ARIADNE

D16.2: First Report on Natural Language Processing

Authors:

Andreas Vlachidis, USW
Doug Tudhope, USW
Milco Wansleeben, LU
Katie Green, ADS
Lei Xia, ADS
Michael Charno, ADS
Holly Wright, ADS

Ariadne is funded by the European Commission's
7th Framework Programme.

Version: 1.0 (*final*) **11 May 2015**

Authors:

Andreas Vlachidis, University of South Wales (USW)

Doug Tudhope, University of South Wales (USW)

Milco Wansleeben, Universiteit Leiden (LU)

Katie Green (ADS)

Lei Xia (ADS)

Michael Charno (ADS)

Holly Wright (ADS)

Versions	Nr.	Authors & changes made	Date
Draft	.5	Holly Wright (ADS) – compilation version for first review	5.02.2015
Draft	.6	Katie Green (ADS)	10.04.2015
Draft	.7	Holly Wright (ADS)	30.04.2015
Draft	.8	Douglas Tudhope, Andreas Vlachidis (USW)	08.05.2015
Final	1.0	Holly Wright (ADS)	11.05.2015

ARIADNE is a project funded by the European Commission under the Community's Seventh Framework Programme, contract no. FP7-INFRASTRUCTURES-2012-1-313193. The views and opinions expressed in this presentation are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.

Table of Contents

Glossary

ADS	Archaeology Data Service
Archaeotools	NLP project to create tools for archaeologists to allow archaeologists to discover, share and analyse datasets
CIDOC-CRM	The CIDOC Conceptual Reference Model (CRM) provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation
CRF	Conditional Random Field
F-Measure	A measure of accuracy calculated from the recall and precision measurements
GATE	A computer architecture framework for NLP
GATEfication	The design and transformation options for translating the original resources (of SKOSified RDF files) into GATE friendly OWL-Lite structures
Gold Standard	A test set of human annotated documents describing the desirable system outcome
Grey liturature	Unpublished reports
IE	Information Extraction
JAPE	Specially developed pattern matching language for GATE
Linked Open Data	A way of publishing structured data that allows metadata to be connected and enriched
NLP	Natural Language Processing
NER	Named Entity Recognition
OBIE	Ontology Based Information Extraction
OWL-Lite	Ontologies in GATE purely support the aims of information extraction and are not stand-alone formal ontologies for logic-based purposes
Polysemy	Multiple, related meanings
RCE	Rijksdienst Cultureel Erfgoed Thesauri
RDF	Resource Description Framework
SENESCHAL	Semantic ENrichment Enabling Sustainability of arCHAeological Links Project
SKOS	Simple Knowledge Organization System
STAR	Semantic Technologies for Archaeological Resources
STELLAR	Semantic Technologies Enhancing Links and Linked data for Archaeological Research.
String matching	The action of matching several strings (patterns) within a larger string or text
SVM	Linear Support Vector Machine
Synonymy	Similar meanings
Text Mining	The process of deriving information from text
Training data/documents	The annotated text used to train NLP classifiers
URI	Unique Resource Identifier
XML	Extensible Markup Language
XSL	Microsoft Excel format

Executive Summary

This document is a deliverable (D16.2) of the ARIADNE project ("Advanced Research Infrastructure for Archaeological Dataset Networking in Europe"), which is funded under the European Community's Seventh Framework Programme. It presents results of the work carried out in Task 16.2 "Natural Language Processing (NLP)".

NLP is an interdisciplinary field of computer science, linguistics and artificial intelligence that uses many different techniques to explore the interaction between human (natural) and computer languages.

The ARIADNE partners involved in this deliverable have explored NLP with the aim of making text-based resources more discoverable and useful, as part of the more research-based workpackages within the project. The partners have specifically focused on one of the most important, but traditionally difficult to access resources in archaeology; the largely unpublished reports generated by commercial or "rescue" archaeology, commonly known as "grey literature".

The partners have explored both rule-based and machine learning NLP methods, the use of archaeological thesauri in NLP, and various Information Extraction (IE) methods in their own language. This includes work by the University of South Wales, in partnership with Leiden University on archaeology thesauri for NLP, which applies Named Entity Recognition (NER) to the Dutch Rijksdienst Cultureel Erfgoed (RCE) Thesauri. The process of importing a subset of RCE thesauri resources into a specific framework (GATE), and the suitability and performance of the selected resources when used for the purposes of Named Entity Recognition (NER) are discussed. This report outlines issues relating to the role of the RCE thesauri in NER, and further development of techniques for the annotation of Dutch compound noun forms.

South Wales have also undertaken a study for a Dutch NER pipeline, which includes the results of the early pilot evaluation based on the input of a single, manually annotated document. The report also presents the results of the vocabulary transformation task from spreadsheets to RDF/XML hierarchical structures, expressed as OWL-Lite (ontology). Observations relate to the vocabulary transformation process and pipeline results, and reveal initial issues that affect vocabulary usage and focus of the NER exercise.

Work has been carried out by the Archaeology Data Service (ADS) at the University of York to develop and evaluate machine learning-based NLP techniques and integrate them into a new metadata extraction web application, which will take previously unseen English language text as input, and identify and classify named entities within the text. The outputs will then be used to enrich the resource discovery metadata for existing and future resources. The final application will include a web-based, user friendly interface that can be used by archaeological practitioners to automatically generate metadata related to uploaded text-based content on a per-file basis, or by using batch creation of metadata for multiple files.

This report presents the results of the work carried out to date, and presents the issues to be addressed during the remainder of the ARIADNE Project.

1 Introduction

1.1 Background

Across Europe, the archaeological domain generates vast quantities of text. This text can range from unpublished fieldwork and specialist reports often referred to as "grey literature", to published journal articles and monographs. Yet access to the valuable information locked within these texts can be difficult to access and can lead to new knowledge not being fed into the wider archaeological domain. The detrimental effect on archaeological knowledge, as a result of the inaccessibility and difficulty of discovery of these texts, has in recent years, begun to be increasingly recognised as a significant problem within the domain, and needs to be addressed.

The online delivery of "born digital" material and "digitised" versions of legacy material can provide a solution to addressing issues of access, however, access is reliant on effective discovery mechanisms, which are in turn reliant upon high-quality metadata. Indexing and metadata creation can be time consuming and may lack consistency when done by hand, and when created it is rarely integrated with the wider archaeological domain data.

The advancement of text mining allows this process of deriving information from large volumes of text to be automated. Text mining indexes all words found in a text file and computes a matrix of frequencies that enumerates the number of times each word occurs in the text. The extracted numeric indices can then be further analysed. The overarching goal is to turn text into data for analysis via the application of Natural Language Processing (NLP) and its subfield, Information Extraction (IE). These processes can be used to identify patterns, trends, and "important" words or terms within text, which can be used to improve archaeological information discovery, retrieval, comparison, analysis, and link texts to other types of data[1].

1.2 Natural Language Processing

NLP is an interdisciplinary field of computer science, linguistics and artificial intelligence, which uses many different techniques to explore the interaction between human (natural) and computer languages. IE is a specific NLP text analysis technique which extracts targeted information from context. This technique analyses textual input to form a new textual output capable of further manipulation. There are two distinct types of information extraction systems; rule-based and machine learning systems.

1.2.1 Rule-Based Systems

Rule-based systems use hand-crafted rules to make deductions or choices that are targeted at creating abstractions that relate to specific IE scenarios. This system requires expert domain knowledge and makes use of domain-independent linguistic syntax to negotiate semantics in context and extract information for a defined problem. Rule-based systems have the ability to achieve high levels of precision when identifying general purpose entities[2], but creating hand-crafted rules is

[1] Richards, J.D., D. Tudhope and A. Vlachidis. (in press). 'Text Mining in Archaeology: Extracting Information from Archaeological Reports'. In, J.A. Barceló and I. Bogdanovic (eds.) Mathematics in Archaeology. Science Publishers, Boca Raton, Florida.

labour intensive, requires extensive domain knowledge, and a complete understanding of the IE problem that the system is trying to resolve. Criticisms of the approach include its costly nature and arguably its limited adaptability to new IE scenarios[3]. Proponents of rule-based systems, claim the value in the rule-based approach is the fact that they do not require training to deliver results, and depending upon the IE task to be carried out, the linguistic complexity can be bypassed and a small number of rules can be used to extract very large sets of variant information[4].

1.2.2 Machine Learning

Machine learning has been proposed as the solution to overcoming the domain expertise dependency of rule-based systems. Machine learning describes the construction and study of algorithms that can "learn" from data and can support supervised and unsupervised learning activities. The supervised learning process is based on a training dataset that has been annotated by human experts, which is used by the machine learning process to deliver generalisations about the extraction rules, which are then able to perform a large scale exercise over a larger corpus[5]. The annotation of a small corpus of training documents is considered to be less labour intensive than the creation of hand-crafted extraction rules, since the latter requires programming expertise and domain knowledge[6]. However, the size of the training dataset may depend on the range and complexity of the desired annotations.

Unsupervised learning refers to the machine learning method that does not require any human intervention at all. The output of the training dataset using this method is not characterised by any desired label, instead a probabilistic clustering technique is employed which partitions the training dataset and describes the output result, with the subsequent generalisation run on a larger collection[7].

Both rule-based and machine learning approaches have their strengths and weaknesses, and work carried out by the ARIADNE partners means to explore both with regard to their usefulness within the archaeological domain.

[2] Feldman, R., Y. Aumann, M. Finkelstein-Landau, E. Hurvitz, Y. Regev and A. Yaroshevich. 2002. A Comparative Study of Information Extraction Strategies. Proceedings (CICLing-2002) Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico city. Mexico, 17–23 February.

[2]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.

[3] Feldman et al. op. cit.

[4] Hobbs, J.R., D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel and M. Tyson. 1993. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. In Proceedings (IJCAI 1993) 13th International Joint Conference on Artificial Intelligence, Chambery, France, 28 August–3 September.

[5] Richards et al. op. cit.

[6] Moens, M.F. 2006. Information Extraction Algorithms and Prospects in a Retrieval Context. Dordrecht, Springer.

[7] Nilsson, N. 2005. Introduction to Machine Learning. Nils J Nilson publications.
http://robotics.stanford.edu/people/nilsson/mlbook.html

1.3 NLP in Archaeology

The archaeological discipline has excellent potential for the exploration of natural language processing techniques because it has a relatively well-controlled set of vocabularies[8]. Significant effort has been put into the development of controlled word lists or thesauri, including the UK MIDAS data standard[9] and the Dutch Rijksdienst Cultureel Erfgoed Thesauri. However, archaeological vocabularies do pose a challenge. Unlike highly specialised domains, which have vocabularies unique to that domain, archaeological terminology consists of common everyday words, for example "wall", and "ditch". In archaeological vocabularies there are also distinctions between descriptions of the present and the archaeological past (for example "ditch" has much more significance if it is a "prehistoric ditch").

Over the past ten years, a number of projects have attempted to use text mining techniques on archaeological material. A pilot application was carried out by Armani, Abajian, Kodratoff and Matte-Tailliez to use string matching to extract information[10]; the *OpenBoek* project experimented with memory-based learning to extract chronological and geographical terms from Dutch archaeological texts[11]; Byrne has explored the application of NLP to extract event information from archaeological texts[12]; the Archaeotools project adopted a machine learning approach, while OPTIMA (which provided the basis of the STAR and STELLAR projects) [13] adopted a rule-based approach. The results of these projects will feed directly into the NLP work carried out as part of ARIADNE. Currently several projects in addition to the ARIADNE project are exploring different NLP techniques for archaeology, including the DADAISM project which is exploring how text mining techniques can be used to extract information about images to improve search and browsing of image archives and improve image labelling[14].

Through experimentation with NLP, the ARIADNE project hopes to make text-based resources more discoverable and useful, as part of the more research-based workpackages. The ARIADNE partners involved in this deliverable have turned their attention to one of the most important, but traditionally difficult to access resources in archaeology; the largely unpublished reports generated by commercial or "rescue" archaeology, commonly known as "grey literature". The partners have adopted a variety of approaches to NLP. The objective of this document is to report upon the work carried out by the partners to date. The following sections of this document present these reports. The partners will continue to pursue and refine these approaches over the duration of the ARIADNE project, which will be discussed in D16.4, *Final report on natural language processing*.

[8] Richards. et al. op. cit.

[9] English Heritage 2007. MIDAS Heritage—The UK Historic Environment Data Standard (Best practice guidelines) http://www.english-heritage.org.uk/publications/midas-heritage/midasheritagepartone.pdf.

[10] Amrani, A., V. Abajian, Y. Kodratoff and O. Matte-Tailliez. 2008. A chain of text-mining to extract information in Archaeology. Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference 1–5.

[11] Paijmans, H. and S. Wubben. 2008. Preparing archaeological reports for intelligent retrieval. pp. 212–217. In: A. Posluschny, K. Lambers and I. Herzog (eds.). Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA) Berlin, Germany, April 2–6, 2007. Kolloquien zur Vor- und Frühgeschichte Band 10, Bonn.

[12] Byrne, K.F. and E. Klein. 2010. Automatic Extraction of Archaeological Events from Text. pp. 48–56. In: B. Frischer, J.W. Crawford and D. Koller (eds.). Making History Interactive. Proceedings of the 37th Computer Application in Archaeology Conference, Williamsburg 2009. Archaeopress, Oxford.

[13] Tudhope, D., K. May, C. Binding and A. Vlachidis. 2011. Connecting Archaeological Data and Grey Literature via Semantic Cross Search, Internet Archaeology 30. http://dx.doi.org/10.11141/ia.30.5.

