named *Basic* and represented the simplest NER system configuration including the intellectually added terms.

The *Basic* configuration was used as the main platform of the evaluation phase where each bespoke NLP module was combined separately i.e. Basic combined with Negation Detection, Basic combined with Noun Phrase Validation and Basic combined with Word Sense Disambiguation. Hence, by using the Basic configuration as the common reference point it is possible to identify the contribution of the NLP modules independently, as shown in figure 8.11.

The No Additions which is the simplest configuration of all, delivers the lowest ratings which though are not disappointing. The configuration manages to deliver adequate Recall rate (81%), however, the Precision score is low (54%) which affects the F-measure score (65%). When the intellectually added concepts are included to form the Basic configuration the Recall rate increases by 8 points reaching 89%. Precision though remain low 55%, indicative of the need to deal with language ambiguities that affect the NER results.

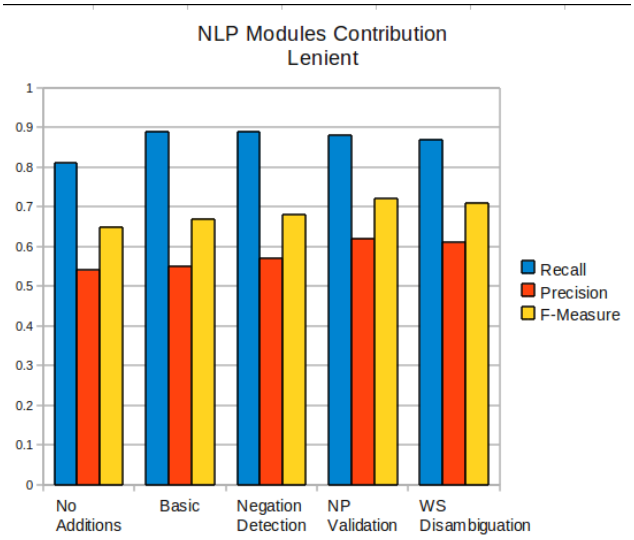


Figure 8.11: Evaluation Metrics of bespoke NLP modules contributing to the NER pipeline

The Precision rate improves when adding the bespoke NLP modules. The Negation Detection module improves Precision by 2% without harming the Recall rate. The improvement brought by the Negation Detection module might not be high but is explained by the limited number of negation phrases included by the evaluation corpus. From the 1099 NER entities delivered by the system, running on the Hypernym expansion mode, only 33 were negated phrases. The limited number of negation examples in the evaluation corpus is because the corpus selection was focused on summary extracts independently of the volume of negation phrases contained.

The Noun Phrase Validation module improved system's Precision by scoring 62% while affecting Recall only by a single unit delivering 88%. Positive was also the contribution of the Word Sense Disambiguation module which increased Precision to 61% while affecting Recall by 2 units delivering 87%. Based on the above evaluation metrics, every single bespoke NLP module makes a positive contribution to the system, mainly improving Precision while affecting Recall only slightly by one or two units.

The combination of all bespoke NLP modules in the pipeline improves the overall Precision by 23%, delivering for the Hypernym mode 78%, compared with the Basic configuration score of just 55%. At the same time the bespoke NLP modules only slightly affect the Recall rate by dropping the overall Recall rank score by 2% from 89% to 87%, thus improving the system performance in terms of F-measure score from 67% to 82%.

8.5.4 Evaluation via Author-based Metadata

An additional phase of evaluation compared the NER semantic annotation result with existing author-based metadata of the OASIS documents. This particular evaluation task has a qualitative basis, hence it does not deliver typical evaluation figures in terms of Recall and Precision. Instead the task compares and contrasts the CRM based semantic annotation of 16 OASIS documents with the available metadata. The list of author-based metadata and semantic annotation of documents can be found in [Appendix D6].

The documents were selected to correspond to the 16 commercial archaeological units which participated in the evaluation corpus. The author-based metadata contain a range of different categories, such as Author, Site, District, County, Parish, Grid References, Monuments and Finds. The volume of available metadata varies between documents. Some documents contain only a few entries, some a couple of dozen of entries and some others do not contain any. The volume of metadata is usually influenced by the document 'richness'; larger reports tend to contain a greater volume of metadata. The metadata types of Finds and Monuments are abstract and vary in size, some are single words while some other are phrases.

The evaluation is based on examining the Monument and Finds metadata of each of the 16 documents, which are semantically close to the E53.Place and E19.Physical_Object annotation types. Appendix D6 holds the metadata and semantic annotations for each document. Each row contains the documents name, the author-based metadata of Monuments, of Finds and the semantic annotations of E53.Place and E19.Physical_Object. Due to the large volume of system generated E19 and E53 semantic annotations for each document, only the three most frequent semantic annotation entries are included. The frequency number is displayed in brackets, for example “deposit(17)” indicates 17 occurrences of “deposit” are annotated. In addition, semantic annotation entries corresponding to the author-based metadata that are not within the three most frequent annotations are also included. For example, for the metadata entry “Garden”, the semantic annotation “Garden(2)” is included, which might not be among the most frequent annotations. In some cases the semantically closest annotation entry is chosen, for example for the Metadata entry “Market garden” the annotation “garden” is selected.

Overall, the majority of semantic annotations per document correspond to the available document metadata and vice versa. In most cases, the semantics of frequent annotations match the semantics of the available metadata, while less frequent annotations are also available corresponding to the range of metadata. For example, *archaeol1-19366_1* has for

the author-based Monument metadata *building, field system, enclosure, kiln, pit, oven, post hole, butchery site* and the corresponding most frequent E53 annotations are *pit(303), ditch(192), deposit(89)* while *building(22); enclosure(70), kiln(17), oven(38), post-hole(66)* are also available. On the other hand, two documents have no available metadata but have semantic annotations.

In two cases the semantic annotations do not fully correspond to available metadata. For the "archenfi2-31470_1" document, *Ceramics* is assigned to Find metadata while the corresponding annotations are *clay(112), iron(33), charcoal(33), slag(19)* and for the "clairefe1-8958" document, *Crop mark* is assigned to Monument metadata while the corresponded annotations are *Pottery(17), finds(13), sherds(4), Village(4), surface(3), road(3), burial(3), trackways(3), rectilinear enclosures(3).* However, the limited assignment of author-based metadata in both cases cannot support any strong conclusions regarding the completeness of the available semantic annotations. On the other hand, the extensive and elaborate assignment of metadata in document "cambridg3-27196_1" required the examination of the CRM-EH Event and Property metadata in order to find corresponding annotations. The document metadata includes phrases such as c*opper alloy bracelets; jet bead necklace; iron hobnails* for describing Finds. An examination of the CRM-EH annotations revealed phrases such as, *several copper-alloy bracelets; copper alloy finger ring; iron nails; preserved necklaces associated with this adult burial;* which correspond to the author-based metadata.

Overall, semantic annotations generated by the OPTIMA pipeline approximate reasonably closely to the author-based metadata. Human authored metadata can be very precise but sometimes very abstract, encapsulating rich semantics in a single phrase. For example the metadata entry "Listed Hall and Gardens" can be well understood by humans but cannot be processed very easily by machines on a semantic level, due to the use of conjunction. On the other hand, semantic annotations which enjoy conceptual and terminological references can be easier to process by machines for supporting semantically oriented queries.

8.6 Conclusion

The evaluation task completed a pilot and a main evaluation phase, delivering a wide range of results, which reflect the system performance on semantic annotation with respect to CIDOC CRM and CRM-EH ontologies. A range of different system configurations participated in evaluating the different types of annotation output and testing the contribution of different NLP modules. Considering the limited time and resources, the task has managed to conduct a full scale evaluation which supports the evaluation aims and objectives of the thesis.

The overall system performance can be considered capable of supporting the cross searching aims of the STAR project in terms of the semantic indexing of grey literature documents with respect to the CRM, CRM-EH ontologies. The system delivers F-measure rates of 82% for the NER task and 80% for the CRM-EH RE and entity specialisation tasks when results are reported in the *Lenient* mode and 75% and 70% respectively when results are reported in *Average* mode. The *Lenient* is considered to be the more appropriate mode of reporting due to the significant difference in the use of moderators and treatment of annotation boundaries between annotators as discussed above (section 8.4.2.5). In terms of Recall the system for the task of NER delivers rates between 65% to 87% (*Lenient)* (table 8.5) depending on the mode of semantic expansion employed by the task. On the other hand the system's performance in terms of NER Precision presents less fluctuation with figures balancing between 78% to 80% (*Lenient)* depending on the mode of semantic expansion. The above figures increase further, 70%-90% Recall and 79%-88.6% Precision, (table 8.7) when the material entity is excluded from the calculation due the significant overlap with physical object entity affecting the overall system performance as discussed in sections 4.3.3 and 8.5.1.

From a simple numerical point of view the overall NER results compare favourably with full scale semantic annotation systems targeted at archaeological context that have yielded F-measure score of 75% (Zhang, Chapman and Ciravegna 2010) and full scale systems targeted at historical text that have delivered F-measure score of 73% (Grover et al. 2008). However this is a relatively broad brush comparison as regards overall system performance since, as discussed in section 8.4, the evaluation task followed a user centred approach that differs from a more prescriptive evaluation approach usually followed by the ML tradition where the annotation criteria are spelled out in detail. In order to be able to have a full comparison between different systems, it is required to take full account of evaluation methodology, the details of which are not always supplied. For example, the

SEKT project (Peters et al. 2005) followed the annotation principles of orthography, topicality and phrasal annotation which are also adopted by the current evaluation task but the focus of SEKT was on proper nouns not archaeological entities and 'rich' phrases. On the other hand, Byrne (2009) followed an evaluation methodology which suggested system delivered annotations for approval by annotators, an evaluation method significantly different from the one followed by this thesis as discussed in section 8.4.2.4.

The system delivers best F-measure score (98%) for the Time Appellation (E49) entity due to the purity and completeness of the thesaurus resources involved, in addition to their enhancement as "skosified" gazetteer resources (section 4.4). On the other hand, the system delivers the lowest F-measure score (50%) on the recognition of the Material (E57) entity, due to the domain related challenges imposed in the identification and extraction of such entities as discussed in section 5.6.4. The results of Physical Object (E19) and Place (E53) entities range from 63% to 85% depending on semantic expansion mode.

The F-measure score is the harmonious balance between Recall and Precision that can be used to reflect the overall system's performance. The system delivers the highest F-measure score (82%) for the task of NER in the case of the *Hypernym* semantic expansion mode, which enables maximum Recall (87%) with the least possible effect on Precision (78%). On the other hand, the best Precision performing mode (80%) is the *Hyponym* mode of semantic expansion. The *Synonym* mode also delivers high Precision score (80%) but less Recall than *Hyponym* (72% vs. 76%). Thus the system is configurable to run in different semantic expansion modes that favour Recall or Precision depending on the aims of a given task.

With regards to the task of Relation Extraction (RE), the best performing mode is the Syntactic-based delivering 75% Recall, 86% Precision and 80% F-measure with all figures reported on the *Lenient* mode (table 8.10), in contrast to the Offset-based mode which, although it delivers higher Recall (83%), is outperformed in Precision (64%) by 22%, thus scoring a lower F-measure score (70%) (table 8.8). From a simple numerical point of view the above results are very comparable with IE driven by machine learning engines targeted at extracting relations from archaeological text, delivering F-measure score of 75% (Byrne 2009) and rule-based, ontology guided systems targeted at biomedicine text delivering F-measure scores between 64% to 76% (Cimiano et al. 2005; Friedman et al. 2001). However, the method and scope of Relation Extraction differs significantly on different projects and the comparison is only indicative. Byrne (2009) 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. 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 (Obio ontology).

The Relation Extraction (RE) phase influences the results of the CRM-EH NER specialisation mode when the system requires co-occurrence of individual entities within 'rich' phrases extracted by the RE task (i.e. Via Events mode). Thus, when the results of RE are combined with the results of NER with regards to the specialisation of CRM-EH entities based on the Via Events mode, then the system delivers 58% Recall and 90% Precision (table 8.9). On the other hand, when the RE results are combined with CRM-EH NER specialisation not requiring co-occurrence of individual entities within 'rich' phrases (i.e. Singleton mode) then the system delivers 78% Recall and 80% Precision.

The system delivers F-measure 78% (table 8.9) when including together with 'rich' phrases the specialised CRM-EH entities derived by the Singleton configuration. This figure drops further (F-measure 66%) when specialisation to CRM-EH entities is restricted only to rich phrases derived by the Via Events configuration. This specialisation technique though, delivers greater Precision rates (81%) than when singletons (entities outside 'rich' phrases) are included (71%). On the other hand, the Singleton configuration delivers higher Recall 78% versus 58% of the Via Events configuration, thus delivering a higher overall F-measure performance (F-measure 78%) than the Via Events (66%) (table 8.9). The Singleton configuration revealed that CRM-EH entity specialisation on archaeological grey literature (excavation and evaluation reports) can be applied without necessarily requiring additional contextual evidence (i.e. occurrence within a relation 'rich' phrase). Thus it is considered to be a better performing configuration for tackling CRM-EH specialisation on archaeological grey literature.

The ability of the two different specialisation techniques (i.e. Singletons – Via Events) to tackle either Recall or Precision is shown on CRM-EH NER results of the EHE0007.Context and EHE0030.ContextFindMaterial entities (Figure 8.9). In the first case the Singletons configuration deliverers 90% Recall whereas the Via Events delivers 30% Recall while in the second case the Via Events configuration deliverers 100% Precision whereas the Singletons delivers 55% Precision. A significant improvement in Precision was also achieved with the use of Syntactic-based patterns for matching 'rich' phrases which outperform the simple offset-based matching by 22% in precision and 10%

in F-measure score, with the offset-based system to deliver better recall score by 8% (Figure 8.9). Overall the best performing configuration for the task of RE and CRM-EH NER specialisation is Syntactic-based mode running on Singleton configuration.

The system's performance is considered sufficient to support the task of semantic indexing since the system delivers higher results than the above scores for the vast majority of entities and events. However, the inclusion in the overall scores of the problematic and ambiguous entities *E57.Material*, *EHE0030.Context Find Material* and the CRM property *P45.consists of* affects the overall system performance. Excluding the above entities from the computation of the overall results, the system delivers F-measure (88% *Lenient*, 81% *Average*) for the task of NER and F-measure (83% *Lenient* and 74 *Average*) for the task of RE including CRM-EH specialisation.

In addition, the evaluation revealed the role of NLP techniques in improving performance of the IE system. The employment of Noun Phrase Validation, Negation Detection and Word Sense Disambiguation modules has improved the overall NER F-measure score of the system by 15%. Every single bespoke NLP module makes a positive contribution to the system. When all modules are combined together, Precision increases by 23% while Recall is only slightly reduced by 2%.

From a qualitative point of evaluation, the semantic indexing results were compared with available author-based metadata. In most of the examined cases (9 out of 10) the automated semantic annotations match or approximate reasonably closely the author-based metadata, which were limited in size and coverage compared to the volume of semantic annotations of each document [Appendix D6].