[14] DADAISM. (2015) Digging into Archaeological Data and Image Search Metadata. http://dadaism-did.org/

2 Archaeology Thesauri for Natural Language Processing: Applying the Rijksdienst Cultureel Erfgoed (RCE) Thesauri in Named Entity Recognition (NER)

2.1 Introduction

Information Extraction (IE) is a specific NLP technique which extracts targeted information from textual context. It is a process whereby a textual input is analysed to form a textual output capable of further manipulation. Rule-based IE systems consist of a pipeline of cascaded software elements that process input in successive stages. Hand-crafted rules make use of domain knowledge and vocabularies together with domain-independent linguistic syntax, in order to negotiate semantics in context.

The employment of rule-based IE and domain vocabulary resources distinguishes this approach from supervised machine learning work, which relies on the existence and quality of training data. The absence of a training corpus coupled with the availability of a significant volume of high quality domain-specific knowledge organization resources, such as a conceptual model, thesauri and glossaries were contributing factors to the adoption of rule-based techniques in this study. Rules invoke input from gazetteers, lexicons, dictionaries and thesauri to support the purposes of Named Entity Recognition (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. In addition, rule-based IE techniques exploit a range of lexical, part of speech and syntactical attributes that describe word level features, such as word case , morphological features and grammar elements that support definition of rich extraction rules, which are employed by the NER process.

Rule-based techniques were employed with available archaeological vocabularies from English Heritage (EH) and Rijksdienst Cultureel Erfgoed (RCE). The GATE framework[15] used for this work is the outcome of a 20 year old project established in 1995 at the University of Sheffield with a world-wide set of users; the GATE community has been involved in a plethora of European research projects. This builds upon previous work with the grey literature digital library from the Archaeology Data Service, which proved capable of semantic enrichment of grey literature reports conforming both to archaeological thesauri and corresponding CIDOC CRM ontology classes representing archaeological entities, such as Artefacts, Features, Monuments Types and Periods. The current pilot system has achieved some promising semantic enrichment of Dutch grey literature reports, for example artefacts such as “pottery/ aaardewerk” (via the RCE Archeologische artefacttypen vocabulary) and other concepts including time periods.

The generalisation of the previous rule based techniques to Dutch language grey literature faces the challenge of a different set of vocabularies. It also faces the issue of differences in language characteristics, for example compound noun forms. These present a challenge for the usual “whole word” matching mechanisms. Compound noun forms examples might include “beslagplaat” where both “beslag” and “plaat” are known to the vocabulary and also “aardewerkmagering” where aardewerk (pottery) is known but “magering” is not. Current work is investigating the development of gazetteers operating on part matching, in order to overcome the ‘whole word’ restriction.

[15] GATE (General Architecture for Text Engineering) https://gate.ac.uk

2.2 GATEfication of RCE Thesauri

This section discusses the process of importing a subset of RCE thesauri resources into the GATE framework and the suitability and performance of the selected resources when used for the purposes of Named Entity Recognition (NER) as carried out by the University of South Wales, in partnership with Leiden University. It discusses issues relating to the role of the RCE thesauri in NER and the further development of techniques for the annotation of Dutch compound noun forms. The discussion is divided into three main sections;

1. The first section discusses the process of "GATEfication", i.e. the design and transformation options for translating the original resources (SKOSified RDF files) into GATE friendly OWL-Lite structures capable of supporting the NER matching mechanism for the Information Extraction pipeline.
2. The second section reveals a range of commonly occurring structural, labelling and coverage issues affecting the capacity of RCE thesauri to support the NER task.
3. The third section proposes several actions for modification and enhancement of the original resources towards a (more) NLP friendly version, which could significantly improve the matching performance and applicability of the RCE thesauri in NER.

   The NER task is focused on the identification of the following concepts (entities) in the context of Dutch archaeological grey literature;

   - Artefacts (finds or physical objects)
   - Features (archaeological context e.g. posthole)
   - Materials
   - Monuments Types
   - Places (focus on place names such as districts)
   - Periods (time appellations including numerical appellations e.g. 480 BC).

   Respectively, the following thesauri have been selected to support NER for the above entities:

   - Archeologische artefacttypen

     http://rce.rnaviewer.net/nl/item?uri=http://www.rnaproject.org/data/2ce46848-3b3b-4371-96d3-7c4011fcd2d6
   - A subset of Archeologische artefacttypen
   - Materialen (Global Thesauri)

     http://rce.rnaviewer.net/nl/item?uri=http://www.rnaproject.org/data/bc3e12d6-ccf3-4f28-b47d-6f9424cb8b17
   - Archeologische complextypen abr+

     http://rce.rnaviewer.net/nl/item?uri=http://www.rnaproject.org/data/548a1b4e-0fa9-4f8d-9d83-f28d0792c3d0
   - Locaties (Global Thesauri)

     http://rce.rnaviewer.net/nl/item?uri=http://www.rnaproject.org/data/8407b36-d75a-4f54-8c7c-ed2cc79f730c

- Archeologische perioden abr+

  http://rce.rnaviewer.net/nl/item?uri=http://www.rnaproject.org/data/c028640a-0359-45e9-a1fb-19d23e939ece

GATE enables Ontology Based Information Extraction (OBIE) techniques using OWL-Lite[16] ontological resources that support the conceptual and glossary requirements of an NER task. Ontologies in GATE provide the necessary conceptual framework for driving the NER task and contribute the glossary input to the matching mechanism. Their main benefit is that they allow the definition of matching rules (JAPE) that exploit the transitive relationships of an ontological structure. As a result matching rules become flexible and capable of exploiting only those parts of the ontological resource that fall within the scope of an entity definition. For example a single line rule can exploit and consequently provide matches from a Monument Type resource, only for those entries that are described as "Defensive Structures", including "castles", "tower" etc. and their sub-types and sub-sub-types. In addition, individual ontological classes or instances benefit from the use of parameters holding spelling variations, synonyms, SKOS identifiers and any other sort of bespoke parameters useful to a particular task. Thus, matches derived from an ontological resource enjoy dimensions that could be useful for further information retrieval and/or interoperability purposes.

The RCE thesauri resources are made available online via API key access:

> http://rce.rnatoolset.net/api/getnodelist.aspx?rna_api_key=e22085bb-3e1b-4b5f-8b49-e6ddf015a1cc&uri=http://www.rnaproject.org/data/c47470f7-0321-4479-9bb2-757c3fa4fb22

Although it could be partially parsed from GATE, the original thesauri structure is not suitable for supporting OBIE approaches, due to the incapacity of the GATE ontology tool to parse (understand) broader/narrower term relationships. As a result, the original resources when loaded into GATE appear as flat structures enabling only two kinds of matching rule definition/exploit the whole resource or define one-to-one rules for every single concept that needs to be matched. The former case can have a severe impact on precision, since the totality of a resource is not always useful, while the latter leads to the definition of too many rules (as many as the concepts to be matched), which in some cases can be thousands.

Transformation of the original SKOSified (RDF) thesauri to OWL-Lite was necessary for:

- exploiting the hierarchical relationships of the resources,
- enabling matching on alternative labels and synonyms,
- enhancing matches with useful interoperable attributes already available in the original resources, such as SKOS unique identifier.

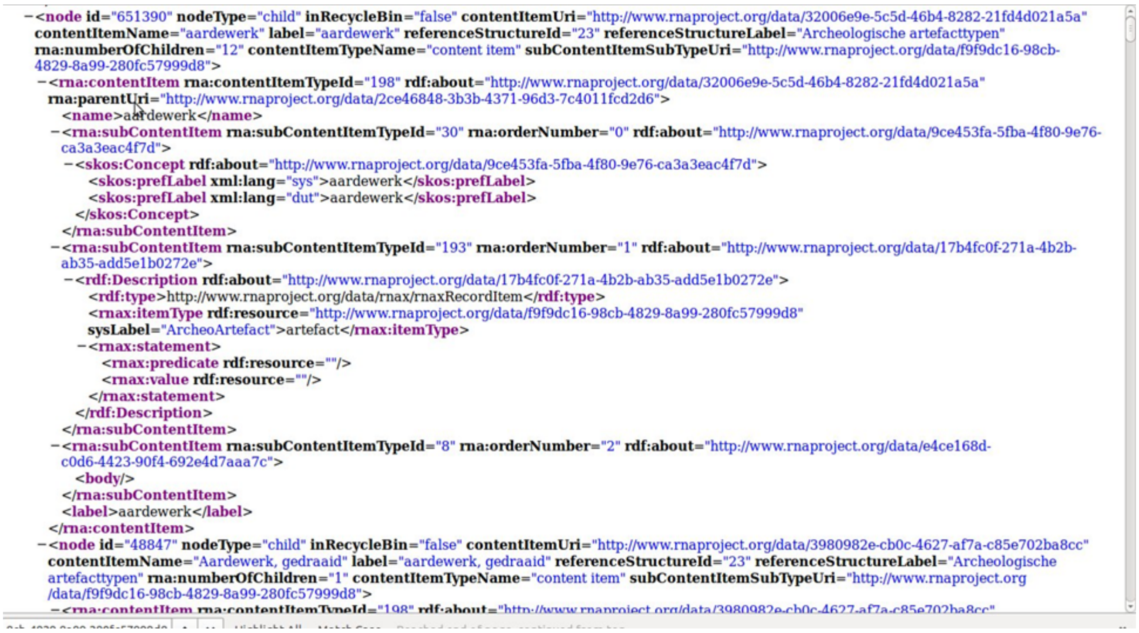
In addition, the transformation process created new human-readable unique resource identifiers (URIs) while maintaining the original **rna:contentItem** and **skos:Concept** (rdf:about) references for individual entries. The necessity to provide new (more) human-readable URIs for classes is dictated by GATE's behaviour towards exposing class URI to JAPE rules. The RCE uses multi-character unique identifiers where individual entries are identified by a common base URI followed by a unique long string identifier e.g. http://www.rnaproject.org/data/35240121-7752-4567-a613-74bca32ff311.

[16] OWL-lite ontologies in GATE purely support the aims of information extraction and are not stand-alone formal ontologies for logic based purposes. For example, thesaurus narrower term relationships are implemented using rdfs:subClassOf for internal GATE purposes only.

Although, the base URI can be left out from the rule definition, still the rule must refer to the class name, which in this case is a long non-human-readable string. Definition of such rules, although not impossible, can have implications particularly during the debugging stage.

The structure of the original thesauri resources encapsulate individual entries under a node shell containing a **rna:contentItem** and two **rna:subContentItem(s)** one reflecting the individual entry and another reflecting thesaurus details (see below).

```
<node …. contentItemUri = “unique URI of node”>
- - <rna:contentItem rdf:about = “same as unique URI of node”  rna:parentUri = “the URI of parent”>
- - - <rna:subContentItem rdf:about = “the unique URI of Item”>
- - - - <skos:Concept rdf:about= “same as unique URI of Item” >
- - - - - ……….
- </skos:Concept> </rna:subContentItem> </rna:contentItem></node>
```



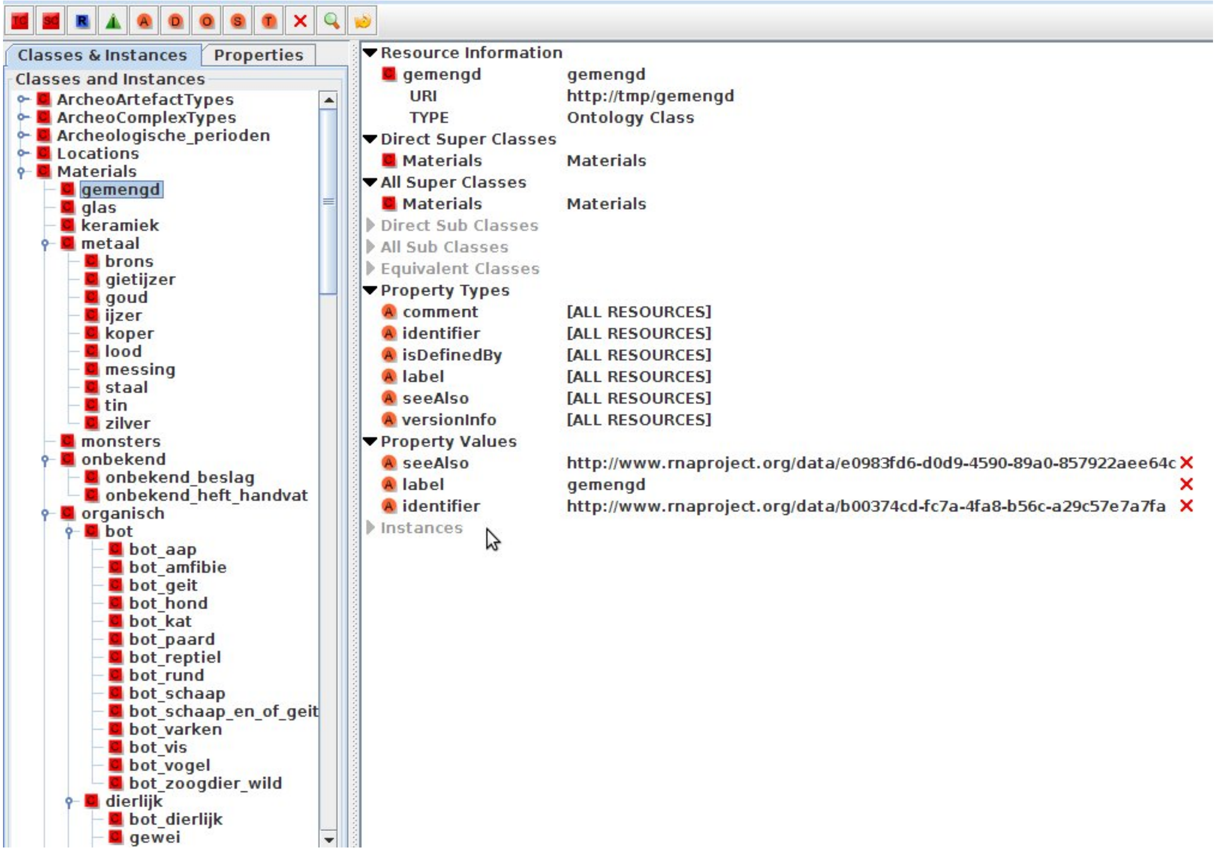
```
−<node id="651390" nodeType="child" inRecycleBin="false" contentItemUri="http://www.rnaproject.org/data/32006e9e-5c5d-46b4-8282-21fd4d021a5a"
 contentItemName="aardewerk" label="aardewerk" referenceStructureId="23" referenceStructureLabel="Archeologische artefacttypen"
 rna:numberOfChildren="12" contentItemTypeName="content item" subContentItemSubTypeUri="http://www.rnaproject.org/data/f9f9dc16-98cb-
 4829-8a99-280fc57999d8">
  −<rna:contentItem rna:contentItemTypeId="198" rdf:about="http://www.rnaproject.org/data/32006e9e-5c5d-46b4-8282-21fd4d021a5a"
   rna:parentUri="http://www.rnaproject.org/data/2ce46848-3b3b-4371-96d3-7c4011fcd2d6">
     <name>aardewerk</name>
   −<rna:subContentItem rna:subContentItemTypeId="30" rna:orderNumber="0" rdf:about="http://www.rnaproject.org/data/9ce453fa-5fba-4f80-9e76-
     ca3a3eac4f7d">
     −<skos:Concept rdf:about="http://www.rnaproject.org/data/9ce453fa-5fba-4f80-9e76-ca3a3eac4f7d">
         <skos:prefLabel xml:lang="sys">aardewerk</skos:prefLabel>
         <skos:prefLabel xml:lang="dut">aardewerk</skos:prefLabel>
       </skos:Concept>
     </rna:subContentItem>
   −<rna:subContentItem rna:subContentItemTypeId="193" rna:orderNumber="1" rdf:about="http://www.rnaproject.org/data/17b4fc0f-271a-4b2b-
     ab35-add5e1b0272e">
     −<rdf:Description rdf:about="http://www.rnaproject.org/data/17b4fc0f-271a-4b2b-ab35-add5e1b0272e">
         <rdf:type>http://www.rnaproject.org/data/rnax/rnaxRecordItem</rdf:type>
         <rnax:itemType rdf:resource="http://www.rnaproject.org/data/f9f9dc16-98cb-4829-8a99-280fc57999d8"
         sysLabel="ArcheoArtefact">artefact</rnax:itemType>
       −<rnax:statement>
           <rnax:predicate rdf:resource=""/>
           <rnax:value rdf:resource=""/>
         </rnax:statement>
       </rdf:Description>
     </rna:subContentItem>
   −<rna:subContentItem rna:subContentItemTypeId="8" rna:orderNumber="2" rdf:about="http://www.rnaproject.org/data/e4ce168d-
     c0d6-4423-90f4-692e4d7aaa7c">
       <body/>
     </rna:subContentItem>
     <label>aardewerk</label>
   </rna:contentItem>
  −<node id="48847" nodeType="child" inRecycleBin="false" contentItemUri="http://www.rnaproject.org/data/3980982e-cb0c-4627-af7a-c85e702ba8cc"
   contentItemName="Aardewerk, gedraaid" label="aardewerk, gedraaid" referenceStructureId="23" referenceStructureLabel="Archeologische
   artefacttypen" rna:numberOfChildren="1" contentItemTypeName="content item" subContentItemSubTypeUri="http://www.rnaproject.org
   /data/f9f9dc16-98cb-4829-8a99-280fc57999d8">
   −<rna:contentItem rna:contentItemTypeId="198" rdf:about="http://www.rnaproject.org/data/3980982e-cb0c-4627-af7a-c85e702ba8cc"
```

Figure 1*:RDF code example of the SKOSified RCE Thesaurus term Aardewerk (pottery).*

Transformation of thesauri to OWL-Lite was performed with XSL templates. The templates produced new human-readable URIs based on a combination of a temporary base URI with the preferred label of individual entries. In order to comply with canonical URI definitions, the preferred labels were cleaned from illegal characters, such as ampersand, slash, etc., while spaces were replaced with underscores. The **dcterms:identifier**, due to its general purpose scope seemed an appropriate choice for holding the unique SKOS reference for individual entries in an OWL-Lite structure instead of the original **skos:Concept,** which is specific to thesauri not to ontology Parent/Child structures. The **rdfs:seeAlso** annotation property is used for holding the unique reference of the RCE node element while the Broader – Narrower term relationships implemented as Parent/Child relationship using **rdfs:subClassOf** structure (see below):

```
    <rdf:Description rdf:about="http://tmp/gemengd">
      <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
      <dcterms:identifier>http://www.rnaproject.org/data/b00374cd-fc7a-4fa8-b56c-
      a29c57e7a7fa</dcterms:identifier>
      <rdfs:seeAlso>http://www.rnaproject.org/data/e0983fd6-d0d9-4590-89a0-
      857922aee64c</rdfs:seeAlso>
     </rdf:Description>
    <rdf:Description rdf:about="http://tmp/gemengd">
      <rdfs:subClassOf rdf:resource="http://tmp/Materials"/>
     </rdf:Description>
```

The resulting OWL file, imported and parsed in GATE can be seen in the figure below.



```
Classes and Instances
  ArcheoArtefactTypes
  ArcheoComplexTypes
  Archeologische_perioden
  Locations
  Materials
    gemengd
    glas
    keramiek
    metaal
      brons
      gietijzer
      goud
      ijzer
      koper
      lood
      messing
      staal
      tin
      zilver
    monsters
    onbekend
      onbekend_beslag
      onbekend_heft_handvat
    organisch
      bot
        bot_aap
        bot_amfibie
        bot_geit
        bot_hond
        bot_kat
        bot_paard
        bot_reptiel
        bot_rund
        bot_schaap
        bot_schaap_en_of_geit
        bot_varken
        bot_vis
        bot_vogel
        bot_zoogdier_wild
      dierlijk
        bot_dierlijk
        gewei
```

```
Resource Information
  gemengd            gemengd
   URI               http://tmp/gemengd
   TYPE              Ontology Class
Direct Super Classes
  Materials          Materials
All Super Classes
  Materials          Materials
Direct Sub Classes
All Sub Classes
Equivalent Classes
Property Types
  comment            [ALL RESOURCES]
  identifier         [ALL RESOURCES]
  isDefinedBy        [ALL RESOURCES]
  label              [ALL RESOURCES]
  seeAlso            [ALL RESOURCES]
  versionInfo        [ALL RESOURCES]
Property Values
  seeAlso            http://www.rnaproject.org/data/e0983fd6-d0d9-4590-89a0-857922aee64c ✕
  label              gemengd ✕
  identifier         http://www.rnaproject.org/data/b00374cd-fc7a-4fa8-b56c-a29c57e7a7fa ✕
Instances
```

***Figure** 2: The OWL-Lite (ontology) structure of RCE Thesauri in GATE environment.*

2.3 RCE Thesauri Structural, Labelling and Coverage issues

The OWL-Lite ontology contributed vocabulary to an NER task aimed at six entity types of interest to archaeology (Artefact, Feature, Material, Monument Type, Place, Period). The resulting annotations were evaluated against a manually annotated corpus (the Gold Standard[17]). The Gold Standard (GS) evaluation revealed two main categories of problematic annotation behaviour; missed annotations (i.e. annotation existing in the GS but failing to be recognised by the pipeline) and spurious annotations (i.e. false positive annotations delivered by the pipeline but not defined in the GS). Spurious annotations are easier to tackle than missed annotations, mainly because they fall under two distinct categories:

- mentions of legitimate annotations that have been overlooked during GS definition;
- mentions of annotations that originate from less relevant parts of the ontology.

Mentions, as for example “vuursteen” (flint), may have been overlooked during GS definition due to their large number of occurrences in document. Thus, it is a matter of reviewing the GS and including such mentions to avoid delivery of false positive matches over legitimate annotations. Other cases, such as “gracht” (canal) may not be of archaeological interest and should be identified and excluded from rule matching.

The main performance drawbacks originate from missed annotation matches influenced by structural, labelling and coverage issues. These issues correlate to the operational behaviour of the ontology to deliver matches over “whole words”. For example, consider the entry “chopping tool”. A match is delivered only when the whole word is encountered, to avoid delivering arbitrary matches on “chopping” and “tool”[18]. This particular behaviour is affected by the labels of ontology classes, which in many cases have been defined with Information Science principles in mind, and not to support NLP operations. For example, the Artefact Type thesaurus contains a class labelled “pot/kookpot/voorraadpot”. A match will only be provided by the ontology when “pot kookpot voorraadpot” is encountered in text as a “whole word”, which is unlikely to happen. Clearly the three separate terms have been conjugated together to form a label within a thesaurus structure, and not to describe a specialised form of pot.

Structural issues relate to the definition of highly specialised classes that lack a parent (broader) class definition. The broader term is more likely to appear in text than the specialised entries. For example the term “nederzetting” (settlement) is absent from the Archeologische complextypen abr+ (Monument Types) thesaurus albeit the “Nederzetting met stedelijk karakter” (Settlement of urban character) and “Nietopgehoogde nederzetting zonder stedelijk karakter” (Settlement without urban character) are available.

Coverage issues relate to terms not included in thesauri but identified as being relevant by the GS definition. Such terms can be grouped into two distinct categories:

[17] Although useful for purposes of this formative exercise, there were issues of consistency with the pilot GS annotations which resulted in some correct machine annotations being unrecognised.

[18] Part word matches in GATE are not supported by ontologies but can be handled in gazetteers. However, the part matching mechanism should be used with caution as it delivers a large number of matches over irrelevant mentions that could harm Precision. The operation of part matches is useful for identifying the constituent parts of compound words, a common noun form in Dutch, such as “aardewerkfragment” (pottery fragment).

- terms for concepts that are totally absent from the resource such as, "bakstene" (brick wall);
- terms for concepts included in the resource, but are a synonym or a spelling variation, such as "laat-middeleeuwse" also written as "Late Middeleeuwen", "Laat Middeleeuwen" etc. (Late Medieval).

The structural, labelling and coverage issues of missed match results can be summarised in the following categories:

- **Broad Concept** missed matches, concerning relevant and frequently used terms such as "Artefact" and "Nederzetting" (settlement) which have not been included in thesauri but are implicitly mentioned by specialised terms. It is indicative that the top class of the Archeologische artefacttypen thesaurus is not "Artefact" but "Archeologische artefacttypen" which is not that helpful for NLP purposes.
- **Conjugated labels** of thesauri terms which in turn became ontology class names are not suitable for NLP purposes. For example "ciborium/hostiekelk", "pot/kookpot/voorraadpot" etc.
- **Compound** noun forms present a significant challenge for the ontology matching mechanism due to the "whole word" matching arrangement. Engagement of gazetteers operating on part matching can significantly help to overcome the "whole word" restriction as discussed in the section below. Compound noun forms can be:
    - of known constituent parts i.e. where all parts of the compound noun are existing in the resource, for example, "beslagplaat" where both "beslag" and "plaat" are available;
    - of part-known constituent parts where only one part is available in the resource e.g. "aardewerkmagering", where aardewerk (pottery) is available but "magering" is not.

  Note that thesauri also contain clear "individual" entries of compound terms e.g. "architectuurfragment". These cases do not present any difficulty in matching, and are not affected by the "whole word" matching arrangement, since they are the "whole word".
- **Synonyms and Spelling variations** - coverage of thesauri resources varies. For example the Archeologische artefacttypen thesaurus has greater coverage for synonyms and spelling variation entries than the Perioden thesaurus.
- **Non-available,** are those terms (concepts) that have been identified by the GS as being relevant but are not available in the resource. Non-available term cases are different from the above case of missed matches, as it is a concept that is totally absent from the resource, not just a spelling variation or a synonym.

2.4 Proposed Actions for Improving the Matching Result

Several actions can be taken for the next iteration and for adapting the ontology resource to the requirements of the NLP task. Some actions that can significantly improve the matching result are easy to implement. Other actions will require a possible structural modification of the resource for NLP purposes, relating to inclusion of new concepts (not just synonyms or spelling variations).

Easy and straight-forward actions which can have a significant impact on matching result are:

- Enhancing (a subset) of existing thesauri terms with synonyms and spelling variations. The GS has already revealed cases that require such enhancement. For example many Period terms such as, “Laat-Middeleeuwse” can be enhanced with spelling variations like “Laat Middeleeuwse” “Late Middeleeuwen” and Late-Middeleeuwen. It is not necessary to examine every single thesauri entry for synonyms and spelling variations. Enhancing the most frequent cases, most of which are revealed in GS evaluation, will have a significant impact in recall.
- Breaking apart conjugated labels into their constituent parts. This action does not require modification of the structure of the resource, but only enhancing the existing classes with alternative labels. So for example the “pot/kookpot/voorraadpot” class will be enhanced with three alternative labels i.e. pot, kookpot, voorraadpot (in the fashion of spelling variations).
- Enhancing broad (non NLP friendly classes) e.g. Archeologische artefacttypen with their NLP friendly equivalent, in this case “Artefact”.