The evaluation process also revealed the difficulty of defining an ambiguous and commonly acceptable Gold Standard. The critical role of the Super Annotator for normalising and reconciling differences between individual annotators was also revealed during the evaluation process. The Pilot evaluation significantly helped towards the definition of the Gold Standard, allowing a test of the manual annotation instructions and the annotation process before committing to a full scale evaluation. However, the Pilot Evaluation did not pursue the delivery of a final version of the Gold Standard; instead it focused on the IAA analysis for the construction of manual annotation instructions and rehearsing the manual annotation process. Although, the delivery of a clear and comprehensive set of instructions is critical for the successful completion of the annotation task, the pilot evaluation should have conducted a full analysis phase beyond the point of delivering annotation instructions and rehearsing the task. A fully completed pilot

evaluation would have revealed earlier the issues of moderators and CRM-EH specialisation of singletons which were addressed by the Super Annotator.

The end-user focus of the evaluation did not give the chance to thoroughly evaluate the performance of specialised NLP modules such as *Negation Detection* and *Word Sense Disambiguation*. The evaluation corpus was selected with criteria that addressed the diversity of document types, richness of discussion and archaeological units participating in the OASIS corpus. Considering the available time and resources, the evaluation aimed to assess the semantic annotation result with respect to the end-user focus directed towards the (cross search retrieval) aims of the broader STAR project and its intended users, archaeology researchers and HE users. Given the necessary time and resources, the evaluation could have completed individualised phases for evaluating the performance of particular NLP modules via specialised Gold Standards. In addition, it could have defined an additional Gold Standard to assess the contribution of manually added matching concepts via a bootstrapping method, i.e. to iterate via different sets of Gold Standard until there was no more significant improvement in Recall via manually added concepts.

Chapter 9

Conclusions and Future Work

9.1 Conclusions

The aim of this thesis has been the development of NLP techniques for automatic indexing of archaeological grey literature for purposes of semantic interoperability. The delivered semantic indices support information retrieval, cross searching and document inspection via ontological and terminological definitions, with respect to the CIDOC CRM, CRM-EH ontologies and the SKOS English Heritage terminological resources. The developed pipeline (OPTIMA) employs the tasks of NER and RE for the semantic annotation and indexing of archaeological grey literature documents with respect to a sub-set of CRM and CRM-EH concepts. The following sections of this chapter summarise the main observations with regards to contributions to knowledge, methodology, generalisation and computational deliverables of the work. The last section discusses possibilities for future work and further research.

9.1.1 Contributions to Knowledge

A range of distinct contributions to knowledge have been delivered by the research work. The contributions are two-fold, covering the broader research fields of Digital Humanities and Natural Language Processing. In particular, contributions are made to the fields of digital archaeology and information extraction respectively. With respect to digital archaeology, the research effort contributes a novel method for the semantic and interoperable indexing of archaeological reports (grey-literature documents) with respect to the CIDOC CRM and CRM-EH ontologies. As discussed in sections 7.3 and 7.4, such semantic indices of grey-literature documents can be employed by software applications to support a range of information seeking activities, such as document retrieval, cross searching and document inspection with respect to semantic properties. Hence, the semantic indices make a novel contribution in the use of interoperable standards for the dissemination of archaeological information, aimed at cross-searching and analysis of multiple digital resources, which according to Richards and Hardman (2008) is not well supported by the current fragmented digital archaeology systems.

Individual contributions are also made with regards to information extraction, particularly focused on the tasks of NER and RE using rule-based, ontology driven techniques. In terms of NER, the research effort explored a novel approach based on the complementary use of ontologies and terminological resources for the definition of semantic annotations. As discussed in section 3.2.4, being able to assign ontological and terminological definitions to annotations enhances significantly cross-searching and information retrieval practises and enables context-dependent information extraction practises. In addition, the employment of terminological resources as "skosified" gazetteers facilitates a dynamic exploitation of vocabulary via broad and narrow concept relationships. This enables the development of a configurable NER pipeline capable of performing in both Precision and Recall enhancing modes. Entity identification also benefits from the domain oriented NLP modules of word-sense disambiguation (section 5.6) and negation detection (section 5.8), which improve the overall precision performance of the NER pipeline as revealed by the evaluation results section 8.5.3.

With regards to the task of RE, the development adopts a novel approach for the definition of extraction rules based on syntactic patterns derived by a corpus analysis study (section 6.3.3). The application of the Zipfian distribution principle (section 6.3.1) is central 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. The syntactic patterns are capable of identifying 'rich' phrases that are annotated as CRM-EH events connecting two or more CRM entities. Evaluation results (section 8.5.2) demonstrate the capability of the method to identify three different types of CRM-EH events and one CRM property. The results show benefits in Precision over 20% when using syntactical pattern over simple token offset rules, as discussed in evaluation conclusions section 8.6. Overall the system delivers competitive F-measure rates reaching 82% for the task of NER and 80% for the task of RE.

Additional contributions are also made with regards to the definition of indices as interoperable computing artefacts and pipeline development in terms of vocabulary enhancement procedures. Such contributions are discussed further in section 9.1.4 Deliverables.

9.1.2 Methodology Reflections

The method of rule-based information extraction techniques equipped with domain vocabulary, is suitable for delivering semantic annotation with respect to domain ontologies (in this case CIDOC CRM and CRM-EH) which can facilitate semantic and interoperable access to grey literature documents.

The OPTIMA pipeline development followed an incremental and iterative process passing through a prototype development cycle that explored initial design and implementation issues. The contribution of the prototype development was beneficial in revealing early problems with regards to vocabulary coverage and usage of ontological and terminological resources. In addition, it helped to explore the capacity and flexibility of information extraction techniques for accommodating the task of semantic annotation of archaeological grey-literature documents. The process iterated through evaluation of the prototype system, revealing performance issues, as well as strengths and weaknesses of the evaluation method. The incremental process made it possible to enhance the system and to improve its performance via revising rules and adding bespoke NLP modules, such as noun phrase validation, word sense disambiguation and negation detection.

The end-user focus of the evaluation aimed to support the validity of the evaluation methodology and the application of the ontological model. Evaluation was directed towards the (cross search retrieval) aims of the broader STAR project, being oriented to the intended users (archaeology researchers and HE users). Thus, annotators were asked to exercise judgement as competent users, following the LIS tradition , rather than following a strict rule-book for annotation decisions, perhaps more common in the Machine Learning tradition, where annotation criteria are spelled out in detail.

This evaluation approach aided delivery of a Gold Standard closer to the end-user needs but at the same time made it harder to reach a single and commonly agreed definition. A pilot evaluation stage benefited the method by providing input on Inter Annotator (IA) agreement scores and revealing vague instruction points. Although, annotation instructions were amended after the pilot evaluation, they continued to have their end-user focus and avoid prescriptive elements.

The annotators used moderators, articles and annotation span of “rich” phrases flexibly, which affected the overall IA agreement score. The role of the Super Annotator was critical for normalising individual differences and deriving a definite Gold Standard of the evaluation phase. Reporting evaluation results in both *Lenient* and *Average* modes

provided a complete picture of the system's performance. The end-user focus of the Gold Standard performance permitted different treatment of annotation spans by individual annotators. Hence, reporting in *Lenient* mode allowed benchmarking of the performance without penalising any mismatch on annotation span.

The end-user focus also helps consider the application of the ontological model. The logical distinction between physical object and material defined by the ontology, proved relatively problematic to extract and identify in grey literature text. Natural language can sometimes make a blurry distinction between objects and materials. For example material can sometimes constitute an archaeological find which ultimately is treated as a physical object. This is also reflected by the vocabulary usage and the large amount of overlapping terms found (section 4.3.3) between terminological resources of physical objects and materials.

Complying with the ontological model and dealing with the above distinction was a challenging task that led to the development of the word-sense disambiguation module (section 5.6.3). The evaluation task has also addressed the distinction between the concepts Material and Physical Object, a challenging task influenced by archaeology domain vocabulary use, which generated significant inter-annotator disagreement. Implementing the ontological arrangement between object and material did not appear to provide any significant advantage other than satisfying a formalistic ontological relationship. The instance of 'pottery' (for example) in the reports often tends to be with reference to an implicit notion of an unknown fragment of pottery and possibly being the material of an implicit object. However, in a real world setting, users are more interested in retrieving information about archaeological finds (either material or physical objects) and less interested in the distinction between them. Possibly it would have been more beneficial if archaeological finds were treated as a hybrid entity containing both materials and objects.

The issue of attaining CRM-EH specialisation of annotations was also highlighted by the end-user evaluation. The results of the prototype system suggested that the CRM-EH specialisation can be reached via exploiting contextual evidence in the form of "rich" phrases (i.e. phrases relating two or more CRM entity types). Due to the use of common English terms to describe archaeological contexts and finds, it was assumed that requiring qualification via rich phrases would support the system's accuracy in identifying cases of CRM-EH specialisation. However, this particular method proved to be strict and restrictive for recall rates. Evaluation results revealed that archaeological grey-literature documents, due to their domain specific focus, can deliver CRM-EH annotations in terms of NER

without their necessary occurrence within "rich" phrases. The technique though of validating singleton entities via "rich" phrase context can support precision of NER and may be beneficial for CRM-EH specialisation on documents that have less archaeological focus. However this assumption requires further investigation by a future project.

9.1.3 Generalisation of the Work

The research effort has managed to deliver a semantic indexing system following a methodology that used NLP techniques in combination with domain-specific terminological and ontological resources for the semantic annotation of archaeology grey-literature documents. Both the semantic indexing system and the development methodology can be generalised within the broader digital humanities domain. Given the necessary availability of terminological and ontological resources, the development method can be generalised outside the domain of digital humanities for delivering systems of semantic annotation focused on the tasks of NER and RE via rule-based IE techniques. However, this requires further investigation by a future project. The following paragraphs discuss some generalisation trials of the OPTIMA pipeline and methodology, aimed at the semantic annotation and indexing of digital humanities documents.

The system has addressed the task of semantic indexing of archaeological grey-literature documents originating from the OASIS corpus. In total the pipeline has processed 2460 documents of various archaeological report types, such as excavation, site evaluation, and watching brief reports. Currently there are more than 7000 archaeology grey-literature reports available from the OASIS corpus. The system is capable of processing a large volume of available OASIS documents without restrictions. The GATE platform is scalable and applications are restricted only by the physical memory of the workstation.

The system has also has processed archaeological documents that did not originate from the OASIS corpus. The Museum of London Archaeology (MOLA) service has made available two monograph publications. The "Black Death Cemetery of East Smithfield" and the "Excavations at the Priory of the Order of the Hospital of St John of Jerusalem" The OPTIMA pipeline has processed the two monographs and delivered semantic indices with respect to CRM and CRM-EH ontologies. The process delivered a large number of annotations due to the extensive length of the monographs. An example of the delivered annotations can be found in [Appendix E1]

Compared to the OASIS grey-literature reports, the MOLA monographs were longer containing up to 500 pages. The system managed to process the long documents without problems. However, the processing time increases significantly for documents that exceed 150 A4 pages. A workstation equipped with a Dual Core 2 GHz CPU and 4 GB RAM needs approximately 3 minutes to run the OPTIMA pipeline (pre-processing, NER, and RE) on a document of 150 pages but will take approximately 30 minutes to process a document of 500 pages.

The development methodology has also been expanded to information extraction of cultural object descriptions, in particular classical vases, originating from collection fascicules. The CASIE (Classical Art Semantics Information Extraction) is a collaborative project between the Hypermedia Research Unit (University of Glamorgan) and the Beazley Archive (Oxford University), which aims to automatically extract information about cultural objects from classical art scholarly texts and represent this information in terms of the CRM. The project applied a methodology comparable to the OPTIMA development, based on the employment of rule-based IE techniques supported by domain vocabulary and driven by CRM ontological arrangements.

In total 12 documents (fascicules – high quality catalogues) were processed, originating from the British Museum, the Ashmolean Museum (Beazely Archive) and the Thessaloniki Archaeological Museum catalogues. The 12 fascicules are part of the Corpus Vasorum Antiquorum (CVA) collection containing over 350 high quality catalogues of mostly ancient Greek painted pottery, illustrating more than 100,000 vases. The CASIE pipeline managed to extract information about the individual artefacts, in terms of their type (fabric), dimensions (height-diameter), catalogue reference and description [Appendix E2] from the set of available fascicules, which contained more structured free text information than the OASIS reports.

The extracted information was expressed in interoperable RDF graphs consistent with the CLAROS project format (Kurtz et al. 2009). CLAROS (Classical Art Research Online Services; www.clarosweb.org) is an international interdisciplinary research initiative led by the University of Oxford, aimed at the semantic integration of world classical art records located in major collections of university research institutes and museums. The role of CIDOC-CRM is central for enabling semantic interoperability across the range of datasets that contribute to CLAROS. The CASIE project delivered CRM compliant RDF graphs of the extracted information [Appendix E2]. Although, it was a pilot and exploratory project, it managed to successfully generalise the OPTIMA method based on

the complementary exploitation of terminological and ontological resources via rule-based information extraction techniques delivering, delivering semantic annotation with respect to the CRM in the broader field of digital humanities.

9.1.4 Deliverables

The research work has produced a range of distinct deliverables. The most significant deliverable is the OPTIMA system, which contains the information extraction rules and the skosified gazetteer resources. Besides OPTIMA, an obvious deliverable is the semantic indices of grey-literature documents. The indices, as discussed previously, are expressed in the interoperable XML and RDF formats. The XML deliverable couples semantic annotations with content, while the RDF deliverable is decoupled from content, containing only the semantic annotation triples of documents. Both indices can be employed and manipulated further by semantic web applications, as seen by the example applications of the Andronikos portal and the STAR demonstrator.

Additional deliverables are the evaluation corpus (Gold Standard) and the proposed concepts for inclusion in the EH terminological resources. The Gold Standard corpus has been produced manually by archaeology domain experts for supporting the benchmarking aims of the evaluation phase. Although, it is not extremely extensive, it contains approximately 11000 words and 1000 semantic annotations. It is one of the few of its kind, if not the only Gold Standard available for archaeological grey-literature with respect to CIDOC-CRM. The evaluation corpus can be used to evaluate other CRM-based information extraction applications and potentially it could be useful in training ML applications. Additionally, a number of concepts were revealed during the development of the Gold Standard that were not included in the EH terminological resources. Such concepts were collected [Appendix D5] and have been made available to EH for further consideration and potential inclusion in the terminological resources.

9.2 Future Work

The semantic indexing results were encouraging since they demonstrated the capacity of rule-based Information Extraction techniques to deliver interoperable semantic abstractions (semantic annotations) with respect to the domain ontologies, CIDOC-CRM and CRM-EH. Such semantic indices were proved capable of aiding semantic aware information retrieval, cross searching and document inspection activities. However, the current semantic indexing system can be further improved and expanded. The following section presents

opportunities for future research and development of the thesis work.

A straightforward future development could be the expansion of the semantic indexing system (OPTIMA) over additional CRM-EH entities and relationships (and also CRM generally). The current system has been focused to identify four entity and four relationship types. The latest version of CRM-EH model consists of 125 classes, which describe different phases of the archaeological work from field research to post excavation analysis. The OPTIMA system has been targeted at the identification of entities and relationships that relate to archaeological find and context (including both individual contexts and larger groupings).