Medium scale actions will require input from Dutch archaeology domain experts:

- Identify the most frequent cases of terms that contribute to compound noun forms. It will not be efficient to produce part-matches via gazetteer from the totality of the “Archeologische artefacttypen” thesaurus, as this will have an impact on precision (generating too many part matches). Instead a selected set of terms that frequently appear in compound forms should be identified and exposed as gazetteer list. For example “aardewerk” has much more chance of appearing as a compound noun than other terms, so it should be prioritised for part-matching.
- Identify entity-type combinations that deliver compound noun forms. Based on the GS results, it appears there are three main combination types a) material+artefact, b) period+artefact and c) artefact+artefact. It would be helpful to discuss these combination forms with Dutch archaeologists before resolving on any NLP matching rules.
- In addition to the above, the annotation approach towards compound entity forms should be discussed and finalised. At this stage it is not clear the number and type of annotations that should be delivered from a compound entity form. For example consider the case of “aardewerkfragment” (pottery fragment). Will it deliver:
    - a single span annotation (aardewerkfragment) associated with two SKOS references one for “aardewerk” and another for “fragment”;
    - two separate annotations each associated with a SKOS reference;
    - three annotations, two separate annotations (as above) and a third for the whole span annotated as “P45.consists_of” property.

- Similarly, some thought should be given towards the annotation of compound entity forms of part-known constituents, where only one of the parts is “known” to the ontology. For example “aardewerkmagering”, or how “magering”, which is not a known (available) term, will be treated. Will it just be ignored and only a single annotation will be delivered i.e. “aardewerk”, or will it be included in the annotation span (which is also possible).
- 'Expert annotator' review of the existing GS for consistency and in light of the automatic output results.

Complex actions will require input for Dutch archaeology domain experts:

- Actions concerning adding new thesauri concepts, and releasing respective SKOS references.
- Rearranging a thesaurus structure for adding new broader terms for a set of specialised terms already included in the resource e.g. “Nederzetting” (settlement).
- With regards to the above, a quick fix for NLP purposes which would not require restructuring the resource, could be adding an alternative label of the broad term to the existing specialised terms e.g. “Nederzetting met stedelijk karakter”. Or to use a general purpose thesaurus that contains the broad term (Nederzetting), such as Erfgoedthesaurus Objecttypen for delivering matches with SKOS reference respective to the broad term not to a specialised term.

3 Early Pilot Evaluation Results: Dutch Named Entity Recognition Pipeline

3.1 Introduction

The section presents the results of the early pilot evaluation of the Dutch NER pipeline based on the input of a single manually annotated document (ARA81_1R_Stichtse Kant bedrijventerrein). It also presents the results of the vocabulary transformation task from spreadsheets to RDF/XML hierarchical structures, expressed as OWL-Lite (ontology) for the purposes of an ODIE task. Observations relate to the vocabulary transformation process and pipeline results, and reveal initial issues that affect vocabulary usage and focus of the NER exercise.

The following vocabulary resources (stored in spreadsheet format) were made available by University of Leiden:

- gemeenten.xls (list of counties)
- BAG_woonplaats.xls (list of cities and towns with their respected municipalities)
- keywords_periode.xls (list of archaeological dates)
- keywords_materialen.xls (list of materials)
- ABR_artefacttypes_combinatie.xls (list of artefact types)
- keywords_grondsporen.xls (list of archaeological features – contexts)
- keywords_complextypes.xls (list of monument types)

The structure and hierarchical information of the vocabularies (spreadsheets) contain inconsistencies. For example the **keywords_materialen** file contains a three-level hierarchical structure of the materials arranged in classes and sub-classes. On the other hand, the **keywords_artefacttypes.xls** file is largely a flat list of artefact types, with few sub-classes. Additional points of inconsistency relate to the use of synonyms and plural versions of the vocabulary. Some files contain rich information with respect to synonyms, such as the **keywords_artefacttypes.xls** but some others contain none, such as the **keywords_complextypes.xls**.

The transformation process made use of the available information contained in files in order to deliver a unified ontological structure covering all vocabulary resources. The resulting ontological classes contain (whenever possible) labels relating to synonyms, plural versions and unique reference codes for the vocabulary entries.

The resulting vocabulary structure (referred as the ontology) contains six top classes, and a large number of sub-classes and sub-sub-classes. The ontology uses a class/sub-class structure for accommodating the volume of vocabulary terms. There is no use of ontology instances/individuals, as all vocabulary terms are expressed as classes of the structure. The six top level classes of the ontology are the following: Artefact, Complex, Grondspoor, Materiaal, Periode, Plaats.

3.2 The Artefact Structure

This particular resource presented significant challenges in transformation. The spreadsheets contained a flat list of artefact types with little information about their hierarchical arrangement. The author imposed a class/sub-class structure based on the information held in the code_alg (main code) column. For example the term “amulet” having main code AMULET has been used as the parent class of all other vocabulary entries (alsengemme, dier amulet, fetisj amulet, godheidsamulet, etc) that also have AMULET as their main code. Whenever possible, terms were grouped under a parent class of the same main code (code_alg), however, there are many "shallow" classes (with no sub-class) in the structure which can potentially be grouped under a parent class. The resulting ontology classes have unique reference codes (from the code_spec column) and alternative labels from the synoniem1, synoniem2, synoniem3 and synoniem4 columns.

The Complex Structure

Transformation of this resource did not impose any significant challenges. The vocabulary resources were grouped under the parent classes of the "groep" column. Classes and sub-classes of the structure have unique reference codes (code column). There original resources did not contain any information relating to synonyms and alternative labels.

The Grondspoor Structure

The original spreadsheet had a structure similar to artefact resource (i.e. a flat list with main and sub-codes assigned to individual entries). The vocabulary terms of the original spreadsheet were grouped under parent classes based on the code_alg assignment. The resulting ontology classes have unique reference codes (from the code_spec column) and alternative labels from the synoniem, meervoud and verkleinwoord columns.

The Materiaal Structure

Transformation of this resource did not impose any significant challenges. The original spreadsheet contained a three-level hierarchical structure (described in columns materiaal1, materiaal2, materiaal3!). The transformation process followed the hierarchical structure already described in the spreadsheet. The resulting ontology classes have a unique reference code (from the code column) and alternative labels from the synoniem1, synoniem2, synoniem3, and bnw columns.

The Periode Structure

Transformation of this resource did not pose any significant challenges. The original spreadsheet contained a four-level hierarchical structure (described in columns datering1, datering2, etc.). The transformation process followed the hierarchical structure already described in spreadsheet. The resulting ontology classes have unique reference codes (from the code column) and alternative labels from the synoniem1, synoniem2, and bnw columns.

The Plaats Structure

This particular structure is the combination of the woonplaats and gemeenten speadsheets. The woonplaats spreadsheet contains city (wooplaats) and municipality (gemente) entries where each city entry is assigned to a municipality. The gemeenten spreadsheet contains a flat list of counties (Provincie). A Google search was conducted for each municipality entry of the wooplaats spreadsheet in order to assign a county to each municipality. The resulting structure has a three-level hierarchy County → Municipality → Town. There are several cases in the Plaats structure where the same "name" is assigned to a County, Municipality and Town (e.g. Utrecht), such cases are distinguished by URIs. For example:

http://www.semanticweb.org/archeologisch_basis_register/obie/provincie/utrecht

http://www.semanticweb.org/archeologisch_basis_register/obie/gemeente/utrecht

http://www.semanticweb.org/archeologisch_basis_register/obie/woonplaats/utrecht

All other ontology classes follow a URI form like:

http://www.semanticweb.org/archeologisch_basis_register/obie/<class name>

The choice of the URI is temporary and helps in constructing the resource at a particular phase of development. There is no use of the # in the URI. OBIE stands for Ontology-Based-Information-Extraction.

3.3 Evaluation

An early evaluation of the pilot Dutch NER pipeline conducted with respect to a single manually annotated document (ARA81_1R_Stichtse Kant bedrijventerrein) was carried out. The document was annotated with respect to the following entities; Actor, Place, Monument, Archaeological Context, Artefact, Material, Period. The NER pipeline is configured to identify the following entities: Place, Physical Thing (i.e. Monument), Physical Object (i.e. Artefact), Time Appellation (i.e. Period), Material, Context. The annotations produced by the pipeline were chosen to help align them to the CIDOC-CRM ontology. Hence, Physical Thing is used instead of Monument, Physical object instead of Artefact, and Time Appellation instead of Period. Some early observations can be made following the initial evaluation task.

Actor

The pipeline is not configured to identify Actor matches. In order to scope Actor the pipeline will need to be equipped with a relevant vocabulary resource. However, from the manual annotations it is not clear why Place name instances are annotated as Actor, for example “Gemeente Almere”.

Place

The Precision of the pipeline with respect to Place entity recognition is reasonably good considering the pilot stage of the development (around 50%). However, Recall is low (26%). This is due to a) limited vocabulary coverage (eg Cirkelbos is not included in the spreadsheet) b) annotation of grid references (eg 149.470/481.309) which are not currently targeted by the pipeline, but it is possible to be targeted by rules c) use of Part-of-Speech tagger for matching cases which are tagged as Noun (e.g. Almere, matched in some cases and missed in others). The Noun restriction can be lifted if it causes more harm than good.

Archaeological Context / Feature

The sample of the manually annotated cases is limited to a single case, hence Recall is 100%. On the other hand, Precision is low at around 25%. It may be that grondsporen, sloten, grachten have been overlooked by manual annotation, or there is a stronger case for excluding such cases from annotation.

Material

The overall pipeline performance for this type of entity is reasonably good considering the pilot stage of the development (72% Recall, 44% Precision). Again the Noun restriction is responsible for a couple of missed cases. There are several cases of False Positives (i.e. cases recognised by the pipeline but not annotated manually, such as vuursteen, stenen, bron. Why these cases have not been included in manual annotation should be explored.

Physical Object (Artefact)

The performance was very problematic for both Recall and Precision (under 10%). This is due to:

- vocabulary coverage (scheepsresten, bewerkingsafval and other) not included in spreasheets;
- delivery of too many false positives (kans, NAP, boor, schaal) which might have been overlooked during manual annotation or original resource contains "noise" (irrelevant terms) held in synonyms; and
- configuration of the pipeline matching mechanism to match only whole vocabulary entries (e.g. bewerkt is not matched because entry is bewerkt steen).

This particular pipeline configuration can be altered to deliver matches from the parts of an entry (e.g. bewerkt-steen) in expense of delivering (possibly) even more false positives.

Physical Thing (Monument)

Results were the same as above; low performance due to the same matching behaviour and results.

Time Appellation (Period)

The overall Precision is better (around 50%) but Recall is poor (around 20%). This is due to:

- too many Partially Correct Matches (e.g. Vroege-Steentijd instead of Midden- en Vroege-Steentijd). From experience working with English grey literature we know that authors use time moderators (later, earlier, mid) quite flexibly. Such moderators are combined with core period descriptions, resulting in compound Time Appellation cases (such as mid early Roman period) where the vocabulary contains just mid Roman or early Roman. Information Extraction rules can be constructed to tackle such cases. Also decision will need to be made about how to annotate compound time appellation cases (e.g. late Roman to Early Medieval), as to whether a single annotation span or two separate spans be delivered;
- use of numerical dates (e.g. 5600 BP), as the current pipeline is not configured to address such cases. Rules can be constructed but we will need to know what moderators are used in Dutch (e.g. BP, voor Christus etc);
- more particular to Dutch than English is the partial annotation of time appellations within a longer string. For example steentijdbewoners where only steentijd is manually annotated. This can be very challenging IE task, mainly because enabling partial matching (wild cards) would open the door to any sort of partial matches or "noise". If there is a Dutch grammar rule upon which some form of partial matching can be based, then it may be possible to match such cases without causing much damage to the system's precision.

4 Machine Learning Applications for the ADS Grey Literature Library

4.1 Introduction

The ADS holds a large corpus of unstructured data in its archives, often referred to as text. This unstructured data can be found in our Grey Literature Library (GLL)[19], our journal back runs, and within general reports, which are a typical component of any archaeological project. This is the case for the vast majority of archaeological reports held around the world. Within ARIADNE, ADS is using the GLL as a resource to develop tools and procedures to help the archaeological domain better access this vast source of largely untapped digital data.

One of the major challenges currently facing the archaeological domain is how to aid users in retrieving relevant data held within these unstructured reports. Traditional information retrieval systems are based on what is referred to as a "bag-of-words" representation, where documents are retrieved by lexical matching (i.e. string matching) using a query to find terms within documents. Due to synonymy (similar meanings) and polysemy (multiple, related meanings), string matching methods often produce imprecise or irrelevant results. Therefore, it would be hugely beneficial to provide better tools for users to search and discover content within this large body of unstructured data. In order to achieve this goal, useful metadata must be obtained from this unstructured data. Traditionally, metadata is created manually, which is extremely time consuming and expensive, so the ADS is attempting to develop a user-friendly web application that will utilise NLP techniques to automatically produce resource discovery metadata (i.e named entities), which can feed into existing data management systems, and improve data retrieval performance.

The ADS first worked with NLP techniques to automatically extract resource discovery metadata from unstructured data as part of the Archaeotools project. Archaeotools was a collaborative project between the ADS and the University of Sheffield's NLP Research Group which ran from 2007-9. The aims of the Archaeotools project were to:

- index the over one million metadata records held by the ADS, describing sites and monuments in the UK, according to the criteria of what, where and when,
- to employ NLP to allow automated tools to search within documents for terms which are part of known classification schemes, adding them to the ADS facetted index, and provide better access to grey literature,
- to explore tools to allow users to impose their own classifications, and index the documents according to their own criteria, adding further user-defined dimensions to the classification,
- and investigate whether it is possible to identify and harvest index terms within older antiquarian literature, such as recently digitised back runs of archaeological journals.

Unfortunately, the results of the Archaeotools project did not reach the unrealistically high expectations for the technology. Essentially, archaeological data proved to be far more complex and varied than was initially anticipated by the project researchers. Despite this result, the NER module developed within Archaeotools, which recognises specific classes of entities in text, such as place names, turned out to be very useful for making sense of large bodies of unstructured archaeological data, and some of the outputs were integrated into ADS systems.

[19] ADS (2015) Library of Unpublished Fieldwork Reports (Grey Literature Library). http://archaeologydataservice.ac.uk/archives/view/greylit/

Most of the NLP tools from Archaeotools were set aside and not investigated again until recently. With time and experience, ADS now understands how the results from the NLP within Archaeotools can be made more useful. Furthermore, new resources, such as SKOSified controlled vocabularies used for classification, developed through the Semantic ENrichment Enabling Sustainability of arCHAeological Links (SENESCHAL) project, led by the Hypermedia Research Unit at the University of South Wales, can now be used to further enhance NLP and the creation of resource discovery metadata. The primary lesson learned from Archaeotools was that the usefulness of NLP is not in creating a human-like understanding of unstructured data, but in helping index unstructured data, which can minimise the task for a human researcher.

4.2 Aims and Objectives

As part of ARIADNE, the ADS is building upon the work and lessons learned from the Archaeotools project, to further develop NLP tools and help the archaeological domain better access the vast resource of unstructured digital data available to archaeologists in the form of text. This text typically exists in PDF, MS Word, or plain text files within the ADS GLL, digitised journal collections, and reports deposited within project archives.

This will be achieved through developing and evaluating machine learning-based NLP techniques and integrating them into a new metadata extraction web application, which will take previously unseen English language text as input, and identify and classify named entities within the text. The outputs will then be used to enrich the resource discovery metadata for existing and future resources. The final application will include a web-based, user friendly interface that can be used by archaeological practitioners to automatically generate metadata related to uploaded, text-based content on a per file basis, or using batch creation of metadata for multiple files.

4.3 Work to Date

The creation of the web application to automatically generate useful archaeological metadata will be based on using NLP techniques to generate the metadata outputs. Based on the successful results from the Archaeotools project with NER, ADS has focussed on developing an effective NER module for the web application to generate the metadata outputs, and has concentrated on exploring additional techniques needed to refine these outputs. NLP techniques such as automatic summarisation and text clustering were also explored. These techniques may be used to add additional functions to the web application.

4.3.1 Training Data

In order to create the NER module for the web application, training data was first required to train the classifiers used. The training data uses annotation to teach the classifier rules that apply to selected concepts. The following concepts were mapped:

<Subject>	topics covered, finds mentioned (e.g. Bronze Ring)
<Placename>	place names related to events, sites and finds (e.g. King’s Manor)
<Temporal>	archaeological dates of interest, e.g. Prehistoric or 800AD
<Grid reference>	grid reference
<Title>	‘Title’ of the given document
<Author>	Document author

Two sets of training data were used. One was produced by human annotators; the other using a rule-based machine annotator. The training data is simply plain text, with XML style tags around the relevant properties, and offsets of the entities that were recorded. Below are examples of a file annotated by a rule-based machine annotator and a human annotator. From these examples it can be seen that the human annotator is much more accurate that the rule-based machine annotator.

Rule-based machine annotator:

> *<placename>alluvial deposits</placename> associated with the River Lee. 4 HISTORICAL AND ARCHAEOLOGICAL BACKGROUND 4.1 No Desk Based Assessment (DBA) has been prepared for the site and no previous archaeological works have been undertaken on the site. 4.2 The site is located on the edge of the floodplain of the River Lea, an area with high potential for <temporal>prehistoric</temporal> activity.4.<subject>3 Finds</subject> in the Leyton area suggest a <temporal>Roman</temporal> <placename>settlement</placename> in the south west of the area. A <temporal>Roman</temporal> <placename>road</placename> probably ran north from London on the line of Leytonstone High Road (Weinreb and Hibbert, 1983).*

Human annotator:

> *Bedfordshire County Council has granted planning permission (2005/39) for alterations and an extension to form a new classroom and music rooms at <placename>Leighton Middle School</placename> <placename>Leighton Middle School</placename> lies in an archaeologically sensitive area, at the western end of the historic core of <placename> Leighton Buzzard</placename> within the town centre Conservation Area. It is bounded by <placename>Bridge Street</placename> to the north and <placename>Church Square</placename> to the east. During <temporal>April 2006</temporal> Albion Archaeology carried out a programme of <subject>fieldwork</subject> on the site of the new classroom to mitigate the archaeological impact of the development. A series of <temporal>post-medieval/modern</temporal> remains were recorded within the development area. These consist of <subject>land boundaries</subject> marked by <subject>ditches</subject>*

The human annotated training data was hand crafted as part of the Archaeotools project by archaeological domain experts. Thirty full-length UK archaeological reports were specially selected for this exercise. The reports varied from five to 120 pages in length, with a total of 225,475 words, resulting in over 5000 annotations for the various entities. There was discussion as to whether it would be useful to create more training data in the context of the ARIADNE project, but it was concluded that there would be an exponentially decreasing benefit relative the amount of work that would be required. However, the ability to annotate new documents was included in the development of the web application.

4.3.2 Classifiers

Following the creation of the training data, or in this case the adoption of training material from the Archaeotools project, the training data needed to be applied to a classifier. A classifier is a machine learning tool that will take data items and place them into classes resulting in a statistical model, which is used to extract entities from entered text. Two classifiers were tested, the Linear Support Vector Machine (SVM) Classifier and the Conditional Random Field (CRF) algorithm, to compare the results and see if one performed better than the other.

The Linear SVM classifier

To verify the implementation of the SVM classifier, it was applied to the human annotated training material. Initially the rule-based training data was also used by the SVM classifier, but because the rule-based annotations were not as accurate as the human annotations, the classifier outputs were also inaccurate. Therefore, the rule-based training data was removed from the SVM classifiers training material. The rule-based system needs more refinement to be used effectively, and further investigation may be done later. The following categories were examined by the classifier:

<subject>	archaeological subject
<temporal>	temporal
<placename>	location
<title>	title
<author>	author

To train the Linear SVM classifier, a window size of five surrounding tokens and the following feature set was used:

- Morphological root of the token
- Exact token string
- Orthographic type (e.g. lowercase, uppercase)
- Token type (e.g. number, word)
- Archaeological Gazetteer (terms existence in)

The training process was repeated several times before it reached its convergence.

CRF classifier

Using the same procedure as with the Linear SMV classifier, the CRF classifier was applied to the human annotated and rule-based training material. As with the Linear SVM classifier the outputs from the rule-based training data were inadequate for use. The following categories from the human annotated training material were examined:

<subject>	archaeological subject
<temporal>	temporal
<placename>	location
<title>	title
<author>	author
<contact>	contact details

To train the CRF classifier, a window size of five surrounding tokens and the following feature set was used:

- N-Grams with max length of six tokens (i.e. contiguous sequence of words)
- Exact token string
- Features from previous word class sequence
- Archaeological Gazetteer

The training process was repeated several times before it reached its convergence.

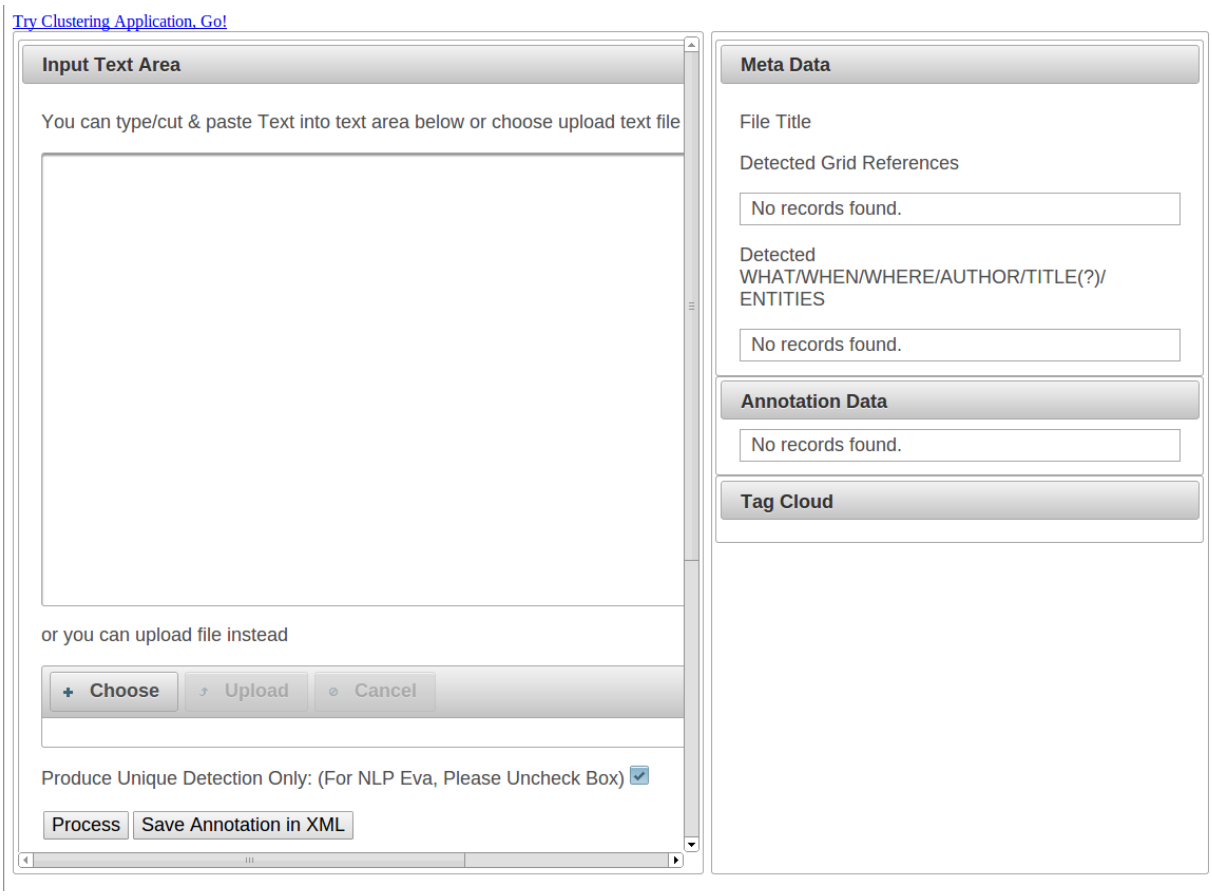
The performance of the classifiers was evaluated using the following measures: Precision, Recall, and F-Measure (a measure of accuracy calculated from the Recall and Precision measurements). The evaluation of classifier performance is typically done using a quantitative method, but as no appropriate Gold Standard or evaluation tool currently exists for the archaeological domain, a more qualitative method was used.

Archaeological domain experts were asked to read a sample of the documents, and were then shown a list of entities extracted from the documents using the classifiers. They were then asked how relevant the concepts, subjects and locations that had been extracted were to the documents. This evaluation found that the entities extracted from the documents were all terms found in the document. However, there were some erroneous entities (spelling mistakes, pluralisations, punctuation marks) and from an archaeological mindset some terms extracted were considered less important than others. However, from a NLP viewpoint the classifiers successfully 'learned' from the training data. Of the two classifiers the CRF classifier was chosen, as it was easier to implement into the web application and required less computing time to produce results.

The models built by the classifier with gazetteers from the SENESCHAL project were then directly applied to the unseen data from grey literature reports. As there is currently no Gold Standard for archaeological grey literature, a group of reports from the North Yorkshire region (knowing there had not been previous training on grey literature from a North Yorkshire dataset) were chosen and manually scored. The gazetteers were especially useful for improving extraction performance, when applied to more unseen corpora. This confirmed there is substantial overlap of information from various corpora within the grey literature.

4.3.3 Web Application

A prototype web application interface is currently under development. The purpose of this application is to allow domain experts to annotate reports, generate resource discovery metadata where none exists, and generate metadata which can be used to further train the classifiers. The application was designed to allow for text to be entered into an "input text area", or a file (PDF or DOC) to be uploaded to the application. When using the latter option the system will extract text out of the PDF or DOC automatically and display it in the 'input text area'. A screenshot of the system can be seen below.



You can type/cut & paste Text into text area below or choose upload text file

or you can upload file instead

Produce Unique Detection Only: (For NLP Eva, Please Uncheck Box) ☑

File Title

Detected Grid References

No records found.