An uncomplicated expansion of the system would be to include classes that relate to archaeological fieldwork sampling and measurement information. The available sampling and measurement glossaries of English Heritage could be imported as gazetteer listings into the OPTIMA system for assisting identification of relevant information, while existing rules can be amended to identify additional CRM-EH entities. In addition, the current system can be expanded to other CRM-EH entities of interest, relating to site, stratigraphic information and other entities.

The current system has delivered the results of semantic indexing in two distinct interoperable formats: as XML annotations coupled with content and RDF triples of annotations decoupled from content. In particular, the RDF triples have followed a CRM-EH based representation of the extracted information tailored for the STAR demonstrator architecture. A future development could adapt other RDF representations such as those proposed for Linked Data.

The Linked Data project proposes a method for publishing data so it can be interlinked and accessed automatically by computers. The project builds on Web standard technologies, such as HTTP, XML/RDF and URI for enabling sharing and information querying from different sources. A potential future direction would be to express the semantic indexing of OPTIMA in Linked Data representations. The STELLAR tool (Semantic Technologies Enhancing Links and Linked data for Archaeological Resources) that enables data to be represented in standard RDF formats following the Linked-Data approach could be employed in such delivery (Tudhope et al. 2011).

A future development of the OPTIMA system could be directed towards the delivery of TEI (Text Encoding Initiative) representations of semantic annotations. TEI is a set of guidelines for encoding textual information in machine readable format. The initiative has been primarily active in the domains of humanities, social sciences and linguistics and it

has been widely used by libraries, museums and publishers to support online research and preservation. It is primarily a semantic representation containing about 500 textual components that are normally employed by an intellectual process of encoding. Intellectual CRM-based annotation has already been explored in relation to TEI (Ore and Eide 2009). Such intellectual indexing can be resource intensive both in terms of time and of human effort. Thus, a future development of the OPTIMA system could assist intellectual encoding with automated tools that overcome resource intensive barriers of larger corpus encoding.

The innovative integration of SKOS references and thesauri structure in GATE gazetteer entries, enabled JAPE rules to exploit thesauri relationships and to define information extraction patterns with respect to domain vocabulary. However, the parameterisation of GATE gazetteer entries was based on a bespoke method which used XML (SKOS enabled) versions of the terminological resources, which were translated into gazetteer entries using XSLT transformation rules. In addition, the definition of JAPE grammars responsible for exploiting terminological relationships was closely associated with the bespoke parameterisation of GATE gazetteer entries. A future development would be to expose thesauri relationships to JAPE rules via a GATE language resource similar to GATE OWLIM. A future "SKOSIM" plug-in for GATE could allow instantiation of any SKOS based thesauri in GATE, similarly to the instantiation of OWL ontologies and enable JAPE grammars to access thesauri vocabulary and relationships as achieved by the bespoke method of OPTIMA.

The evaluation of the OPTIMA system can be enhanced further. The current evaluation process had an end-user focus which delivered a Gold Standard definition based on the input of archaeology experts. The evaluation corpus originated from the same selection of 2460 OASIS documents which participated in the corpus analysis of syntactical patterns (section 6.3.3) and were processed by the pipeline. A future evaluation process could seek to benchmark the OPTIMA performance on OASIS archaeological grey literature documents which have not participated in the corpus analysis task. This way the performance of syntactical patterns of the Relation Extraction phase would be evaluated against new "unseen" cases. In addition, the generalisation of syntactical patterns can be tested in fields other than archaeology, aimed at identifying CRM events such production, move and beginning of existence. The potential of deep parsing could be explored by a future development for revealing patterns, which are not based only on syntactical aspects but also on writing practices and styles.

Future evaluation should also seek to evaluate the modules of Negation Detection and Word-sense Disambiguation. The current evaluation process, based on the end-user focus and constraints of the available resources, has benchmarked all individual processes of the pipeline under a single unified Gold Standard. Future evaluation of the pipeline could be focused on the definition of individual Gold Standards targeted at benchmarking the performance of specific modules of the information extraction pipeline, such as Negation Detection and Word-sense Disambiguation.

Information retrieval results originating from the semantic indices, as in the case of STAR Demonstrator, do not contain the same degree of certainty as the results originating from controlled fields in databases, since results from free text are potentially influenced by natural language ambiguities (i.e. false positive results). Currently the Demonstrator does not distinguish the provenance of information (data sources are identified but not the data type nor extraction method) and so treats equally information originating from natural language text and data extracted from datasets.

A future research direction that emerges from the experience gained adapting CIDOC-CRM in semantic annotation could aim to investigate the issue of modelling provenance and the degree of confidence of information. There is an inherent uncertainty in natural language information, which frequently makes statements about the world (that cannot always be assumed as facts), using grammatical moderators and making implicit reference to other statements. Being able to model contextual moderators and to assign a level of confidence could improve the information retrieval utility of semantic annotations. For example, different weights might be assigned to semantic annotations that correspond to different information extraction rules and pipeline stages, which might be utilised as an attribute of ranking retrieval results.

Finally, further investigation of the applicability of the current system within digital humanities beyond the domain of archaeology would be beneficial to reveal the potential for generalisation of the research work. A further development could investigate the adaptability of the current system via expanding current rules and domain vocabulary. A useful direction for both NER and RE tasks could be to focus on annotation of the CRM parent classes instead of the specialised CRM-EH currently used by OPTIMA.

References

APPAG (2003) 'The Current State of Archaeology in the United Kingdom', *First report of the All-Party Parliamentary Archaeology Group.* Available at*:* http://www.appag.org.uk/report/report.html (Accessed: 12 June 2012).

Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Dumais, S., Fuhr, N., Harman, D., Harper, D.J., Hiemstra, D., Hofmann, T., Hovy, E., Kraaij, W., Lafferty, J., Lavrenko, V., Lewis and Jones, D., Liddy, L., Manmatha, R., McCallum, A., Ponte, J., Prager, J., Radev, D., Resnik, P., Robertson, S., Rosenfeld, R., Roukos, S., Sanderson, M., Schwartz, R., Singhal, A., Smeaton, A., Turtle, H., Voorhees, E., Weischedel, R., Xu, J., and Zhai, C.X. (2003) 'Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval', *ACM SIGIR Forum*, 37(1), pp. 31–47.

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

Andronikos Web Portal (2012) Available at: http://hypermedia.research.glam.ac.uk/resources/andronikos/ (Accessed: 12 June 2012)

Aubin, S. and Hamon, T. (2006) 'Improving term extraction with terminological resources', in Salakoski, T., Ginter, F., Pyysalo, S. and Pahikkala, T. (eds), *Advances in Natural Language Processing: In Proceedings of the 5th International Conference on NLP*, FinTAL. Springer

Bach, N. and Badaskar, S. (2007) 'A survey on relation extraction' Technical report, *Language Technologies Institute, Carnegie Mellon University.*

Baeza-Yates, R., and Ribeiro-Neto, B., (1999) *Modern Information Retrieval*. Boston: Addison-Wesley Longman Publishing Co., Inc.

Bateman J., and Jeffrey S. (2011) 'What Matters about the Monument: reconstructing historical classification' *Internet Archaeology* (29) [Online]. Available at: http://intarch.ac.uk/journal/issue29/bateman_index.html (Accessed: 12 June 2012).

BBC (2007) *Miserable failure' links to Bush George W. Bush has been Google bombed.* Available at: http://news.bbc.co.uk/2/hi/americas/3298443.stm (Accessed: 12 June 2012).

Berners-Lee T., Hendler J., Lassila O. (2001) 'The Semantic Web', *Scientific American,* 284(5), pp. 28–37.

Berners-Lee T. (2007) 'Looking Back, Looking Forward: The process of Designing Things in a Very Large Space', *Lovelace Lecture, British Computer Society, London, UK.* [Video Lecture] Available at: http://www.w3.org/2007/Talks/0313-bcs-tbl/, http://mazine.ws/node/543 (Accessed: 12 June 2012).

Binding C., Tudhope D., May K. (2008) 'Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM', *In Proceedings (ECDL 2008) 12th European Conference on Research and Advanced Technology for Digital Libraries,* Aarhus, Denmark, 12-18 September, pp. 280–290.

Bontcheva K., Tablan V., Maynard D., Cunningham H. (2004) 'Evolving GATE to Meet New Challenges in Language Engineering', *Natural Language Engineering,* 10(3/4), pp. 349-373.

Bontcheva K., Cunningham H., Kiryakov A., Tablan V. (2006a) 'Semantic Annotation and Human Language Technology', in Davies, J., Studer, R. and Warren, P. (ed.) *Semantic Web Semantic Web Technology: Trends and Research in Ontology Based Systems.* Chichester*:* John Wiley and Sons Ltd.

Bontcheva K., Duke T., Glover N., Kings I. (2006b) 'Semantic Information Access', in Davies, J., Studer, R. and Warren, P. (ed.) *Semantic Web Semantic Web Technology: Trends and Research in Ontology Based Systems.* Chichester*:* John Wiley and Sons Ltd.

Brants T. (2000) 'Inter-annotator agreement for a German newspaper corpus', *In Proceedings (LREC 2000) 2nd International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May – 2 June.

Brill E. (1995) 'Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging', *Computational Linguists*, 21(4), pp. 543–565.

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

Byrne, K. (2009) *Populating the Semantic Web—Combining Text and Relational Databases as RDF Graphs*. PhD thesis, University of Edinburgh.

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

Chapman W.W., Bridewell W., Hanbury P., Cooper G.F. and Buchanan B.G. (2001) 'A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries', J*ournal of Biomedical Informatics*, 34(5), pp.301–310.

Ciravegna, F. and Lavelli, A. (2004) 'LearningPinocchio: adaptive information extraction for real world applications', *Natural Language Engineering*, 10(02), pp. 145–165.

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

Cowie, J., and Lehnert, W. (1996) 'Information extraction', *Communications ACM*, 39(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 Technical Paper.* Available at: http://www.cidoc-crm.org/technical_papers.html (Accessed: 12 June 2012).

Crofts, N., Doerr, M., Gill, T., Stead, S., and Stiff, M. (2009) 'Definition of the CIDOC Conceptual Reference Model', *FORTH Greece*. Available at http://cidoc.ics.forth.gr/docs/cidoc_crm_version_5.0.1_Mar09.pdf (Accessed: 12 June 2012).

Cunningham H., Wilks Y., and Gaizauskas R. (1996) 'GATE-a General Architecture for Text Engineering', *Computers and the Humanities*, pp. 1057–1060.

Cunningham, H. (1999) *'A definition and short history of Language Engineering'*. Natural Language Engineering, 5(1), 36, pp. 223–254.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan V. (2002) 'GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications', *In Proceedings (ACL2002) 40th Anniversary Meeting of the Association for Computational Linguistics*, Stroudsburg, Philadelphia, July 2012. Available at http://gate.ac.uk/sale/acl02/acl-main.pdf (Accessed: 12 June 2012).

Cunningham, H., Scott, D. (2004) 'Software Architecture for Language Engineering', *Natural Language Engineering*, 10(3-4), pp. 205–209.

Cunningham, H. (2005) 'Information Extraction, Automatic' in Brown, K. (ed.) *Encyclopedia of Language and Linguistics.* Cambridge: Elsevier.

Dale, R., Moisi, H., and Harold, S. (2000) *Handbook of Natural Language Processing*. New York: Marcel Dekker.

Debachere M.,C. (1995) 'Problems in Obtaining Grey Literature', *Journal of the International Federation of Library Associations and Institutions*, 21(2), pp. 94-106.

Department of the Environment (2010) *Planning Policy Statement 5: Planning for the Historic Environment (PPS5).* [Online]. Available at :

http://www.communities.gov.uk/publications/planningandbuilding/pps5 (Accessed: 12 June 2012).

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004) 'The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation', *In Proceedings (LREC 2004) 4th International Conference on Language Resources and Evaluation,* Lisbon, Portugal , 26-28 May, pp. 837–840.

Doerr, M., (2003) 'The CIDOC CRM – An ontological approach to semantic interoperability of metadata', *AI Magazine,* 24(3), pp. 75–92.

Doerr M., Gradmann S., Hennicke S., Isaac A., Meghini C., Sompel van de H.(2010) '*The Europeana Data Model (EDM)*', *In World Library and Information Congress: 76th IFLA General Conference and Assembly*, Gothenberg, Sweden, 10-15 August.

Drożdżyński, W., Krieger, H.-U., Piskorski, J., Schäfer, U., and Xu, F. (2004) 'Shallow Processing with Unification and Typed Feature Structures – Foundations and Applications', *Künstliche Intelligenz*, 1(1), pp. 17-23.

English Heritage Recording Manual (2006), English Heritage internal document

English Heritage National Monuments Record Thesauri (2007). Available at http://thesaurus.english-heritage.org.uk/ (Accessed: 12 June 2012).

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 [Online]. Available at: http://intarch.ac.uk/journal/issue17/falkingham_index.html (Accessed: 12 June 2012).

Farace, D.J. (1997) 'Grey Literature and Publishing', *Publishing Research Quarterly,* 13(2), pp. 3-4.

Feldman, R., Aumann, Y., Finkelstein-Landau, M., Hurvitz, E., Regev, Y., and Yaroshevich, A. (2002) 'A Comparative Study of Information Extraction Strategies', *In Proceedings (CICLing-2002) Third International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico city, Mexico, 17-23 February.

Ferrucci D., and Lally A. (2004) 'Building an example application with the Unstructured Information Management Architecture', *IBM Systems Journal* 43(3) pp. 455-475

Fernandez, M.P., and Garcia-Serrano, A.M. (2000*)* 'The role of knowledge-based technology in language applications development', *Expert Systems with Applications*, 19(1), pp. 31–44.

Friedman, C., Kra, P., Yu, H., Krauthammer, M., and Rzhetsky, A. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 17, S74–S82

Gaizauskas R. and Wilks Y. (1998) 'Information extraction: beyond document retrieval', *Journal of Documentation,* 54(1), pp. 70–105

Gazdar, G. (1996*)* 'Paradigm merger in natural language processing' in Milner, R. and Wand, I. (ed.), *Computing Tomorrow: Future Research Directions in Computer Science*. New York: Cambridge University Press, pp. 88-109.

Golub, K., Hamon, T. and Ardö, A. (2007) 'Automated classification of textual documents based on a controlled vocabulary in engineering', *Knowledge Organization*, 34(4), pp. 247-263

Golub, K. (2006) 'Automated subject classification of textual web documents', *Journal of Documentation*, 62(3), pp. 350-371

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: 12 June 2012).

Grishman, R. and Sundheim B. (1996) 'Message Understanding Conference-6: a brief history', *Proceeding (COLING 1996) 16th International Conference on Computational*

Linguistics , Copenhagen, Denmark, 5-9 August.

Grover C., Givon S., Tobin R., and Ball J. (2008) 'Named entity recognition for digitised historical texts', *In Proceedings (LREC 2008) 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 28-30 May.

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

Guarino, N., Oberle, D., and Staab, S. (2009) 'What is an Ontology?', in Staab, S. and Studer, R. (ed.) *Handbook on Ontologies - Second Edition.* Berlin: Springer Verlag, pp. 1-17.

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

Hargrave S (2007) 'Is Google a Grinch or a good guy?', *The Guardian*, 13 December [Online]. Available at: http://www.guardian.co.uk/technology/2007/dec/13/internet.google (Accessed: 12 June 2012).

Hepple, M. (2000) 'Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers', *In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, Hong Kong, 1-8 October.

Hobbs, J.R. (1993) 'The Generic Information Extraction System', *In Proceedings (MUC-5) 5th Message Understanding Conference*. Baltimore, Maryland, 25-27 August.

Hobbs, J. R., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M., and Tyson, M. (1993) 'FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text', *In Proceedings (IJCAI 1993) 13th International Joint Conference on Artificial Intelligence*, Chambery, France, 28 August – 3 September.

Horn L.R. (1989) *A Natural History of Negation*. Chicago : University of Chicago Press.

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H., and Wilks, Y. (1998) 'University Of Sheffield: Description Of The Lasie-II System As Used For MUC-7', *In Proceedings (MUC-7) 7th Message Understanding Conference.* [Online]. Available at: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html (Accessed: 12 June 2012)

Isaac A., Summers E. (2009) *SKOS Simple Knowledge Organization System Primer. [*Online]. Available at: http://www.w3.org/TR/skos-primer (Accessed: 12 June 2012)

Jansen B., and Spink A. (2006) 'How are we searching the world wide web? A comparison of nine search engine transaction logs', *Information Processing and Management,* 42(1), pp. 248-263

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"*, pp.2507–2519

Jones K.S. (1999) 'What is the role of NLP in text retrieval?', in Strzalkowski, T. (ed) N*atural language information retrieval.* New York: Kluwer, pp.1-25.

Jurafsky, D., and Martin, J.H. (2000) *Speech and Language Processing.* New Jersey: Prentice Hall

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*, 2(1), pp. 49–79

Kurtz D., Parker G., Shotton D., Klyne G., Schroff F., Zisserman A., and Wilks Y (2009) CLAROS - Bringing Classical Art to a Global Public, *In Proceedings 5th IEEE*

International

Conference on e-Science . Oxford, UK, 9-11 December

Lee, B., Hendler, J. and Lassila, O. (2001) 'The Semantic Web', *Scientific American*, 284(5), pp. 28-37.

Lewis, D, and Jones K.S., (1996*)* 'Natural language processing for information retrieval', *Communications of the ACM,* 39(1), pp.92–101.

Li, W. (2002) 'Zipf's Law Everywhere', *Gottometrics*, 5, pp.14–21

Li Y., Bontcheva K. (2007) 'Hierarchical, Perceptron-like Learning for Ontology Based Information Extraction', *In Proceedings of the 16th International World Wide Web Conference*, Banff, Canada, 8-12 May

Liakata M., Q C., and Soldatova S. (2009) 'Semantic annotation of papers: Interface & enrichment tool (sapient)', *In Proceedings of BioNLP 2009*, Boulder, Colorado, 4-5 June.

Liddy, E.D. (2003) 'Natural Language Processing' in Drake, M. (ed.), *Encyclopedia of Library and Information Science*, London: Taylor and Francis, pp. 2126–2136

Lin, D. (1995) 'University of Manitoba: description of the PIE system used for MUC-6', *In Proceedings (MUC 6) 6th Message Understanding Conference*, Columbia, Maryland, 6-8 November.

Linguistic Data Consortium (LDC) (2004) '*Automatic Content Extraction'* (ACE). [Online]. Available at: http://projects.ldc.upenn.edu/ace/ (Accessed: 12 June 2012).

Linked Data [Online]. Available at: http://linkeddata.org/ (Accessed: 12 June 2012).

Maynard D., Funk A. (2011) 'Automatic detection of political opinions in tweets', *In Proceedings of (MSM 2011), Making Sense of Microposts Workshop at 8th Extended Semantic Web Conference (ESWC 2011)*. Heraklion, Greece, 29 May - 2 June.

Maynard D., Yankova M., Kourakis A., and Kokossis A. (2005) 'Ontology-based information extraction for market monitoring and technology watch', *In Proceedings (ESWC 2005) Workshop "End User Apects of the Semantic Web*, Heraklion, Crete, 29 May – 1 June

Maynard D., Peters W., Li Y. (2006) 'Metrics for Evaluation of Ontology-based Information Extraction', *In Proceedings WWW Conference 2006, Workshop on "Evaluation of Ontologies for the Web*", Edinburgh, Scotland, 23-26 May.

McGilvray, J. (1999) *'Chomsky: Language Mind and Politics'*. Cambridge: Polity Press

McQuire A.R., Eastman C.M. (1998) 'The ambiguity of negation in natural language queries to information retrieval systems', J*ournal of the American Society for Information Science*, 49(8), pp. 686–692

Medelyan, O., and Witten, I. (2006) 'Thesaurus based automatic keyphrase indexing', *In Proceedings (JCDL 2006), Joint Conference on Digital Libraries.* Chapel Hill, NC, 11-15 June.

Moens, M.F., (2006) *Information Extraction Algorithms and Prospects' in a Retrieval Context*, Dordrecht: Springer

MUC-7 (2001) 7th Message Understanding Conference. [Online]. Available at: http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html (Accessed: 12 June 2012)

Mutalik P.G., Deshpande A., Nadkarni P.M.(2001) 'Use of General-purpose Negation Detection to Augment Concept Indexing of Medical Documents: A Quantitative Study Using the UMLS'. *Journal of the American Medical Informatics Association 2001,* 8(6), pp. 598-609

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

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

Newman M (2005) 'Power laws, Pareto distributions and Zipf's law', *Contemporary*

Physics, (46) pp. 323–351

Nilsson, N. (2005) *Introduction to Machine Learning*. [Online]. Nils J Nilson publications.

Available at: http://robotics.stanford.edu/people/nilsson/mlbook.html (Accessed: 12 June 2012)

OPTIMA (2012) Project Resources Available at: http://sourceforge.net/projects/optimacidoc/ (Accessed: 12 June 2012)

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

Peters W., Aswani N., Bontcheva K., and Cunningham H. (2005) 'Quantitative Evaluation Tools and Corpora v1. Technical report', *SEKT project deliverable D2.5.1*

Popescu A.M., Etzioni O. (2005) 'Extracting product features and opinions from reviews', *In Proceedings(HLT/EMNLP-2005) Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing* , pp.339–346, Vancouver, Canada, 6-8 October.

Popov B, Kiryakov A, Ognyanoff D, Manov D, Kirilov A. (2004) 'KIM ; a semantic platform for information extraction and retrieval', *Natural Language Engineering*, 10(3-4), pp. 375–392

Ramshaw L., and Marcus M. (1995). Text Chunking Using Transformation-Based Learning. *In Proceedings of the Third ACL Workshop on Very Large Corpora*, Dublin, Ireland

Reeve, L., and Han H. (2005). 'Survey of semantic annotation platforms', *In Proceedings 20th Annual ACM Symposium on Applied Computing*, Santa Fe, New Mexico, 13-17 March

Resnik P. (1997) 'Selectional preferences and sense disambiguation', *In Proceedings ACL SIGLEX Workshop on Tagging Text with Lexical Semantics*, Columbia, Washington, D.C, 4-5 April.

Richards, J. and Hardman, C. (2008) 'Stepping back from the trench edge', in Greengrass, M. and Hughes, L. (ed.) *The Virtual Representation of the Past*, Farnham England: Ashgate.

Rokach, L., Romano, R., Maimon, O. (2008) 'Negation recognition in medical narrative reports', *Information Retrieval*, 11(6), pp. 499–538

Ross, S. (2003) 'Position Paper in Towards a Semantic Web for Heritage Resources', *Digital Culture (DigiCULT) Thematic Issue* 3, pp. 7–11.

Sanderson M (2000) 'Retrieving with Good Sense', *Information Retrieval*, 2(1), pp. 49–69

Savary, A., Waszczuk, J., and Przepiórkowski, A. (2010) 'Towards the Annotation of Named Entities in the National Corpus of Polish', *In Proceedings (LREC'10) Fourth International Conference on Language Resources and Evaluation*, Valletta, Malta, 13-17 May.

Shaw, R., Troncy, R., and Hardman, L., (2009) 'LODE: linking open descriptions of events', 4*th Asian Semantic Web Conference (ASWC'09)* Shanghai, China, 6-9 December

Smeaton, A.F. (1997) 'Information Retrieval: Still Butting Heads with Natural Language Processing?' in Pazienza M.T. (ed.) *International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology (SCIE '97),* London, Springer-Verlag

Smith D. A., Rydberg-Cox J.A., Crane G.R. (2000) 'The Perseus Project: a Digital Library for the Humanities', *Literary and Linguistic Computing*, 15(1), pp. 15-25

Soderland, S., Fisher, D., Aseltine, J., and Lehnert, W. (1995) 'CRYSTAL inducing a conceptual dictionary', *In Proceedings (IJCAI-95) 14th International joint conference on Artificial Intelligence* (2). Montreal, Canada, 20-25 August.

Soderland, S., Fisher, D., and Lehnert, W., (1997) 'Automatically Learned vs. Hand-crafted Text Analysis Rules', *CIIR Technical Report T44*

Soldatova L., Batchelor C., Liakata M., Fielding H., Lewis S., and King R. (2007) 'ART: An ontology based tool for the translation of papers into Semantic Web format', *In Proceedings of the Bio-Ontologies SIG Workshop 2007,* Vienna, Austria, 20 July.

Text Encoding Initiative .[Online]. Available at: http://*www.tei-c.org (*Accessed: 12 June 2012)

Ting Wang, Yaoyong Li, Kalina Bontcheva, Hamish Cunningham and Ji Wang. (2006) 'Automatic Extraction of Hierarchical Relations from Text', *In Proceedings of the Third European Semantic Web Conference (ESWC 2006),* Budva, Montenegro, 11-14 June.

Tsujii J., Ananiadou S. (2005) 'Thesaurus or logical ontology, which one do we need for text mining?', *Language Resources and Evaluation*, 39(1), pp. 77-90.

Tudhope D., Koch T., Heery R. (2006) 'Terminology Services and Technology: JISC State of the art review', JISC. [Online]. Available from: http://www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_technology_review_sep_06.pdf (Accessed: 12 June 2012).

Tudhope D., Binding C., May K. (2008) 'Semantic interoperability issues from a case study in archaeology', in Kollias, S. and Cousins, J. (ed), *Semantic Interoperability in the European Digital Library, Proceedings of the First International Workshop SIEDL 2008*, associated with 5th European Semantic Web Conference, Tenerife, Spain 1-5 June.

Tudhope D, May K, Binding C, Vlachidis A. 2011. 'Connecting archaeological data and grey literature via semantic cross search', *Internet Archaeology*, (30) [Online]. Available at: http://intarch.ac.uk/journal/issue30/tudhope_index.html/ (Accessed: 12 June 2012).

Tudhope D, Binding C, Jeffrey S, May K, Vlachidis A. 2011. 'A STELLAR Role for Knowledge Organization Systems in Digital Archaeology', in Greenberg, J. (ed.) *Special Section on Knowledge Organization Innovation: Design and Frameworks*, 37(4), pp.15-18.

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*, 4(1), pp. 14–28.

US NIST (1996) 'The TIPSTER architecture', *US National Institute for Standards and Technology (NIST),* [Online]. Available at: http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/ (Accessed: 12 June 2012).

US NIST (2003). 'The ACE 2003 Evaluation Plan', *US National Institute for Standards and Technology (NIST),* [Online]. Available at: http://www.itl.nist.gov/iad/mig/tests/ace/2003/ (Accessed 12 June 2012).

US NIST (2004). 'The ACE 2004 Evaluation Plan', *US National Institute for Standards and Technology (NIST),* [Online]. Available at: http://www.itl.nist.gov/iad/mig/tests/ace/2004/ (Accessed 12 June 2012).

US NIST (2005). 'The ACE 2004 Evaluation Plan', *US National Institute for Standards and Technology (NIST),* [Online]. Available at: http://www.itl.nist.gov/iad/mig/tests/ace/ace05/ (Accessed 12 June 2012).

US NIST (2007). 'The ACE 2004 Evaluation Plan', *US National Institute for Standards and Technology (NIST),* [Online]. Available at: http://www.itl.nist.gov/iad/mig/tests/ace/ace07/ (Accessed 12 June 2012).

Voorhees E. (1993) 'Using WordNet to Disambiguate Word Senses for Text Retrieval', *In Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.171-180. Pittsburgh, 27 June - 01 July.

Voorhees, E.M. (1999) 'Natural Language Processing and Information Retrieval', *Information Extraction: Towards Scalable, Adaptable Systems, Lecture notes in Artificial Intelligence,* 1174, pp. 32-48, Springer-Verlag: London.

Westbury, C. (2000) 'Just say no: The evolutionary and developmental significance of negation in behavior and natural language', *3rd Conference The Evolution of Language,* Paris, France 3-6 April.

Weintraub, I. 2000 'The Role of Grey Literature in the Sciences', *Brooklyn College Library,* [Online]. Available at: http://library.brooklyn.cuny.edu/access/greyliter.htm. (Accessed 12 June 2012).

Weizenbaum, J. (1966) 'ELIZA – A computer program for the study of natural language communication between man and machine', *Communications of the ACM*, 9(1), pp. 36-45.

Wilks Y., and Brewster C. (2009) 'Natural Language Processing as a Foundation of the Semantic Web', *Foundations and Trends in Web Science*, 1(3-4), pp. 199–327.

Zelenko D., Aone C., Richardella A. (2003) 'Kernel methods for relation extraction', *Journal of Machine Learning Research,* (3), pp. 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 pp.301–315, Springer-Verlag: London.

Zipf G.K. (1935) *The Psycho-biology of language: An Introduction to Dynamic Biology*, second edition (1965), Cambridge: MIT Press.