Detected WHAT/WHEN/WHERE/AUTHOR/TITLE(?)/ ENTITIES

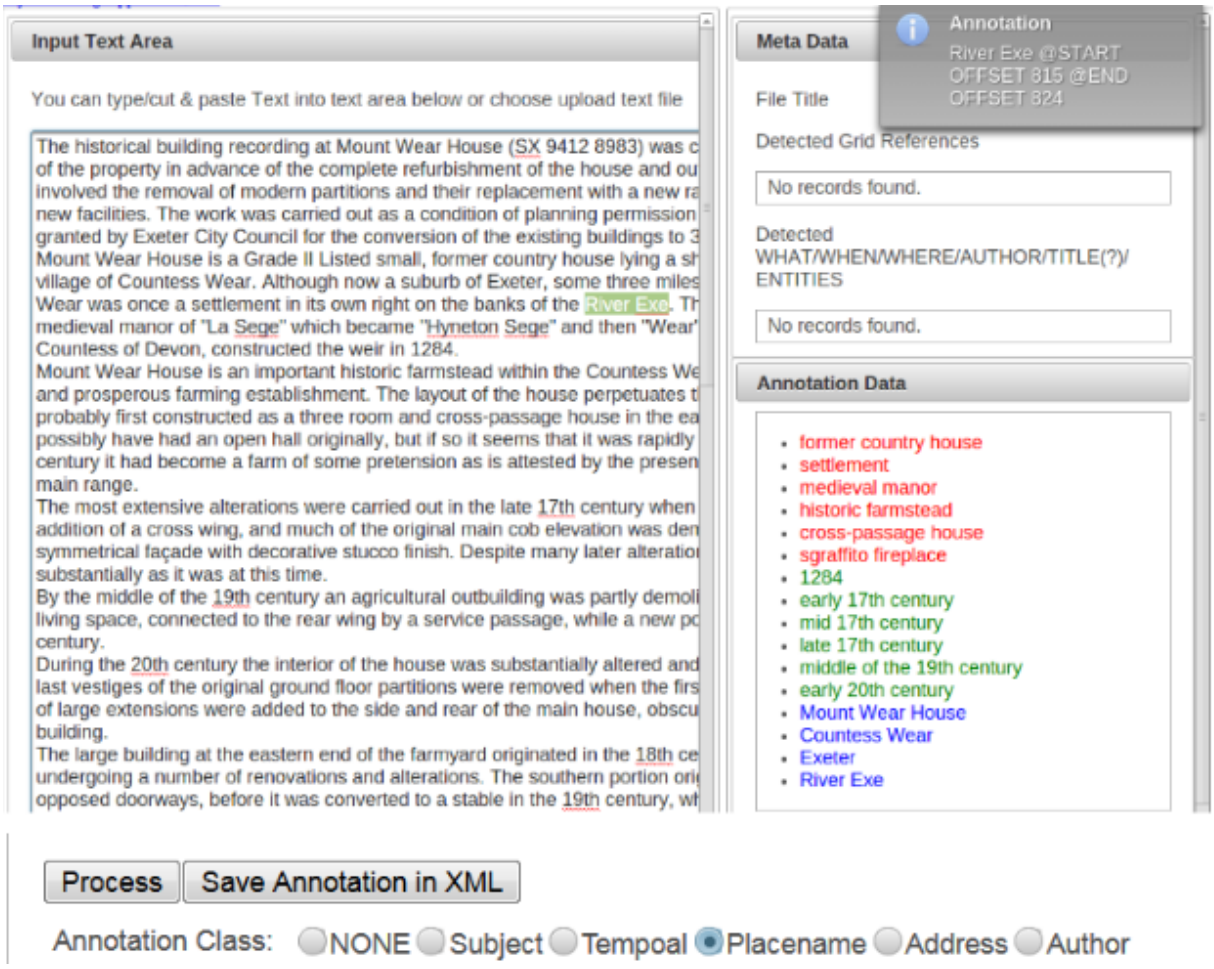
No records found.

No records found.

Figure 3: Input screen shot of the ADS prototype web application.

The performance of the NER integrated into the web application is somewhat dependent on the quality of the content extraction module, where plain text is extracted from the PDF or DOC. During the extraction process the format and structure of the document, which may provide valuable information for the identification of entities by the NER module, is largely lost. Experiments have shown that the cleaner the content the better the result, therefore, by improving the performance of the content extraction module, it can be expected that the NER performance can also be improved. Improvement of this extraction module will be investigated later in the project.

An integrated annotation tool has also been included in the web application to assist human annotators in producing additional training data which, in turn can be used to retrain the CRF classifier used for better performance. This tool means a user can upload a document or text, and then go through the process of annotating the document by selecting appropriate classes as shown in the figure below. The screenshot below shows a sample of the annotation process, with an annotated entity, the notification from the application, and a list of the annotated data on the right.



Input Text Area

You can type/cut & paste Text into text area below or choose upload text file

The historical building recording at Mount Wear House (SX 9412 8983) was c
of the property in advance of the complete refurbishment of the house and ou
involved the removal of modern partitions and their replacement with a new ra
new facilities. The work was carried out as a condition of planning permission
granted by Exeter City Council for the conversion of the existing buildings to 3
Mount Wear House is a Grade II Listed small, former country house lying a sh
village of Countess Wear. Although now a suburb of Exeter, some three miles
Wear was once a settlement in its own right on the banks of the River Exe. Th
medieval manor of "La Sege" which became "Hyneton Sege" and then "Wear'
Countess of Devon, constructed the weir in 1284.
Mount Wear House is an important historic farmstead within the Countess We
and prosperous farming establishment. The layout of the house perpetuates t
probably first constructed as a three room and cross-passage house in the ea
possibly have had an open hall originally, but if so it seems that it was rapidly
century it had become a farm of some pretension as is attested by the presen
main range.
The most extensive alterations were carried out in the late 17th century when
addition of a cross wing, and much of the original main cob elevation was den
symmetrical façade with decorative stucco finish. Despite many later alteratio
substantially as it was at this time.
By the middle of the 19th century an agricultural outbuilding was partly demoli
living space, connected to the rear wing by a service passage, while a new pc
century.
During the 20th century the interior of the house was substantially altered and
last vestiges of the original ground floor partitions were removed when the firs
of large extensions were added to the side and rear of the main house, obscu
building.
The large building at the eastern end of the farmyard originated in the 18th ce
undergoing a number of renovations and alterations. The southern portion ori
opposed doorways, before it was converted to a stable in the 19th century, wh

Meta Data

Annotation
River Exe @START OFFSET 815 @END OFFSET 824

File Title

Detected Grid References

No records found.

Detected WHAT/WHEN/WHERE/AUTHOR/TITLE(?)/ ENTITIES

No records found.

Annotation Data

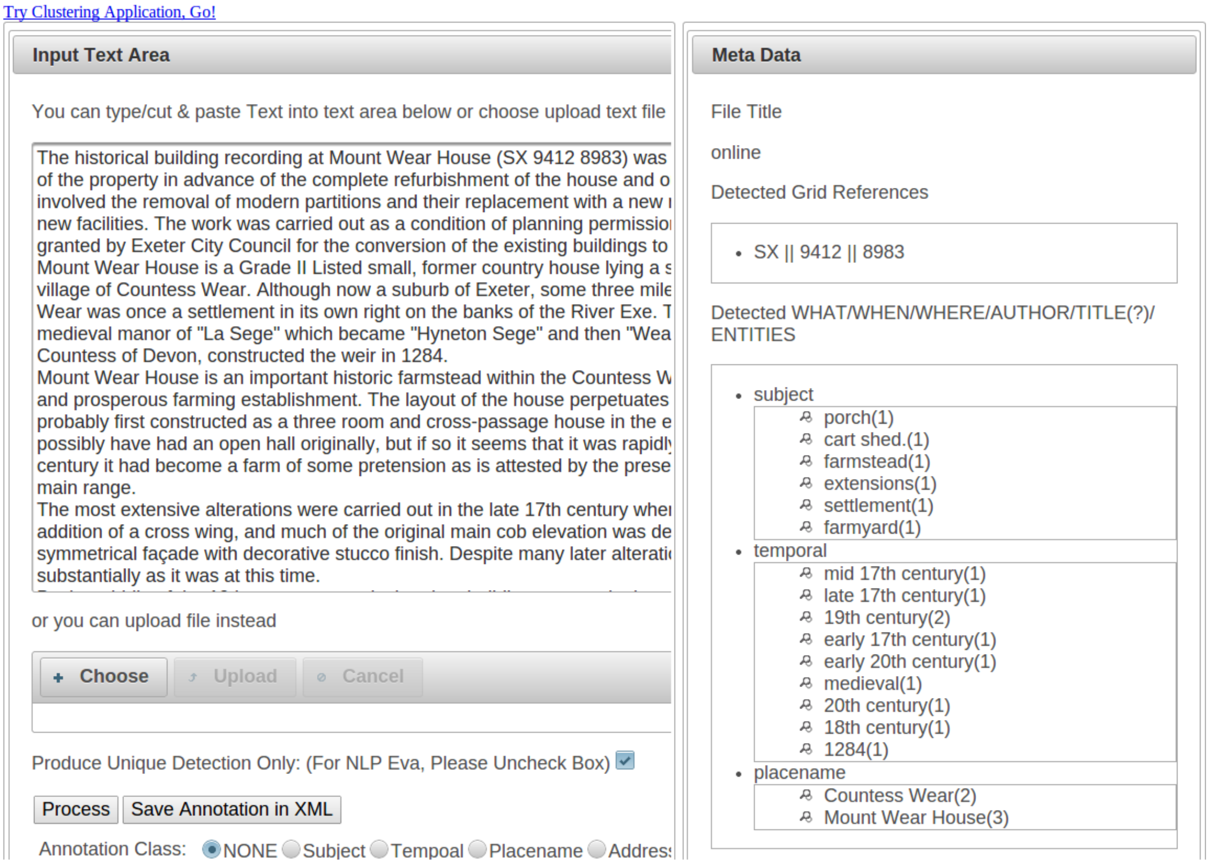
- former country house
- settlement
- medieval manor
- historic farmstead
- cross-passage house
- sgraffito fireplace
- 1284
- early 17th century
- mid 17th century
- late 17th century
- middle of the 19th century
- early 20th century
- Mount Wear House
- Countess Wear
- Exeter
- River Exe

Annotation Class: ○ NONE ○ Subject ○ Tempoal ● Placename ○ Address ○ Author

Figure 4: Screen shot of the annotation process in the prototype ADS web application.

This will be an extremely useful feature which can be used to produce more training data in the future, and also provide an intuitive interface for users to correct results which can then be used by the training classifier.

To extract the possible metadata from the uploaded documents, an NER module was created for the web application. A simple Java application was written that utilised the CRF classifier. When text is entered into the “input text area” entities are extracted from the text using the NER module based on the CRF classifier. The extracted entities are displayed as suggested metadata to the right of the entered text where users can assess the relevance of the extracted entities. The web application can also detect and extract UK grid references using manually crafted regular expressions. Extracted grid references are automatically verified using UK Geospatial data held within an Oracle Spatial database, where incorrect grid references can be filtered out from the result. A sample of the web application outputs can be seen when the text from the introduction of a randomly chosen archive (DOI:10.5284/1027059) is pasted into the “input text area”. The results generated can be seen in the image below in the right-hand column. By clicking on the magnifying glass icons beside each entity generated, a user can jump directly to the word in the text from which the result was derived.



Try Clustering Application, Go!

Input Text Area

You can type/cut & paste Text into text area below or choose upload text file

The historical building recording at Mount Wear House (SX 9412 8983) was
of the property in advance of the complete refurbishment of the house and o
involved the removal of modern partitions and their replacement with a new
new facilities. The work was carried out as a condition of planning permissio
granted by Exeter City Council for the conversion of the existing buildings to
Mount Wear House is a Grade II Listed small, former country house lying a s
village of Countess Wear. Although now a suburb of Exeter, some three mile
Wear was once a settlement in its own right on the banks of the River Exe. T
medieval manor of "La Sege" which became "Hyneton Sege" and then "Wea
Countess of Devon, constructed the weir in 1284.
Mount Wear House is an important historic farmstead within the Countess W
and prosperous farming establishment. The layout of the house perpetuates
probably first constructed as a three room and cross-passage house in the e
possibly have had an open hall originally, but if so it seems that it was rapidly
century it had become a farm of some pretension as is attested by the prese
main range.
The most extensive alterations were carried out in the late 17th century whe
addition of a cross wing, and much of the original main cob elevation was de
symmetrical façade with decorative stucco finish. Despite many later alterati
substantially as it was at this time.