Appendix A

Terminology Resources Listings

A1. Physical Object and Material Overlapping Terms

1. MDA Object Type thesaurus and Main Building Materials thesaurus
"Brass", "Ceramic", "Clinker", "Cork", "Daub", "Marble", "Mosaic", "Pantile", "Paper", "Rubber", "Shingle", "Slag", "Slate", "Terracotta", "Tessera", "Textile", "Tile"

2. MDA Object Type thesaurus and Box Index Form: Material glossary
"Animal Bone", "Human Bone", "Slag"

3. MDA Object Type thesaurus and Bulk Finds Material List glossary
"Animal Bone", "Brick", "Human Bone", "Slag", "Slate", "Tile"

4. Main Building Materials thesaurus and Small Finds Form glossary
"Brick", "Flint", "Glass", "Mortar", "Plaster", "Shell", "Slag", "Slate", "Tile", "Wood"

5. Small Finds Form glossary and Box Index Form: Material glossary
"Animal bone", "Bone", "Fired clay", "Flint", "Glass", "Human bone", "Plaster", "Pottery", "Shell", "Slag", "Stone", "Wood"

6. Small Finds Form glossary and Bulk Finds Material List glossary
"Fired clay", "Flint", "Glass", "Shell", "Wood"

A2. Supplementary Gazetteer Listings

Time Prefix

Past, Past-, Past -, past, past-, past -, Pre, Pre-, Pre -, pre, pre-, pre -, mid, mid-, mid -, Mid, Mid-, Mid -, Early, Early-, Early -, early, early-, early -, Earlier, Earlier-, Earlier -, earlier, earlier-, earlier -, Later, Later-, Later -, later, later-, later -, Late, Late-, Late -, late, late-, late -, Middle, Middle-, Middle -, middle, middle-, middle -, post, post-, post -, Post, Post-, Post -, latter half, earlier half, latter half of the, earlier half of the, first half of the, second half of the, first half, second half, midearly, mid-early, mid- early, mid -early, mid - early, midlate, mid-late, mid- late, mid -late, mid – late

Date Suffix

Years, Year, Months, Month, Mnths, Mnth, Milleniums, Millenium, Millenia, Decades, Decade, Century, Centuries, Periods, Period, yrs, yr, years, year, months, month, mnths, mnth, milleniums, millenium, millenia, decades, decade, century, centuries, periods, period, BC, bc, AC, ac

Ordinal

1st, First, first, 1st2nd, 2nd, Second, second, 2nd3rd, 3rd, Third, third, 3rd4th, 4th, fourth, Fourth, 4th5th, 5th, Fifth, fifth, 5th6th, 6th, Sixth, sixth, 6th7th, 7th, Seventh, seventh, 7th8th, 8th, Eighth, eighth, 8th9th, 9th, Ninth, ninth, 9th10th, 10th, Tenth, tenth, 10th11th, 11th, Eleventh, eleventh, 11th12h, 12th, Twelfth, twelfth, 12th13th, 13th, Thirteenth, thirteenth, 13th14th, 14th, Fourteenth, fourteenth, 14th15th, 15th, Fifteenth, fifteenth, 15th16th, 16th, Sixteenth, sixteenth, 16th17th, 17th, Seventeenth, seventeenth, 17th18th, 18th, Eighteenth, eighteenth, 18th19, 19th, Nineteenth, nineteenth, 19th20th, 20th, Twentieth, twentieth

A3. Frequent Noun Phrase List

archeological trial, Baked, beamish, braunstone, brickearth, brickearth deposit, brickwork, brown clayey loam, cattle horncore, circular posthole, clayey loam, clayey silt patches, clayey soil, collapsed superstructure, cologne/frechen stoneware, Corn, corn cockle, corn gromwell, cornwall, current ploughsoil, dark brown clayey loam, dark greyish brown clayey loam, datestone, Drainage, drainage channel, drainage function, drainage pipe, drainage service, embanked, embanked material, embanked subsoil, english stoneware, flintwork, flod enbankment, floorboard, frechen stoneware, friable mid-brown sandy-clayey silt, Furrow, furrows, furrow cultivation, furrow remains, furrow ridge, furrow type, german stoneware, golden age, goldsmith, hole tree, holystone, holystone history group, Hoof, horse-bone, Human, human activityies, human habitation, human occupation, human presence, human skeletal, human skeletal remains, humanities data service, humberstone, infrastructure, Intrusion, large posthole, large tree, later intrusion, Leaded, Lenses, leytonstone, light greyish brown clayey silt deposit, london stoneware, matt claydon, Metallicled, metalwork, metalwork style, metalworking debris, mid greyish brown clayey silt deposit, mid-greyish brown clayey-sandy silt, natural brickearth, no floorboard, occasional posthole, old tree, oval posthole, past human societies, pinkish brown clay, Plastered, plasterwork, ploughed, ploughmark, ploughsoil, postholes, posthole group Potteryies, raeren stoneware, ravenstone, rectangular posthole, redeposited brickearth, ridge-and-furrow, salt-glazed stoneware, segment tree, shallow posthole, single posthole, small posthole, snibstone, sparkenstone estate, staffordshire salt-glazed stoneware, stonebond limited, stoneware, stonework, superstructure, Tree, tree bole, tree canopy, tree clearance, tree cover, tree hole, tree preservation order, tree roots, Trial, trial excavation, trial holes, undated posthole, Walls/ed, west humberstone, westerwald stoneware, wheelchair, wheelhouse, Wooden

A4. Added Synonyms in Gazetteers

Human skeletal	Synonym	Human Bone ehg019.10
Horse bone	Synonym	Animal Bone ehg019.2
Drainage	Synonym	Drain (ehg003.24)
Cattle horncore	Synonym	Horn (ehg026#14)
Human skeletal remains	Synonym	Human Bone ehg019.10
ploughmark	Synonym	Plough-mark ehg003.60
Potteryies	Synonym	Pottery ehg027.2
Furrow	Synonym	Ridge and Furrow EHT1 68628
Furrow ridge	Synonym	Ridge and Furrow EHT1 68628
Ridge-and-furrow	Synonym	Ridge and Furrow EHT1 68628

A5. Enhancements for the term "cesspit: fill" (example case)

cesspit: fill@skosConcept=ehg003.57
cesspit fill@skosConcept=ehg003.57
fill of cesspit@skosConcept=ehg003.57
cess-pit: fill@skosConcept=ehg003.57
cess-pit fill@skosConcept=ehg003.57
fill of cess-pit@skosConcept=ehg003.57
cess- pit: fill@skosConcept=ehg003.57
cess- pit fill@skosConcept=ehg003.57
fill of cess- pit@skosConcept=ehg003.57
cess -pit: fill@skosConcept=ehg003.57
cess -pit fill@skosConcept=ehg003.57
fill of cess -pit@skosConcept=ehg003.57
cess - pit: fill@skosConcept=ehg003.57
cess - pit fill@skosConcept=ehg003.57
fill of cess - pit@skosConcept=ehg003.57
cesspits: fill@skosConcept=ehg003.57
cesspits fill@skosConcept=ehg003.57
fill of a cesspits@skosConcept=ehg003.57
cess-pits: fill@skosConcept=ehg003.57
cess-pits fill@skosConcept=ehg003.57
fill of cess-pits@skosConcept=ehg003.57
cess- pits: fill@skosConcept=ehg003.57
cess- pits fill@skosConcept=ehg003.57
fill of cess-pits@skosConcept=ehg003.57
cess -pits: fill@skosConcept=ehg003.57
cess -pits fill@skosConcept=ehg003.57
fill of a cess - pits@skosConcept=ehg003.57
cess - pits: fill@skosConcept=ehg003.57
cess - pits fill@skosConcept=ehg003.57
fill of cess - pits@skosConcept=ehg003.57
cesspit: fills@skosConcept=ehg003.57
cesspit fills@skosConcept=ehg003.57
fills of cesspit@skosConcept=ehg003.57
cess-pit: fills@skosConcept=ehg003.57
cess-pit fills@skosConcept=ehg003.57
fills of cess-pit@skosConcept=ehg003.57
cess- pit: fills@skosConcept=ehg003.57
cess- pit fills@skosConcept=ehg003.57
fills of cess-pit@skosConcept=ehg003.57
cess -pit: fills@skosConcept=ehg003.57
cess -pit fills@skosConcept=ehg003.57
fills of cess - pit@skosConcept=ehg003.57
cess - pit: fills@skosConcept=ehg003.57
cess - pit fills@skosConcept=ehg003.57
fills of cess - pit@skosConcept=ehg003.57
cesspits: fills@skosConcept=ehg003.57
cesspits fills@skosConcept=ehg003.57
fills of a cesspits@skosConcept=ehg003.57
cess-pits: fills@skosConcept=ehg003.57
cess-pits fills@skosConcept=ehg003.57
fills of cess-pits@skosConcept=ehg003.57
cess- pits: fills@skosConcept=ehg003.57
cess- pits fills@skosConcept=ehg003.57
fills of cess-pits@skosConcept=ehg003.57
cess -pits: fills@skosConcept=ehg003.57
cess -pits fills@skosConcept=ehg003.57
fills of a cess - pits@skosConcept=ehg003.57
cess - pits: fills@skosConcept=ehg003.57
cess - pits fills@skosConcept=ehg003.57
fills of cess - pits@skosConcept=ehg003.57

A6. Verb Vocabulary (Context Find Deposition Event)

accompany, accompanies, accompanied, accompanying, align, aligns, aligned, aligning, allocate, allocates, allocated, allocating, allow, allows, allowed, allowing, appear, appears, appeared, appearing, associate, associates, associated, associating, bases, based, build, builds, built, bury, buries, buried, carry, carries, carried, collect, collects, collected, come from, came from, compose, composes, composed, composing, comprise, comprises, comprised, comprising, consist, consists, consisted, consisting, construct, constructs, constructed, constructing, contain, contains, contained, containing, cover, covers, covered, covering, deposited, depositing, derive, derives, derived, deriving, discard, discards, discarded, discarding, discover, discovers, discovered, discovering, ditching, ditched, encounter, encounters, encountered, encountering, examine, examines, examined, examining, excavate, excavates, excavated, excavating, exist, exists, existed, existing, expose, exposes, exposed, exposing, filled, filling, found, identify, identifies, identified, identifying, include, includes, included, including, incorporate, incorporates, incorporated, incorporating, indicate, indicates, indicated, indicating, involve, involves, involved, involving, lay, lays, laid, laying, locate, locates, located, locating, lying, measure, measures, measured, measuring, note, notes, noted, noting, observe, observes, observed, observing, overburden, overburdens, overburdened, overburdening, overlay, overlie, overlies, overlain, overlying, places, placed, placing, present, presents, presented, presenting, preserve, preserves, preserved, preserving, prevalent, produce, produces, produced, producing, project, projects, projected, projecting, propose, proposes, proposed, proposing, quantify, quantifies, quantified, quantifying, record, records, recorded, recording, recover, recovers, recovered, recovering, relate, relates, related, relating, remove, removes, removed, removing, represent, represents, represented, representing, retain, retains, retained, retaining, retrieve, retrieves, retrieved, retrieving, reveal, reveals, revealed, revealing, samples, sampled, sampling, scatter, scatters, scattered, scattering, see, sees, seen, seeing, show, shows, showed, showing, shaped, shaping, situate, situated, situating, suggest, suggests, suggested, suggesting, survive, survives, survived, surviving, uncover, uncovers, uncovered, uncovering, underlay, underlays, underlaid, underlying, viewed, viewing

Appendix B

Negation Detection Listings

B1. Pre-negation list

no, none, non, not, do not, don't, does not, doesn't, did not, didn't, was not, wasn't, were not, weren't, have not, haven't, cannot, can't, could not, couldn't, would not, wouldn't, lack of, absence of, fails to reveal, failed to, with no, without, free of, negative for, to exclude, excluding, unremarkable for, rules out, rather than, preclude

B2. Post-negation list

unlikely, improbable, was ruled out, is ruled out, are ruled out, have been ruled out, has been ruled out, is excluded, are excluded, was excluded, were excluded, has been excluded, have been excluded, is precluded, are precluded, was precluded, were precluded, has been precluded, have been precluded, are unknown, was unknown, were unknown, has been unknown, have been unknown

B3. Negation Verbs list

appear, appeared, encountered, established, examined, excavated, exposed, found, identified, included, inspected, investigated, located, marked, mentioned, observed, obtained, produced, raised, reached, recorded, recovered, resolved, retained, reveal, seen, shown, survived, tested, traced, verified

B4. Stopclause-negation list

and, however, nevertheless, withal, still, yet, all the same, even so, nonetheless, notwithstanding, at the same time, but, although, than, though, therefore, hence, thence, therefrom, thereof, thus, so, so far, thus far, up to now, hitherto, heretofore, as yet, til now, until now, insofar, in so far, so far, to that extent, to that degree, furthermore, moreover, what is more

Appendix C

Relation Extraction Event Spans and Extraction Patterns

C1. Analysis of the "ObjectTime" Event Spans

Number of Tokens	Occurrences	Log Num. of Tokens	Log Occurrences
2	1631	0.3	3.21
3	1224	0.48	3.09
4	1012	0.6	3.01
5	922	0.7	2.96
6	973	0.78	2.99
7	999	0.85	3
8	920	0.9	2.96
9	831	0.95	2.92
10	1021	1	3.01
11	678	1.04	2.83
12	611	1.08	2.79
13	511	1.11	2.71
14	487	1.15	2.69
15	414	1.18	2.62
16	352	1.2	2.55
17	329	1.23	2.52
18	284	1.26	2.45
19	274	1.28	2.44
20	209	1.3	2.32
21	200	1.32	2.3
22	177	1.34	2.25
23	149	1.36	2.17
24	144	1.38	2.16
25	132	1.4	2.12
26	117	1.41	2.07
27	91	1.43	1.96
28	95	1.45	1.98
29	67	1.46	1.83
30	54	1.48	1.81

Table Appx.1: ObjectTime pairs of span size in number of tokens in actual and logarithmic values

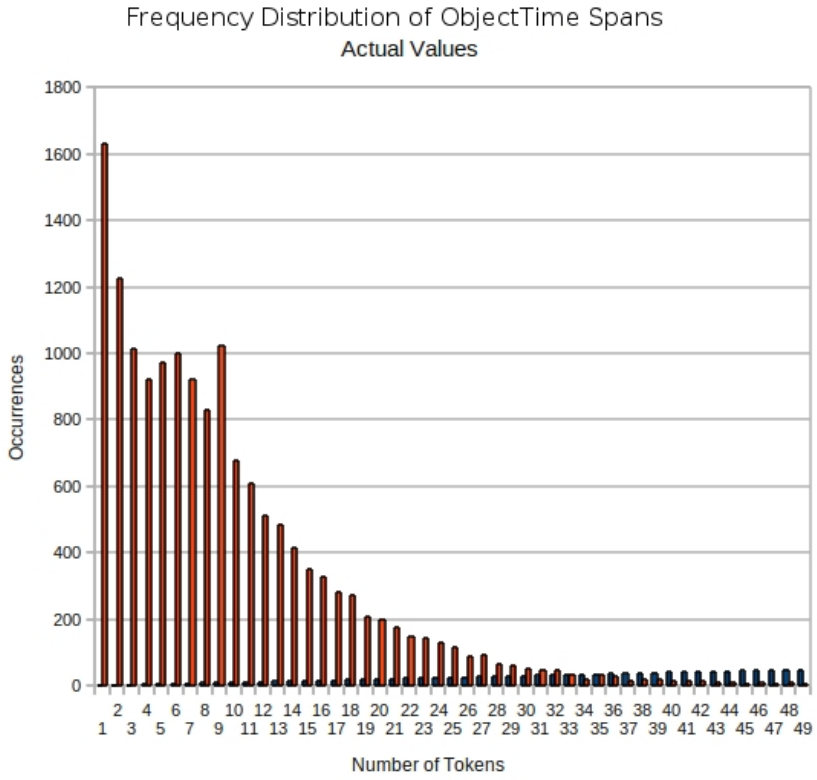


Figure Appx.1: ObjectTime Spans, distribution actual values

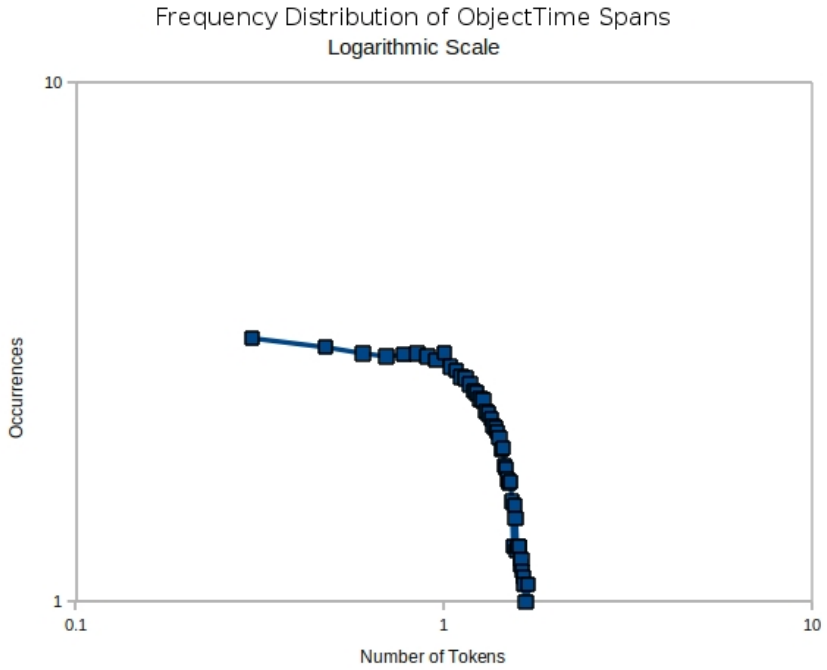


Figure Appx.2: ObjectTime Spans, distribution on the logarithmic

Span Length	2	3	4	5	6	7	8	9	10
Unique Patterns	24	139	373	620	757	809	763	762	677

Table Appx.2: ObjectTime Spans, unique patterns for 9 different span lengths

C2. Analysis of the "PlaceTime" Event Spans

Number of Tokens	Occurrences	Log Num. of Tokens	Log Occurrences
2	3391	0.3	3.53
3	3469	0.48	3.54
4	3003	0.6	3.48
5	3231	0.7	3.51
6	2881	0.78	3.46
7	2776	0.85	3.44
8	2714	0.9	3.43
9	2559	0.95	3.41
10	2105	1	3.32
11	1794	1.04	3.25
12	1556	1.08	3.19
13	1410	1.11	3.15
14	1303	1.15	3.11
15	981	1.18	2.99
16	920	1.2	2.96
17	826	1.23	2.92
18	710	1.26	2.85
19	562	1.28	2.75
20	518	1.3	2.71
21	400	1.32	2.6
22	396	1.34	2.6
23	348	1.36	2.54
24	306	1.38	2.49
25	273	1.4	2.44
26	238	1.41	2.38
27	187	1.43	2.27
28	177	1.45	2.25
29	145	1.46	2.16
30	134	1.48	2.13

Table Appx.3: PlaceTime pairs of span size in number of tokens in actual and logarithmic values

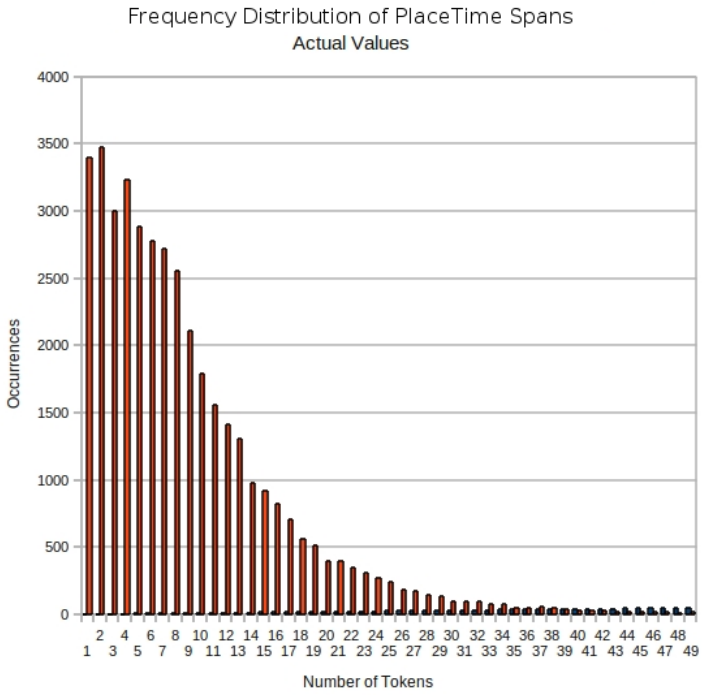


Figure Appx.3: PlaceTime spans, distribution actual values

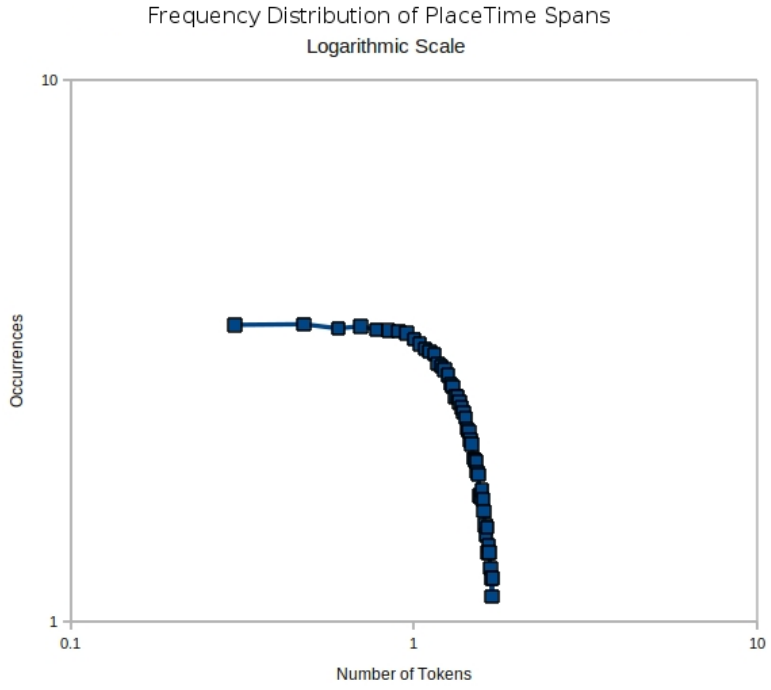


Figure Appx.4: PlaceTime spans, distribution on the logarithmic scale

Span Length	2	3	4	5	6	7	8	9	10
Unique Patterns	24	193	722	1381	1830	2033	2139	2083	1932

Table Appx. 4: PlaceTime spans, unique patterns for 9 different span lengths

C3. Analysis of the "ObjectMaterial" Property Spans

Number of Tokens	Occurrences	Log Num. of Tokens	Log Occurrences
2	1078	0.3	3.03
3	5222	0.48	3.72
4	4622	0.6	3.66
5	4258	0.7	3.63
6	3626	0.78	3.56
7	3088	0.85	3.49
8	2848	0.9	3.45
9	2214	0.95	3.35
10	1906	1	3.28
11	1532	1.04	3.19
12	1620	1.08	3.21
13	1156	1.11	3.06
14	986	1.15	2.99
15	990	1.18	3
16	882	1.2	2.95
17	754	1.23	2.88
18	718	1.26	2.86
19	616	1.28	2.79
20	534	1.3	2.73
21	376	1.32	2.58
22	428	1.34	2.63
23	352	1.36	2.55
24	346	1.38	2.54
25	298	1.4	2.47
26	220	1.41	2.34
27	218	1.43	2.34
28	190	1.45	2.28
29	182	1.46	2.26
30	154	1.48	2.19

Table Appx.5: ObjectMaterial pairs of span size in number of tokens in actual and logarithmic values

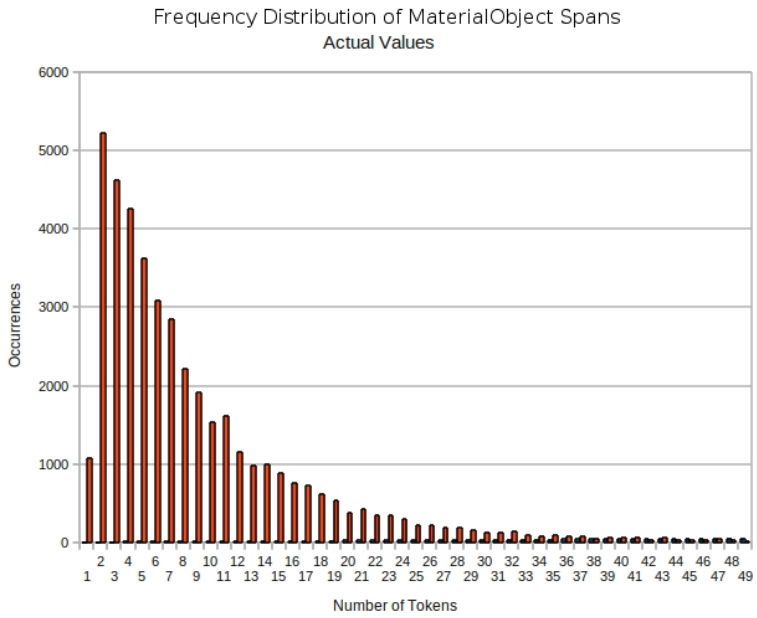


Figure Appx.5: ObjectMaterial spans, distribution actual values

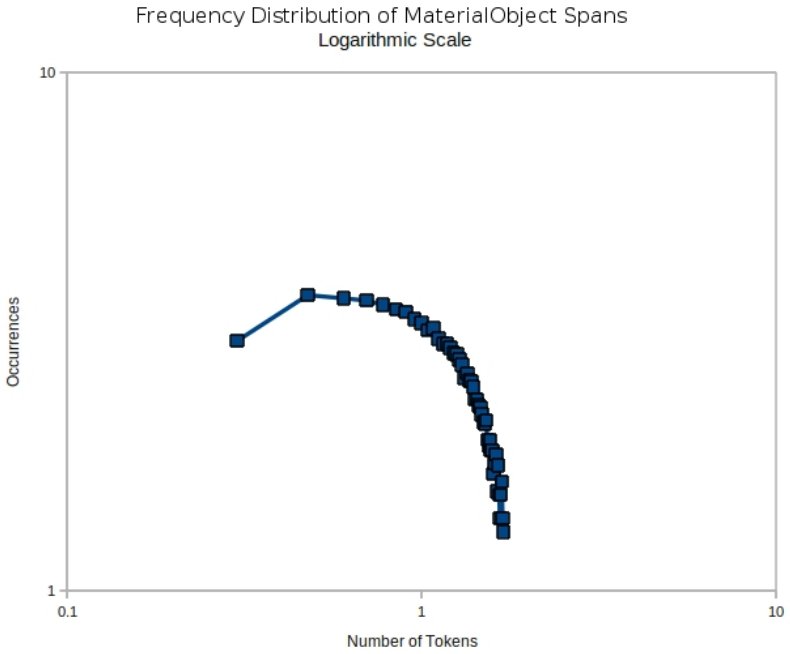


Figure Appx.6: ObjectMaterial spans, distribution on the logarithmic scale

Span Length	2	3	4	5	6	7	8	9	10
Unique Patterns	28	163	518	926	1146	1188	1108	949	853

Table Appx.6: ObjectMaterial spans, unique patterns for 9 different span lengths

C4. Sample of selected patterns denoting a Deposition Event

Phrase	POS Pattern
deposit contained lenses	NN VBD NNS
nails in burials	NNS IN NNS
brick from wall	NN IN NN
finds from this layer	NNS IN DT NN
deposit of crushed brick	NN IN VBN NN
ditch containing burnt flint	NN VBG JJ NN
Amphora incorporated into context	NNP VBD IN NN
bottle were recovered from contexts	NN VBD VBN IN NNS
artefacts retrieved from these deposits	NNS VBD IN DT NNS
animal bone in rubbish pits	NN NN IN JJ NNS
artefacts were recovered from grave contexts	NNS VBD VBN IN JJ NNS
finds were recovered from seven contexts	NNS VBD VBN IN CD NNS
brick was also recovered from pit	NN VBD RB VBN IN NN
pot was recovered from the grave	NN VBD VBN IN DT JJ
flint flakes were recorded in three contexts	NN NNS VBD VBN IN CD NNS
animal bone were recovered from the fill	NN NN VBD VBN IN DT NN
artefacts were recovered from any of these features	NNS VBD VBN IN DT IN DT NNS
animal bone fragments were collected from seven contexts	NN NN NNS VBD VBN IN CD NNS
flint tools have also been recovered from several sites	NN NNS VBP RB VBN VBN IN JJ NNS
deposit contained regular fragments of ceramic building materials including brick	NN VBD JJ NNS IN JJ NN NNS VBG NN

Table Appx.7: Sample patterns denoting a Deposition Event (EHE1004)

C5. Sample of selected patterns denoting a Production Event

Phrase	POS Pattern
Roman glass	NNP NN
Roman finds	NNP NNS
Plate 44 Georgian	NNP CD JJ
glass is modern	NN VBZ JJ
coins are Roman	NNS VBP NNP
bag dated Roman	NN VBN NNP
Prehistoric 174 Animal bone	NNP CD NNP NN
Mesolithic and Neolithic flint	NNP CC NNP NN
medieval unglazed ware bowl	NN JJ NN NN
glass is relatively modern	NN VBZ RB NN
finds of the Roman period	NNS IN DT NNP NN
artefacts predating the 19th century	NNS VBG DT JJ NN
coin of roughly 18th century	NN IN RB JJ NN
vessels dating to the 19th century	NNS VBG TO DT JJ NN
floors dating from the 18th century	NNS VBG IN DT JJ NN
animal bone and a sherd of late post-medieval	NN NN CC DT NN IN JJ JJ
large silver brooches of the 10th century AD	JJ NN NNS IN DT JJ NN NNP
artefacts that date to the medieval and post-medieval periods	NNS WDT NN TO DT NN CC JJ NNS

Table Appx.8: Sample patterns denoting a Production Event (EHE1002)

C6. Sample of selected patterns denoting a Context Event

Phrase	POS Pattern
prehistoric ditch	JJ NN
modern rubbish pit	JJ JJ NN
medieval boundary ditches	NN NN NNS
Gully Probably medieval	NNP RB NN
Road is post-medieval	NNP VBZ JJ
Modern 12 2 Deposit	NNP CD CD NNP
thick layer of modern date	JJ NN IN JJ
deposits were clearly modern	NNS VBD RB JJ
sites are all post-medieval	NNS VBP DT JJ
streets are probably medieval	NNS VBP RB NN
medieval field and enclosure ditches	NN NN CC NN NNS
sites pre-dating the 19th century	NNS JJ DT JJ NN
small gullies of probable post-medieval	JJ NNS IN JJ JJ
feature is of relatively modern	NN VBZ IN RB JJ
layers dating from the 18th century	NNS VBG IN DT JJ NN
layer dating to the 14th century	NN VBG TO DT JJ NN
layer date from the Roman period	NN NN IN DT NNP NN
Saxon and medieval occupation of the site	NNP CC NN NN IN DT NN
boundaries were formed in the 13th century	NNS VBD VBN IN DT JJ NN
archaeological deposits dating to the post-medieval	JJ NNS VBG TO DT NN NN
enclosures that had developed from the 2nd century	NNS WDT VBD VBN IN DT CD NN
human burials dating back to the Bronze Age	JJ NNS VBG RB TO DT NNP NNP

Table Appx.9: Sample patterns denoting a Context Event (EHE1001)

C7. Sample of selected patterns denoting a Consists of property

Phrase	POS Pattern
pottery finds	NN NNS
Ceramic ball	NNP NN
Ceramic artefacts	NNP NNS
artefacts of gold	NNS IN NN
small brick tiles	JJ NN NNS
floor is concrete	NN VBZ JJ
squared blocks of limestone	VBN NNS IN NN
pin was of iron	NN VBD IN NN
finds of animal bone	NNS IN NN NN
small flint rounded pebbles	JJ NN JJ NNS
pebbles and flecks of chalk	NNS CC NNS IN NN
wall is composed of stone	NN VBZ VBN IN NN

Table Appx.10: Sample patterns denoting a Consists of property (P45)

Appendix D

Evaluation Support Documents

D1. Use case Scenarios

Sample use cases transcript from notes as discussed with archaeology experts during project meetings. The example use cases group together uses case scenarios that carry common question problems. The problems are expressed as search scenarios which are loosely associated to CRM and CRM-EH entities and relationships.