or you can upload file instead

Produce Unique Detection Only: (For NLP Eva, Please Uncheck Box) ☑

Annotation Class: ● NONE ○ Subject ○ Tempoal ○ Placename ○ Addres

Meta Data

File Title

online

Detected Grid References

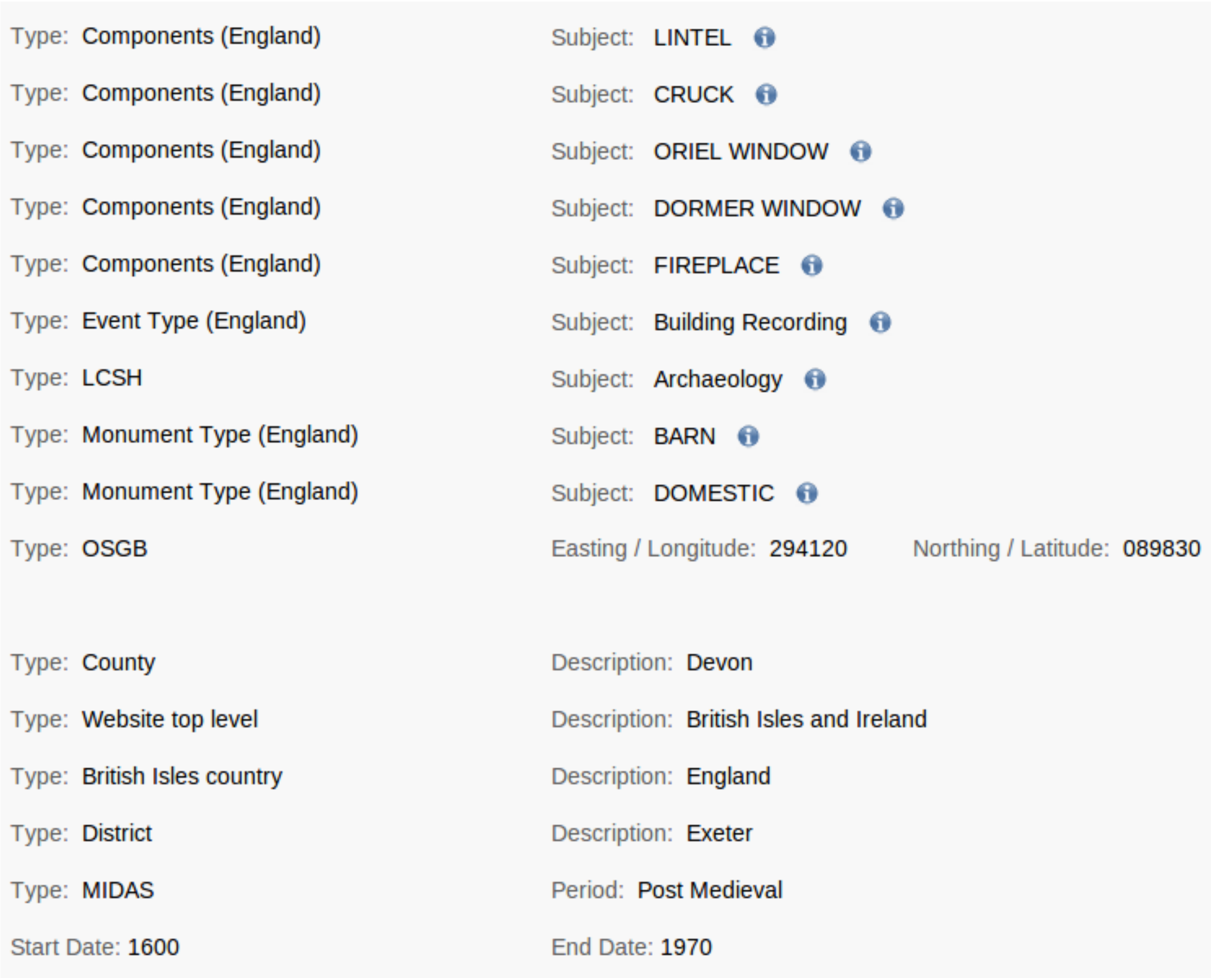
- SX || 9412 || 8983

Detected WHAT/WHEN/WHERE/AUTHOR/TITLE(?)/ ENTITIES

- subject
  - porch(1)
  - cart shed.(1)
  - farmstead(1)
  - extensions(1)
  - settlement(1)
  - farmyard(1)
- temporal
  - mid 17th century(1)
  - late 17th century(1)
  - 19th century(2)
  - early 17th century(1)
  - early 20th century(1)
  - medieval(1)
  - 20th century(1)
  - 18th century(1)
  - 1284(1)
- placename
  - Countess Wear(2)
  - Mount Wear House(3)

Figure 5: Screen shot of the prototype ADS web application showing entities extracted from the text.

By comparing these outputs of the web application tool to the metadata for the same archive provided by the depositor (see figure below), it can be seen that the identified entities are relevant to the archive. There are, however, some noticeable differences between the metadata provided and the extracted entities. It should be noted though that the introductory text does not represent a complete archive unlike the depositor metadata. The "subject" terms returned by the tool can be seen to be similar to the depositors metadata, although they contain less detail than the original metadata. Conversely, the "temporal" terms are more specific, with multiple period terms extracted, whereas the depositor metadata has combined these multiple period terms into a single "post medieval" term. The "locational" terms extracted by the web application were also on a more specific local level, whereas the metadata provided by the depositor included a full locational hierarchy. However, when a correct local placename is extracted from the text, higher level locational metadata can be automatically extrapolated using Linked Open Data at a later stage, during the ingestion of the metadata into the ADS Collections Management System.



Type: Components (England)	Subject: LINTEL	
Type: Components (England)	Subject: CRUCK	
Type: Components (England)	Subject: ORIEL WINDOW	
Type: Components (England)	Subject: DORMER WINDOW	
Type: Components (England)	Subject: FIREPLACE	
Type: Event Type (England)	Subject: Building Recording	
Type: LCSH	Subject: Archaeology	
Type: Monument Type (England)	Subject: BARN	
Type: Monument Type (England)	Subject: DOMESTIC	
Type: OSGB	Easting / Longitude: 294120	Northing / Latitude: 089830
Type: County	Description: Devon	
Type: Website top level	Description: British Isles and Ireland	
Type: British Isles country	Description: England	
Type: District	Description: Exeter	
Type: MIDAS	Period: Post Medieval	
Start Date: 1600	End Date: 1970	

Figure 6: Example of metadata in the ADS Collection Management System.

The entities extracted by the NER module using this method (using a relatively short piece of text) specifically composed to provide an introductory overview of an archive), produces very successful results, and the relatively small number of entities are easy to view and manage within the web application by a user. This becomes more complicated when tested with a larger body of text.

The following outputs were returned when a randomly chosen PDF report: (DOI:10.5284/1027271) was selected from the ADS Grey Literature Library and uploaded to the web application to be processed. In this case much more information was extracted from the text, as the length of the report (the file was 13 Mb) was much longer than the short archive introduction text used in the exemplar above.

Grid Refs (6 entities):

TQ || 0825 || 7765

TQ || 0825 || 7765

TQ || 08300 || 77600

TQ || 07900 || 77800

TQ || 08550 || 77720

TQ || 08550 || 77550

As an additional feature of the web application six grid references were all correctly identified using pattern matching. Two of these entities are duplicates. In future, the application can be developed so that duplicates can be combined, and a quantity indicated next to the grid reference.

Author (1 entity):

Lorraine Mepham

Using the NER module the single author of the text was successfully identified by the web application.

Title (7 unique entities):

- Imperial College Sports Ground, Sipson Lane, Harlington, London Borough of Hillingdon - Archaeological Excavation (1)
- GLSMRlRCHME NMR ARCHAEOLOGICAL REPORT FORM (1)
- Imperial College Sports Ground, Sipson Lane; Harlington, London Borough of Hillingdon Archaeological Excavation (1)
- Imperial College Sports Ground Sipson Lane, Harlington Archaeological Excavation (1)
- Late Late Early Iron Late Iron LIA/ER-B Early R-B Late R-B ?(1)
- Imperial College Sports Ground, Sipson Lane, Harlington, London Borough of Hillingdon - An Archaeological Evaluation (1)
- Imperial College Sports Ground, Sipson Lane, Harlington, London Borough of Hillingdon Archaeological Excavation (3)

Seven unique titles were extracted from the PDF text by the NER module of the web application. The most identified "Title" was the "Imperial College Sports Ground, Sipson Lane, Harlington, London Borough of Hillingdon Archaeological Excavation", which was found three times. This was the correct title of the report. Seven suggested titles is a manageable number for the web application to effectively display to a user, but less would be preferable. Within these results there were several PDF conversion artefacts that future development of the web application can filter out by improving the web applications text extraction tool. This would reduce the number of suggested titles.

Subject (288 unique entities):

stone(1)

slag.(1)

armlet(1)

cremation cemetery(10)

arrowhead(1)

settlement enclosure(1)

rolled waste flint flake(1)

linear cemetery(1)

rim s.herds(1)

square enclosure(2)

ditched cursus(1)

spur stamp(1)

walled vessels(1)

settlement centre(3)

stone axes(1)

ritual' boundaries(1)

polished stone axe(1)

human occupation(1)

greywares(2)

Burnt Flint(2)

cremation vessels(1)

timber- lined well(1)

Brickearth(1)

shells(3)

barrow(1)

gravel pits(5)

style roundhouses(1)

gravel quarrying(1)

ditched enclosures(1)

worked flint(3)

Brickearth.(1)

planks(1)

burials(5)

strip fragments, one Romano-British coin(1)

slag(1)

Stone(2)

ditches(14)

sheep(3)

rectilinear enclosures(1)

flint(6)

animal bones(3)

Pottery(2)

field systems.(1)

bead rim jars(1)

seeds(2)

ironwork(1)

brick(1)

clay pipes(1)

worked quartz sandstone(1)

charred remains(3)

trackway ditches(1)

mollusc columns(1)

gravels(1)

aerial photographs(1)

field systems(6)

trackway(19)

causewayed enclosure(1)

ditch(16)

pits.(1)

wells(9)

Lynch Hill Gravels(1)

cattle(5)

field boundary(1)

vessels(5)

gully(1)

evaluation(1)

roundwood pegs.(1)

artefacts(4)

weed seeds(4)

worked timbers(1)

pits(39)

furrow(3)

snail shells(1)

iron-working slag(1)

cattle.(1)

field boundary ditches(2)

enclosure(37)

enclosure structure(1)

enclosure ditches(2)

roundhouse(1)

worked timber(1)

copper working slag.(1)

samian(1)

cropmarks(1)

ceramic(2)

knife blade(1)

boundary(2)

sheep.(1)

charred cereal grains(1)

Pits -.(1)

Peterborough Ware pottery(1)

wares(1)

rural settlement(1)

horse(4)

Ceramic(1)

pig bones(1)

Waterlogged plant remains(2)

domestic cattle(1)

daub(1)

blades(1)

field systems(1)

Charred weed seeds(1)

copper alloy(3)

gravel pit(1)

post-holes(1)

port(2)

archaeological evaluation(3)

charcoal(5)

long barrows(1)

settlements(1)

animal bone(2)

whiteware mortaria(1)

brickearth(1)

Long Mortuary Enclosure(1)

rectangular enclosure(3)

defended enclosure(1)

sports pitches(2)

imbrices(1)

roof tile(2)

Enclosure(1)

fired clay(1)

backed knife(1)

fineware' vessel(1)

saucepan pots(1)

wood.(1)

postholes(2)

rectangular enclosures(1)

pitcher handle(1)

settlement enclosures(1)

Fired Clay(3)

field boundaries(1)

axe(1)

flint tools(3)

settlement centres(1)

vessel(1)

long barrow(1)

field boundaries(3)

grog-tempered wares(1)

grain stores(1)

jar rims(1)

gravel flint(1)

hearth(1)

minimum bone(1)

metalwork(1)

pit(12)

Glass(3)

cremation pits(2)

bridge(1)

pollen(1)

ceremonial(1)

charcoals(3)

rectangular ditched 'ritual' enclosure(1)

Alice Holt industry(1)

tools (scrapers(1)

grain(1)

nails(1)

pig burials(1)

gullies(4)

inhumation(3)

tree line(1)

calcareous (shelly) wares(1)

field boundary(1)

inhumation grave(1)

Human Bone(3)

Ceramic Building Material(3)

burnt flint(1)

funerary pyres(1)

midden(8)

charred plant remains(2)

colour coated wares(1)

crosses(1)

cemetery(2)

enclosure ditch(1)

flakes(6)

cremation(3)

loom weights(1)

flint arrowheads(1)

market gardening(1)

enclosures(16)

Pit(2)

linear features(1)

pottery(23)

serrated blade(1)

fence lines(1)

point(2)

bowl(1)

rectilinear enclosures(1)

ritual(18)

ceramic building material(1)

animal pavvprint(1)

buff wares(1)

rural settlement(3)

ritual' enclosure(1)

burning(1)

water shells(1)

mortuary enclosure(2)

features(1)

cremation burial(1)

straw(1)

walls(1)

leaf arrowhead(1)

rim sherd(1)

track(1)

sherds(16)

horseshoe(1)

cremation pits(1)

enclosure ditch(3)

gravel terrace(1)

faunal remains(1)

enclosed settlement(2)

coarse oxidised wares(1)

post-excavation analysis(1)

Hmnan bone(1)

findspots(1)

floor(1)

gravel extraction(1)

jars(1)

Shelly wares(1)

waterlogged timbers(2)

subrectangular enclosures(1)

charred grain(4)

rolled waste flake(1)

horse metacarpal(1)

rectangular ditched 'ritual' enclosure(1)

Metalwork(2)

burial(5)

trial trenches(1)

cremations(9)

hearths(1)

bank(1)

Mortlake style(1)

worked flints(1)

small mammal bones(1)

timber-(1)

Peterborough Ware (1)

excavation(3)

greensand(1)

faunal evidence(1)

field system(8)

teeth(3)

stock enclosures(2)

tegulae(1)

dog(1)

Posthole(1)

cereals(1)

well(1)

flue tile(1)

rural settlement straddles(1)

ridge(3)

ring ditches(2)

bowls(1)

droveway(1)

worked stone(2)

Animal Bones(1)

metal(2)

wooden stakes(1)

timber(1)

globular urn(1)

road(2)

pottery sherd(1)

cores(2)

chaff(4)

funerary(1)

Charred plant remains(3)

iron(2)

mound(1)

Taplow Gravel(1)

gravel quarries(1)

bone(12)

ecofact(1)

Mortlake Ware sherds(1)

ditched divisions(1)

agricultural settlement(1)

Road(1)

farming economies(1)

plainware(1)

environmental evidence(2)

cremated human bone(1)

gravel terraces(1)

Oxfordshire fine wares(1)

settlements?(1)

Flint(1)

barbed-(1)

whitewares(1)

cemeteries(1)

flint flakes(1)

brickslfloor tiles(1)

settlement(57)

288 subject entities were identified by the NER module of the web application for this single PDF. The above table lists all 288 entities extracted with the number of times extracted in brackets next to each entity. All terms when evaluated by a domain expert in combination with the original text were identified as correct entities for the subject class. However, there were a quite a few erroneous entities which could be greatly reduced by improving the text extraction tool.

Despite the success of the web application in using NLP techniques to extract suggested metadata, from a data management perspective and from a web application user perspective, 288 entities, even when the erroneous entities are removed, is too large a number of metadata terms to manage. To be successful, other techniques will have to be applied to improve the effectiveness of the web application for its intended purpose.

The number of entities presented to a user by the web application could be reduced, if the outputs from the NER module are then compared to an archaeological gazetteer using the Levenshtein algorithm (a metric for measuring the differences between words). This will help reduce the number of outputs displayed in the web application by using “fuzzy string matching” to combine singular and pluralised terms, slight spelling mistakes, and differences in punctuation, into a single output with multiple values.