1) Find CRM-EH name entities such as "Archaelogical Contexts" and "Physical Objects"
e.g.: Find all entities of type Archaeological Context.

2) Find CRM-EH name entities of a given attribute
e.g.: Find all entities of type Archaeological Context that have attribute *simple name* equal to post-hole

3) Find CRM-EH name entities that might have a given attribute and have a specific relationship to another name entity type which might have a give attribute.
e.g.: Find all Contexts containing a Find
e.g.: Find all Contexts of type "Hearth" containing a Find
e.g.: Find all Contexts of type "Hearth" containing a Find of type "Coin"
e.g.: Find all Contexts containing a Find of type "Coin"
e.g.: Find all Contexts consist of Material of type "pottery"
e.g.: Find all Contexts of type "Corn Mill" were Sample was taken
e.g.: Find all Contexts of type "Corn Mill" associated with Time Appellation of type "Roman"
e.g.: Find all context Finds of "brooch" associated with Time Appellation of type "Roman"

4) Find Stratigraphic Relations of CRM-EH entities
e.g.: ditch/fill – bellow context

5) Geospatial indication of CRM-EH entities
e.g: waterlogged ditch in same context with pit

6) Find Archeobotanical Evidence
e.g: macroscopic plant remains evidence

7) Finds evidence of Activity
e.g: site used for farming

8) Search by existing CRM-EH "Groups" (and definite features)
e.g: Find all contexts within a group that have certain features, eg simple name = post-hole

D2. Prototype System - Instructions for Manual Annotators

Manually annotate the following ten summary extracts with respect to the archaeological concepts of *Period, Physical Object, and Archaeological Context*.

Periods can be single or multi-worded phrases relating to a particular age or time such as 'Iron Age', 'Early Roman', 'Medieval' or ordinal related phrases like 1st, 2nd century etc

Objects can be single or multi-worded phrases relating to a physical object (artifacts) such as 'pottery sherds', 'flint flakes' etc., having a particular interest from an archaeological point of view.

Contexts can be single or multi-worded phrases relating to archeological places such as 'Context, Cut, Pit, Ditch, Trench' and their plural forms. It can include the place related outcomes of grouping/phasing activity, such as building, structures or monuments.

Use colour coding for the different entity type. For example highlight Period with blue colour, Context with green and Physical object with yellow.

Underline terms (single or multi-worded phrases) representing any of these single entities; aim to identify terms enjoying a rich meaning, so instead of pit, underline large pit, or instead of Roman annotate Early Roman if this is the case .

Use Italics for larger phrases involving two or possibly all three of the above concepts such as the phrase (annotation example)

- *22 pits containing Roman pottery sherds*
- **RomanoBritish boundary ditch**

D3. Manual Annotation Instructions (Main)

Thank you for taking part in this exercise – your assistance with our research is greatly appreciated. Please try to complete the whole exercise but you are free to take as many small breaks as required in order to progress smoothly with the task. Your names and IDs will not be recorded with the data. Please ask the researchers if there is anything you do not understand, or if any of the colours pose difficulties.

1. Aim

Your aim is to highlight key archaeological concepts from a selection of OASIS grey literature document summaries. The purpose is to provide a set of reference data (the 'gold standard') against which the performance of an automatic Natural Language processing system will be evaluated. Therefore, it is important to be as systematic and as consistent as possible.

Different colours will be used to highlight four different types of concepts, using Microsoft Word, as explained below. We interested in highlighting individual concepts and **also** phrases that express meaningful connections between individual concepts. When highlighting a concept, please include any modifiers in form of adjectives and adverbs *(large, circular, narrow etc.)* **excluding** any colours (*brownish, grey*). Also highlight lists of concepts as if a single multi-concept. It is not required to highlight any determiners *(a, the)* but if you do this will not affect the validity of the exercise. In addition to highlighting the individual concepts, please underline the immediate phrase that connects the concepts.

2. The four types of archaeological concepts

Finds: Physical objects that can be described as finds of archaeological interest. They may be manmade (eg amphora) or naturally occuring (eg flint flakes). Finds can be described as items of a material nature that are units for documentation and have physical boundaries that clearly separate them from other objects. Examples of Finds are: *a coin, coins, the bottle, a brick, the Aphrodite of Milos* as well as small finds, such as *bone fragments, flint flakes,* etc.

Time Periods: All kinds of names or codes for historical periods. Time Periods may vary in their degree of precision and may be expressed relative to other time frames. Examples of Time Periods are *Early Roman, Medieval, 10th Century*, etc. **Do not** highlight exact dates such as 1749, 6 of August and 10/09/1980 etc.

Contexts: Spatial elements that constitute an individual archaeological unit of excavation (and a basic site location in the excavation database). For this exercise, include both primitive contexts and larger groupings of contexts. Examples of Contexts are *pits, a cut, the deposit, context* (itself) as well as larger contexts, such as *enclosures, post-holes, hearths, a well, the floor, a structure, buildings, roads*, etc.

Materials: Materials that have an *archaeological* interest and are associated with physical objects (finds). Examples of Materials are *iron, copper, charcoal etc.* Sometimes a word might be treated as a Material if it modifies a Find but treated as a Find otherwise. For example, *the brick* is a Find, whereas *the brick oven* is a Material followed by a Find. Similar examples are *flint flakes* and the *pottery fragments* versus simply *flints* and *pottery*.

3. Further guidelines

Please highlight with the following principles in mind.

Negation Detection: Concepts or phrases that are negated should NOT be highlighted. For example, *No context was found to contain pottery* should NOT be highlighted. Since *context* is not highlighted because of negation, this means that *pottery* is also not highlighted

Relevance: Consider how relevant the entity is to the overall discourse.

Disambiguate between Finds and Materials. For example the term 'brick' can either refer to a material ie (a brick wall) or to a physical object ie (a brick found in context). You should decide on the conceptual alignment of terms that can be either materials or physical objects.

Endings: You should consider plural, gerund and possessive ending when applicable. For example you should consider 'bricks' 'panning' etc.

Phrases: You should consider conjunct and adjectival phrases. For example you should consider conjunctions of the kind 'Early Roman to Late Roman', 'Pottery and brick', as well as 'worked flint', 'small finds' etc.

Annotations should define meaningfully connections between entities by marking boundaries of phrases in text that normally begin with a single entity type and end with another entity type. It is irrelevant which entity type begins and which ends the phrase since phrases can use entity types interchangeably. For example annotators should aim to annotate both phrases; 'Roman period Find' and 'find of Roman period'. In addition, phrases might involve more than two conceptual entities. For example the above phrase could have been 'Iron finds of the Roman period' involving a Material (Iron), a Physical Object (Find) and a Time Period (Roman period), or even the above phrase could have been 'Deposit containing iron finds of the Roman period' involving as well a Place (Deposit) together with the other entities. Annotators may annotate phrases that engage at least two concepts to as many concepts as possible. The connection phrases can be very short having to two concepts next to each other (eg Roman Find) or connections can be larger spanning to a dozen or more words. For example annotators should annotate as one phrase the case '*Iron finds of the Roman period have been associated with the pit'*,

4. Colour Coding

Find {Grey 50% : #999999}
Time Period {Turquoise : #00FFFF }
Archaeological Context {Bright Green: #00FF00 }
Material {Red : #FF0000 }
Underline connection between concepts
Colour Coding (Colour Blind Version)
Find {Grey 50% : #999999}
Time Period {Blue : #0000FF }
Archaeological Context {Bright Green: #00FF00 }
Material {Red : #FF0000 }
Underline connection between concepts
5. Meaningful connections between concepts
Find and Time Period
i.e. Mediaeval Pottery
Archaeological Context and Time Period
i.e. Roman deposits
Find and Archaeological Context
i.e. Ditch containing coins
Find and E57 Material
i.e. Copper alloy artefacts
More Meaningful Examples
Two pits, a posthole and a linear cut, which are broadly dated from the Neolithic period to the Late Bronze Age
A quantity of human bone was recovered from its fill
Ceramic artefacts included pottery sherds, roof tiles and bricks all dated to the Roman period
deposit was medium brownish grey silty sand that also contained frequent charcoal
.....a broad range but three fragments off glass bottles in (13/007) were dated to the late 18th early 19th century and this is probably a.....
.....spits 35 Postmedieval and possibly medieval deposits, to be taken down carefully in one to three spits, down to c.11.1010.75m OD .....
....a total of 7 copper alloy artefacts were recovered from various Bronze Age contexts during the excavation/evaluation.....
*NOTE Example 5 mentions archaeological finds by a reference number for example (13/007). Mentions to archaeological finds and contexts should not be highlighted.

D4. Principles of Annotations Transfer

During transfer, archaeological place annotations (highlighted in green) were expressed as CRM E53.Place, archaeological find annotations (highlighted in grey) were expressed as CRM E19.Physical_Object, archaeological find material annotations (highlighted in red) were expressed as CRM E57.Material and time appellation annotations (highlighted in blue) were expressed as CRM E49.Time_Appellation. In addition, underlined phrases denoting 'rich' discussion which involved place and time appellation annotation were expressed as CRM-EH EHE1001.ContextEvent, when phrases involved find and time appellation were expressed as CRM-EH EHE1002.ContextFindProductionEvent, when phrases involved place and find were expressed as CRM-EH EHE1004.ContextFindDepositionEvent, and when phrases involved find and material were expressed as CRM P45.consists of property

Major principle: Transfer annotations "as is" following the given colour coding; thus do not alter any concept alignment, even if an obvious mistake exists.

Transferring Concepts: 'For all concepts 'and', 'or' and 'comma' are used to distinguish different terms that enjoy different terminological references. So even if annotator has provided a single unified highlighting for three concepts , during transferring three individual annotations are created. Example: beakers, jars and flagons beakers, jars and flagons.

For Time Appellations , 'to', 'dash -', 'slash /' are used to join time appellations into a single annotation (which though enjoys a dual terminological reference) example: late 2nd to early 4th century AD

Transferring 'Rich' phrases: Transferring 'rich' phrases as CRM-EH events required knowledge of the ontology and some comprehension of the phrase within the overall discourse. Four different events were addressed by the transferring process. EHE1001 Context Event involving a Place (archaeological context) and a Time Appellation, EHE1002 Context Find Production Event involving a Physical Object and a Time Appellation, EHE1004 Context Find Deposition Event involving a Place (archaeological context) and a Physical Object and P45 consists of involving a Physical Object and Material.

Transferring is based on the principle of unfolding the clauses contained within the identified by the annotator phrase. Each event is generated from a single clause. For example: worked flint recovered from topsoil and subsoil contexts and a small number of pits containing pottery of this date. Two clauses one (worked flint with topsoil and subsoil -conjunctions taken into account) and another (pits with pottery),

However a single clause might generate more than one but always different types of event. For example single pit containing Iron Age pottery The phrase creates two events, deposition between pit-pottery, and production between iron age- pottery.

There are cases were 'rich' phrase do not translate to CRM-EH event, for example a circular stone and timber building. No event between material -place

Re-annotating concepts to CRM-EH entities: A straightforward re-annotation (the CIDOC CRM annotation is also kept) of CRM entities to CRM-EH for those phrases identified as expressing CRM-EH events. The process is considering conjunctions for example Roman pits and postholes – two EHE0007.contexts are identified, pit and postholes

D5. Identified Terms for Inclusion after Pilot Evaluation

Terms not included by the terminological resources

aisle, blocks, ceiling, chalkbuild , chalklined, cropmarks, debris, features, frontage, glassworks, horse mandible, landscape, lane, lithics, nave, objects, occupation, paleo-channel, plot, remains, scatter, soil, subsoil, substratum, tarmac, topsoil, ware

Manually Added Terms in Matching Rules

Term	Broad Term
Bowl	Food Serving Container
Castle	Defence
Cemetery	Funerary Site
Flake	Debitage
Fort	Defence
Garden	Dwelling
Hollow	-
House	Dwelling
Inhumation	Burial and Funerary Site
Knife	Food and Preparation Equipment
Monument	Commemorative Monument
Nail	Fastening
Plough	Cultivation Object
Pond	Water Supply and Drainage
Playground	Dwelling
Ridge and Furrow	Cultivation Marks
Ring	-
Settlement	-
Shed	Building
Tool	Tools and Equipment
Weapon	Armour and Weapon

Table Appx.11: Manually Added Terms in Matching Rules

D6. OASIS Documents Metadata

Title	OASIS Metadata		CRM - Annotations	
	Finds	**Monuments**	**E19.Physical Object**	**E53.Place**
aocarcha1-4139	vessel	boundary ditch; extended inhumation; market garden	pottery(6); animal remains(5); copper alloy(4); finds (4); brick(4)	deposit(17); ceramic vessel(13); burial(12); gardens(2);
archaeol1-19366_1	core; window; cheese press; pot	building; field system; enclosure; kiln; pit; oven; post hole; butchery site	pottery(108); flint(108); charcoal(68); artefacts(49); pot(12);	pit(303); ditch(192); deposit(89); building(22); enclosure(70); kiln(17); oven(38); post-hole(66)
archenfi2-31470_1	ceramics	-	Clay(112); iron(33); charcoal(33); slag(19)	Context(142); layer(94); structure(51)
birmingh2-36136_1	-	-	Pottery(82); sherds(33); charcoal(22)	Layer(27); surface(20); pit(20); ditch(18)
borderar1-39096_1	roundhouse	-	Pottery(20); charcoal(22); sandstone(15)	Pit(45); deposit(28); gully(24); house(1)
cambridg3-27196_1	pottery cremation urns; glass bead necklace; human remains; copper alloy bracelets; shale bracelet; coffin nails; pottery beakers and jars; copper alloy finger	Pit; inhumation; cremation;	skeleton(91); bone(74); beads(68); necklace(31); human bone(13); bracelet(39); nails(60); pottery(45); beakers(7); ring(45); P45.consists_of(109) -several copper-alloy bracelets;	Grave(285); burial(223); cemetery(129); pit(87); cremation(86)

	rings; jet bead necklace; iron hobnails; copper alloy earring; st neots ware pottery		-copper alloy finger ring; -iron nails; -preserved necklaces associated with this adult burial;	
clairefe1-8958	Beaker; bowl; cup; mortarium; rotary quern;	Crop mark;	Pottery(17); finds(13); sherds(4)	Village(4); surface(3); road(3); burial(3); trackways(3); rectilinear enclosures(3)
colchest3-27821_1	Pottery; human bone; church foundations	Villa; church;	Pottery(33); sherds(12); finds(6); clay(5); human bone(3)	Foundation(32); church(30); layer(13);
compassa1-5431	-	-	Pottery(4); artefacts(3); glass(2)	Surface(4); deposit(3); trial trenches(3)
essexcou1-13888_1	Pottery	Garden out-building; pits; terracing	Pottery(9); finds(6); small sherds(4)	Deposits(8); castle(5); road(4); surface(4); structure(4); garden(2); building(2); isolated pits(3)
foundati1-4768	-	Cess Pit	Finds(16); chalk(14); flint(8); clay(6); pottery(6); sherds(6)	Structure(39); layer(21); foundation(19); cesspit(5);
norfolka1-16657_1	Buckle; pottery; burnt flint; debitage; animal remains;	Pit; post-hole; ditch; natural feature;	Charcoal(28); pottery(18); flake(15); flint(8);	Pit(79); ditch(34); deposit(19); post-hole(12)
northamp3-15136_1	Pottery; rivet wheat seed	Ditches; pits; post-holes; construction trenches;	Pottery(29); sherds(5); flint(5); bone(3)	Ditch(38); pit(14); context(6); post-hole(4)
suffolkc1-25638_2	Pottery	Listed halls and Gardens	Flint(5); pottery(5); charcoal(4);	Deposits (18); ditch(7); hall(5); enclosed garden(1)

suttonar1-8318	pot	Commercial office	Recorded pottery(2); finds(1); red brick(1)	Concrete floor(5); wall(3); deposits(3)
universi1-13308	Animal bone; pottery; tile	Collegiate church; precinct	Pottery(5); slate(5); tile(4);	Building(16); road(13); archaeological deposits(9); church(1); precinct walls(1)
wessexar1-6381	Pottery; animal bone	Pits; post-holes; ridge and furrow; ditched enclosure	Clay(39); charcoal(8); bone(6); pottery(3)	Secondary fill(17); pits(13); ditch(10); enclosure(7); ridge and furrow (2)

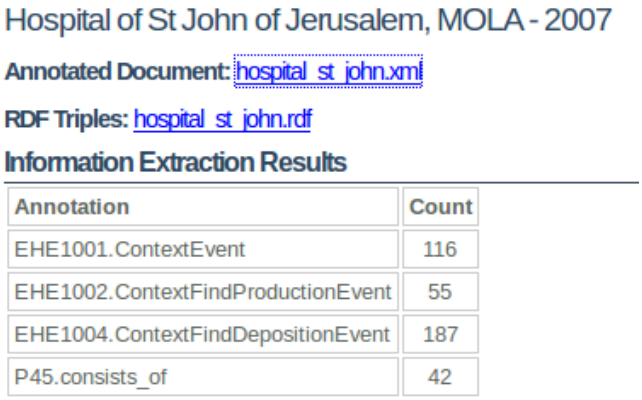
Table Appx.12: OASIS Metadata and CRM Annotations

Appendix E

Connected Projects

E1.MOLA Semantic Annotations of Monographs

Sample results of semantic annotation of the MOLA document "Excavations at the priory of the Order of the Hospital of St John of Jerusalem".



Hospital of St John of Jerusalem, MOLA - 2007

Annotated Document: hospital_st_john.xml

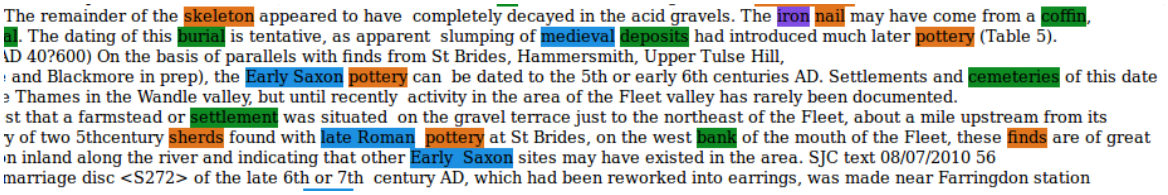
RDF Triples: hospital_st_john.rdf

Information Extraction Results

Annotation	Count
EHE1001.ContextEvent	116
EHE1002.ContextFindProductionEvent	55
EHE1004.ContextFindDepositionEvent	187
P45.consists_of	42

Figure Appx.7: MOLA Relation Extraction Annotations

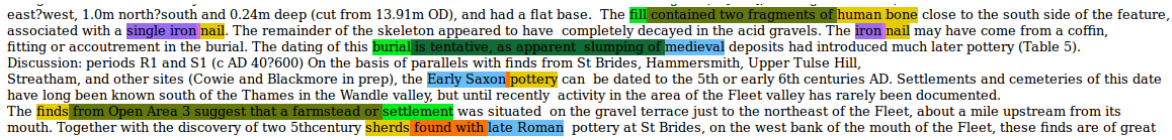
Example of contextual NER result with regard to CRM entities Physical Object (Orange), Place (Green), Time Appellation (blue) and Material (purple)



The remainder of the skeleton appeared to have completely decayed in the acid gravels. The iron nail may have come from a coffin,
ll. The dating of this burial is tentative, as apparent slumping of medieval deposits had introduced much later pottery (Table 5).
AD 40?600) On the basis of parallels with finds from St Brides, Hammersmith, Upper Tulse Hill,
and Blackmore in prep), the Early Saxon pottery can be dated to the 5th or early 6th centuries AD. Settlements and cemeteries of this date
Thames in the Wandle valley, but until recently activity in the area of the Fleet valley has rarely been documented.
st that a farmstead or settlement was situated on the gravel terrace just to the northeast of the Fleet, about a mile upstream from its
y of two 5thcentury sherds found with late Roman pottery at St Brides, on the west bank of the mouth of the Fleet, these finds are of great
n inland along the river and indicating that other Early Saxon sites may have existed in the area. SJC text 08/07/2010 56
marriage disc <S272> of the late 6th or 7th century AD, which had been reworked into earrings, was made near Farringdon station

Figure Appx.8: MOLA NER sample results

Example of NER and RE with regard to a range of CRM-EH entities



east?west, 1.0m north?south and 0.24m deep (cut from 13.91m OD), and had a flat base. The fill contained two fragments of human bone close to the south side of the feature,
associated with a single iron nail. The remainder of the skeleton appeared to have completely decayed in the acid gravels. The iron nail may have come from a coffin,
fitting or accoutrement in the burial. The dating of this burial is tentative, as apparent slumping of medieval deposits had introduced much later pottery (Table 5).
Discussion: periods R1 and S1 (c AD 40?600) On the basis of parallels with finds from St Brides, Hammersmith, Upper Tulse Hill,
Streatham, and other sites (Cowie and Blackmore in prep), the Early Saxon pottery can be dated to the 5th or early 6th centuries AD. Settlements and cemeteries of this date
have long been known south of the Thames in the Wandle valley, but until recently activity in the area of the Fleet valley has rarely been documented.
The finds from Open Area 3 suggest that a farmstead or settlement was situated on the gravel terrace just to the northeast of the Fleet, about a mile upstream from its
mouth. Together with the discovery of two 5thcentury sherds found with late Roman pottery at St Brides, on the west bank of the mouth of the Fleet, these finds are of great

Figure Appx.9: MOLA RE sample results

E2. CASIE Project Deliverables

Diagram of the E22_Man_Made_Object RDF triples. The diagram presents the type of information pieces extracted by the CASIE system.

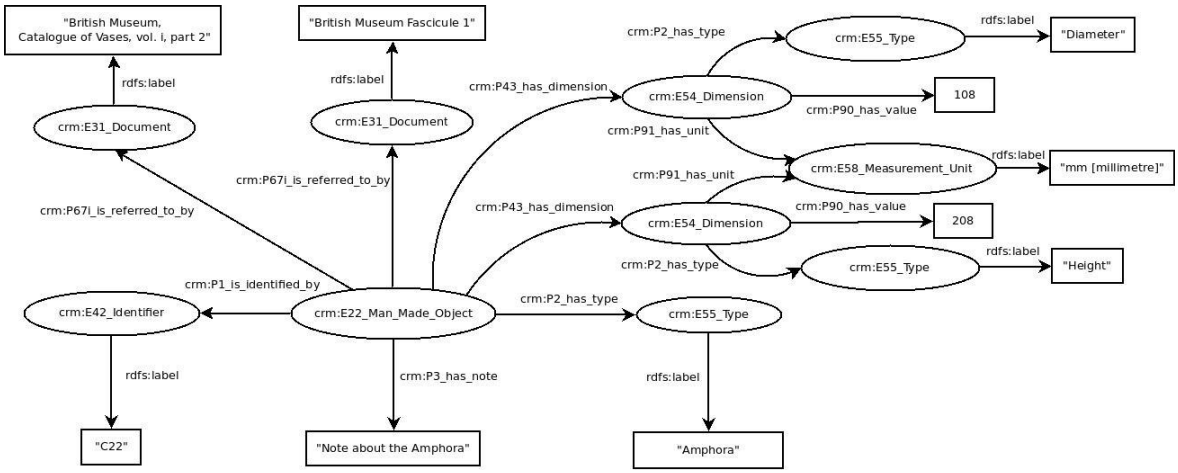


Figure Appx.10: RDF graph of CASIE project for E22_Man_Made_Objects

Example of contextual information chunks, *cultural object* highlighted in light blue, *height* highlighted in yellow, *catalogue reference* highlighted in red and the *description passage* highlighted in dark blue.

Askos with neck-spout and bird-tail projection; two suspension-handles at sides. Drab clay and slip ; dull black nds and wave line. Ht. 103. Ohnefalsch-Richter, 1884.Bibl. Cat. C 809.
Lekythos. Buff clay and polished slip ; grey-black bands and, on shoulder, concentric circles. Ht. 70. Pierides, 68.Bibl. Cat. C 791, pi. IV.
Small stamnos. Drab clay and slip ; grey-black bands. Ht. 67.Inv. 1908, 4-11, 40.
Jug with spout. Similar technique ; bands and con·centric hooks (on this ornament see Myres, op. cit., p. 81). Ht. 97. , W. Franks, 1879.Bibl. Cat. C 772.
Mastos with spout and two small suspension-handles. Similar technique ; lattice triangles. Ht. 91. nefalsch-Richter, 1884.Bibl. Cat. C 801, pi. IV.
c2.
Flask. Similar technique ; concentric rings. Ht. 110. From Kouklia, 1899.Bibl. Cat. C 799.
Jug. Similar technique ; bands ; on base, Maltese cross within rings. Ht. 60. From Curium, 1895, tomb 56.Bibl. t. C 775 ; Excav., p..74, fig. 129.
Jug. Buff clay and polished slip ; grey-brown paint. On lip, triangles ; herring-bone down handle ; on shoulder, ncentric hooks ; below, bands. Ht. 81.Inv. 1908, 4-11, 37.

Figure Appx.11: Annotation examples in context of CASIE project

Example of RDF document presenting triples of a E22_Man_Made_Object.

```
−<rdf:RDF>
  −<crm:E22_Man-Made_Object rdf:about="http://purl.org/NET/Claros/#British_Museum_Fascicule_1.566707">
    −<crm:P2_has_type>
      −<crm:E55_Type>
          <rdfs:label>Cup on three legs </rdfs:label>
          <crm:P127_has_broader_term rdf:resource="http://purl.org/NET/Claros/vocab#Shape"/>
        </crm:E55_Type>
      </crm:P2_has_type>
    −<crm:P3_has_note>
        Cup on three legs ; elementary spout. Height 179 millimetres. From Phoenikiais.Bibl. Cat. C 13.
      </crm:P3_has_note>
```

Figure Appx.12: RDF sample of E22_Man_Made_Object

Appendix F

F1. Part-of-Speech Tags used in the Hepple Tagger

CC - coordinating conjunction: ‘and’, ‘but’, ‘nor’, ‘or’, ‘yet’, plus, minus, less, times (multiplication), over (division). Also ‘for’ (because) and ‘so’ (i.e., ‘so that’).

CD - cardinal number

DT - determiner: Articles including ‘a’, ‘an’, ‘every’, ‘no’, ‘the’, ‘another’, ‘any’, ‘some’, ‘those’.

EX - existential ‘there’: Unstressed ‘there’ that triggers inversion of the inflected verb and the logical subject; ‘There was a party in progress’.

FW - foreign word

IN - preposition or subordinating conjunction

JJ - adjective: Hyphenated compounds that are used as modifiers; happy-go-lucky.

JJR - adjective - comparative: Adjectives with the comparative ending ‘-er’ and a comparative meaning. Sometimes ‘more’ and ‘less’.

JJS - adjective - superlative: Adjectives with the superlative ending ‘-est’ (and ‘worst’). Sometimes ‘most’ and ‘least’.

JJSS - -unknown-, but probably a variant of JJS

-LRB- - -unknown-

LS - list item marker: Numbers and letters used as identifiers of items in a list.

MD - modal: All verbs that don’t take an ‘-s’ ending in the third person singular present: ‘can’, ‘could’, ‘dare’, ‘may’, ‘might’, ‘must’, ‘ought’, ‘shall’, ‘should’, ‘will’, ‘would’.

NN - noun - singular or mass

NNP - proper noun - singular: All words in names usually are capitalized but titles might not be.

NNPS - proper noun - plural: All words in names usually are capitalized but titles might not be.

NNS - noun - plural

NP - proper noun - singular

NPS - proper noun - plural

PDT - predeterminer: Determiner like elements preceding an article or possessive pronoun; ‘all/PDT his marbles’, ‘quite/PDT a mess’.

POS - possessive ending: Nouns ending in ‘’s’ or ‘’’.

PP - personal pronoun

PRPR$ - unknown-, but probably possessive pronoun

PRP - unknown-, but probably possessive pronoun

PRP$ - unknown, but probably possessive pronoun,such as ‘my’, ‘your’, ‘his’, ‘his’, ‘its’, ‘one’s’, ‘our’, and ‘their’.

RB - adverb: most words ending in ‘-ly’. Also ‘quite’, ‘too’, ‘very’, ‘enough’, ‘indeed’, ‘not’, ‘-n’t’, and ‘never’.

RBR - adverb - comparative: adverbs ending with ‘-er’ with a comparative meaning.

RBS - adverb - superlative

RP - particle: Mostly monosyllabic words that also double as directional adverbs.

STAART - start state marker (used internally)

SYM - symbol: technical symbols or expressions that aren’t English words.

TO - literal “to”

UH - interjection: Such as ‘my’, ‘oh’, ‘please’, ‘uh’, ‘well’, ‘yes’.

VBD - verb - past tense: includes conditional form of the verb ‘to be’

VBG - verb - gerund or present participle

VBN - verb - past participle

VBP - verb - non-3rd person singular present

VB - verb - base form: subsumes imperatives, infinitives and subjunctives.

VBZ - verb - 3rd person singular present

WDT - ‘wh’-determiner

WP$ - possessive ‘wh’-pronoun: includes ‘whose’

WP - ‘wh’-pronoun: includes ‘what’, ‘who’, and ‘whom’.

WRB - ‘wh’-adverb: includes ‘how’, ‘where’, ‘why’. Includes ‘when’ when used in a temporal sense.

:: - literal colon

, - literal comma

$ - literal dollar sign

- - literal double-dash

“ - literal double quotes

´ - literal grave

(- literal left parenthesis

. - literal period

- literal pound sign