The next step in the development for the web application will be to determine how to weight and rank the results. This may be done by removing common entities that appear only once or identify them as less important in a ranking system. This will be a complicated task, and will require careful deliberation with domain experts as to what is considered a “common” term. For example, “Morelake Ware sherd” only occurs once in the list above, as does “mound” but it is likely that the more specific nature of the entity “Morelake Ware sherd” is a more informative term than the more general term “mound”. Displaying large quantities of extracted entities in a user friendly and intuitive manner that will allow a user to select entities they want to use as final metadata, will be essential to the success of the final web application.

Temporal (114 unique entities):

Romano British(1)

late Roman(1)

Late Upper Palaeolithic / Mesolithic; 12,000 - 8,500 BC).(1)

early Upper Palaeolithic(2)

Prehistoric(1)

Romano-British(10)

Late Neolithic(9)

Four Bronze Age(1)

Palaeolithic Roman I Meselitaie Saxon (pre-AD 1066(1)

Neolithic(40)

17th/early 18th century(1)

Medieval(7)

Iron Age (700 BC - AD(1)

Saxon(13)

Late Neolithic to post-medieval.(1)

Early Iron Age(7)

4th centuries AD.(1)

3rd- century(1)

1st(1)

post-Deverel-Rimbury(1)

Palaeolithic(6)

Middle(2)

Late Pleistocene(1)

early 1 st millennium BC(1)

12thl13th- century(1)

Early Romano-British(1)

MiddlelLate Iron Age(1)

Middle Bronze Age(7)

Early Iron(1)

Age - Late Neolithlc(1)

early Verulamium(1)

post- medieval(1)

late Romano-British(1)

Bronze Age/ Early Iron Age(1)

Later Saxon(2)

earlier Neolithic(2)

Late Pleistocene (Lower to Upper Palaeolithic; 500,000 -10,000 BC).(1)

Late Iron Age(12)

Early Romano-British(2)

Middle Palaeolithic(3)

Late Romano-Brltlsh Earfy Romano-Brltlsh Late Iron Age Early Iron Age(1)

post-medieval(5)

Medieval (AD 1066 - 1500(1)

post- Roman(1)

Early Bronze Age (2,400-i,500BC(1)

11 Post-medleYal/Nodern M.dleval Saxon(1)

Late Iron Age/early Romano-British(3)

Late Bronze Age/Early Iron Age(1)

Early Romano-Brltlsh _ Late Iron Age _ Early Iron Age(1)

Romano-British (AD 43 - 410(2)

Victorian(1)

Early Neolithic(1)

Romano- British(2)

Medieval (AD 1066(1)

medieval(19)

later Bronze Age(1)

Mesolithic(5)

Roman (43-410 AD)(1)

modern(2)

Late Bronze Age(1)

Iron Age (700 BC - AD 43)(1)

Neolithic Medieval (AD 1066-1485) Bronze Age Post-Medieval I Iron Age(1)

Early Bronze Age(8)

late 12th to early 13th century(1)

500,000 - 4,000 BC(2)

Middle Neolithic(1)

Late Bronze Age(14)

Bronze Age(21)

later Neolithic(1)

AD 1500(1)

Later Romano-British(2)

Middle to late Bronze Age(1)

Early Iron Age(1)

Iron Age-(1)

Post-medieval(7)

later Neolithic(2)

Roman(17)

Holocene(1)

AD) - Post l(1)

Middle Bronze AgelIron Age(1)

early 2nd century AD.(1)

5th century AD(1)

Neolithic (c. 4,000 - 2,400 BC(1)

Late Romano-British(3)

Mesolithic communities (8,500-4,00 BC(1)

Neolithic to Bronze Age(1)

Early Romano-Brltlsh Late Iron Age Early Iron(1)

Late Roman(1)

1 st century BC(1)

Neolithic/Bronze Age(1)

early 5th century(1)

Early Post-Glacial(1)

Late NeolithicEarly Bronze 132 - 561531481721- Age 1 - 1515 Middle/Late Bronze(1)

early Romano-British(1)

Late Iron Age/ Romano-British(1)

Later Bronze Age(3)

Early Rom Late Iron Age - Early Iron Age - Late Bronze .(1)

prehistoric(10)

Late Romano-British 4.7.9.(1)

Bronze Age (c. 2,400 - 700 BC(1)

Late Neolithic(1)

1st century AD.(1)

Middle Palaeolithic (500,000- 30,000 BC(1)

Late Iron Age(1)

early 2nd century(1)

late Romano-Brltlsh Early Romano-Brltfsh late Iron Age Early Iron Age(1)

Saxon Medieval Neolithic Bronze Age Age Age(1)

Iron Age(19)

Late Iwn Age(1)

mid 3rd century AD(1)

Bronze Age (c. 2,400 -700 BC(1)

Pleistocene(2)

Modern (AD 1500(1)

Saxon (AD 410 - 1066(2)

114 unique “temporal” entities were generated from the uploaded PDF document. As in the case of the “subject” entities there are spelling mistakes and erroneous entities returned, but all are correct “temporal” terms. The number of entities displayed in the web application can be reduced in the same manner described above.

Ranking “temporal” entities simply based on the number of times they are referred to in the text, could be more effective than in the case of “subject” entities, as the multiple use of a temporal term such as “Neolithic” would suggest the report is largely about the neolithic, but the single instance of the term “modern” would be deemed less important. Many of the entities extracted relate to the same temporal period, but have been written differently. During the next stage of the web application development, the possibility of grouping the extracted entities into more general time periods will be investigated.

Placenames (47 unique entities):

River Crane(4)

Brentford(2)

London Borough of Hillingdon(2)

Mayfield Farm(1)

River Colne(5)

Wall Garden Farm(1)

Oxfordshire(1)

Long Enclosure(1)

Peterborough Ware(1)

Lynch Hill Gravels(2)

Surrey(2)

Great South West Road(1)

London(5)

Harlington(5)

London Clay(1)

Cranford Lane(4)

Lower Thames(1)

East Anglian(1)

Mill Road(1)

Middle Thames Valley(1)

Bath Road(2)

Lynch Hill(1)

River Thames(2)

Sipson Lane(1)

Staines(1)

Rivenhall(2)

Hillingdon(1)

Horton(1)

Middle Thames(1)

Colne Valley(1)

Imperial College(1)

West Dray(1)

Normanton Down(2)

Perry(1)

Runnymede Bridge(1)

Walled Garden Farm(1)

Burrows Hill(1)

Purley Way(2)

Prospect Park(1)

Greater London(2)

Cranford(1)

Sipson Lane(1)

Heathrow(2)

West London(1)

Taplow Gravels(1)

Sipson's Lane(1)

Colne(1)

Forty-seven unique “placenames” were extracted from the PDF text by the NER module of the web application. All 47 entities were correct placenames. Future development of the web application may allow locational terms to be ranked by “placename” type, with countries and counties at the top of list, followed by possible cities, towns and villages, then by more local terms such as roads, rivers and farms.

Contact (1 unique entity):

Portway House, Old Sarum Park, Salisbury, Wiltshire

These were the correct contact details for the company responsible for writing the report. These details are important to extract as they are often required in the ADS Collection Management System.

4.4 Evaluation of Work to Date

The NER module works successfully and produces correct entities for the classes it has been trained to identify. The NLP tools currently under development will be very useful for extracting resource discovery metadata from unstructured archaeological data, particularly grey literature reports, for resource discovery indexing, where little or no metadata currently exists. From a data management perspective however, the large quantities of entities extracted by the NER module can be too large to effectively manage. The annotation tool built into the web application will allow users to produce more training data to better train the module.

The inclusion of an annotation tool within the web application will also be particularly beneficial to other ARIADNE partners who may wish to use it. By including an annotation tool in the web application we have removed the need for the annotation process described above to be conducted by each ARIADNE partner. Instead the web application could be used to create rules for each language, which could then theoretically be used as training material to train a machine-based classifier, and therefore further enhance the output.

4.5 Future Development

Future development of the web application will include:

- The refinement of the interface style to make it easier and more intuitive to use. This is very important, as crowdsourcing may be explored to process large quantities of unstructured data.
- Improvement of the text extraction module as discussed above.
- The development of a module to export the selected metadata in a variety of formats.
- As a test of the module in a working system, the integration of the web application in the redevelopment of the OASIS system (the online system for indexing archaeological grey literature in the UK). The aim of this will be to allow an archaeologist to upload a report to OASIS, and by choosing to use the web application, they will be able to automatically extract suggested metadata for the report. The metadata will be accepted or rejected by the user and then automatically populated into the correct fields within OASIS.
- Techniques will be explored to “tidy”, group and rank the entities outputted from the NER module using text clustering, and generating cluster labels based on the content in respective clusters.

Regarding future developments, the following issues will be explored:

Issue 1

Negation in unstructured archaeological text content was observed during the evaluation. It is crucial for any practical implementation this is recognized. For example, if a named entity occurs inside the scope of a negation then that named entity should not be included in the output. Negation detection will be explored (which can be as simple as matching negative words).

Issue 2

Feedback from domain experts suggests that although the entities detected by the system are valid terms, some are not considered important. Although, we do not think this is a problem from an NLP perspective, nevertheless, it is desirable for the system to be more selective. So far, work has been focused on NER, but it may be possible to solve this issue using techniques from Entity Linking (EL). The problem is distinct from the NER module, as it does not identify the occurrence of the “names”, but their reference. In order to build such a system, a knowledgebase is needed. It may be possible to develop this knowledgebase from available archaeological Linked Open Data. We can define entity linking as matching a textual entity detected by the NER module, to a knowledgebase entry, such as a Linked Data node that is a canonical term for that entity. However, entities are often detected by the NER module which have different surface forms, including abbreviations, shortened forms, or aliases. Therefore, EL must find an entry despite changes in the detected string by the NER module. Entity Ambiguity (EA) resolution is another problem that will need to be resolved when using this technique. For instance, “Roman”, can match multiple Linked Data entries as either “subject” or “temporal”. The last difficulty is the absence of the entity in the knowledgebase. Processing large text collections guarantees that many entities will not appear in the Linked Data, so the system may not be able to cope with this situation. Addressing a “negative sample” could be used when creating the training data, as opposed to the positive samples taken during the original training of the data used for this tool.

5 Conclusion

The ARIADNE partners involved in this deliverable will continue to explore NLP with the aim of making text-based resources more discoverable and useful. The partners have specifically focused on one of the most important, but traditionally difficult to access resources in archaeology; the largely unpublished reports generated by commercial or "rescue" archaeology, commonly known as "grey literature".

The partners have explored aspects of rule-based and machine learning approaches, the use of archaeological thesauri in NLP, and various Information Extraction (IE) methods. This includes work by the University of South Wales, in partnership with Leiden University, on archaeology thesauri for NLP, which applied Named Entity Recognition (NER) to the Dutch Rijksdienst Cultureel Erfgoed (RCE) Thesauri. The process of importing a subset of RCE thesauri resources into a specific framework (GATE), and the suitability and performance of the selected resources when used for the purposes of Named Entity Recognition (NER) were discussed.

This revealed issues relating to the role of the RCE thesauri in NER and further development of techniques for the annotation of Dutch compound noun forms. Several actions will be taken for the next iteration of this work and for adapting the ontology resource to the requirements of the NLP task. Some actions that can significantly improve the matching result are easy to implement. Other actions will require a possible structural modification of the resource for NLP purposes, relating to inclusion of new concepts (not just synonyms or spelling variations). These include easy and straightforward actions which can be carried out quickly, and medium to complex actions that will require input from Dutch archaeology domain experts.

South Wales also undertook a study for a Dutch NER pipeline, which included the results of the early pilot evaluation based on the input of a single, manually annotated document. The report also presented the results of the vocabulary transformation task from spreadsheets to RDF/XML hierarchical structures, expressed as an OWL-Lite (ontology). Observations related to the vocabulary transformation process and pipeline results, and revealed initial issues that affect vocabulary usage and focus of the NER exercise. The document was annotated with respect to the following entities; Actor, Place, Monument, Archaeological Context, Artefact, Material, Period. The NER pipeline is configured to identify the following entities: Place, Physical Thing (i.e. Monument), Physical Object (i.e. Artefact), Time Appellation (i.e. Period), Material, Context. Each entities produced differing levels of results, which in some cases were good, but others need to be explored further for improvement.

Work was carried out by the Archaeology Data Service (ADS) at the University of York, to develop and evaluate machine learning-based NLP techniques and integrate them into a new metadata extraction web application, which takes previously unseen English language text as input, and identifies and classifies named entities within the text. The outputs will be used to enrich the resource discovery metadata for existing and future resources. The final application will include a web-based, user friendly interface that can be used by archaeological practitioners to automatically generate metadata related to uploaded text-based content on a per-file basis or using batch creation of metadata for multiple files.

Early work has revealed the NER module works successfully and produces correct entities for the classes it has been trained to identify. The NLP tools currently under development will be very useful for extracting resource discovery metadata from unstructured archaeological data, particularly grey literature reports, for resource discovery indexing, where little or no metadata currently exists. From a data management perspective however, the large quantities of entities extracted by the NER module can be too large to effectively manage. The annotation tool built into the web application will allow users to produce more training data to better train the module. ADS will continue to work to refine the tool, especially with regard to the interface to make it easier and more intuitive to use, exploring crowdsourcing for processing large quantities of unstructured data, improvement of the text extraction module, development of a module to export the selected metadata in a variety of formats, integration of the web application in the redevelopment of the OASIS system (the online system for indexing archaeological grey literature in the UK), and techniques will be explored to “tidy”, group and rank the entities output from the NER module using text clustering, and generating cluster labels based on the content in respective clusters.

To date, the partners have successfully explored a variety of NLP techniques to make text-based archaeological resources more discoverable and useful. However, there are still several key issues which need to be addressed to fully achieve the potential of the techniques explored. The partners will continue to pursue and refine these techniques over the rest of the project, which will be reported in D16.4, *Final report on natural language processing*.