Knowledge-based Named Entity Recognition of Archaeological Concepts in Dutch

Andreas Vlachidis[1][0000-0003-0986-4430], Douglas Tudhope[2][0000-0002-5222-0430] and Milco Wansleeben[3]

[1] Department of Information Studies, UCL, Gower Street, WC1E 6BT, London, UK
[2] School of Computing and Mathematics, University of South Wales, CF37 1DL, UK
[3] Faculty of Archaeology, Leiden University, Einsteinweg 2, 2333 CC, Leiden, The Netherlands
a.vlachidis@ucl.ac.uk
douglas.tudhope@southwales.ac.uk
m.wansleeben@arch.leidenuniv.nl

Abstract. The advancement of Natural Language Processing (NLP) allows the process of deriving information from large volumes of text to be automated, making text-based resources more discoverable and useful. The attention is turned to one of the most important, but traditionally difficult to access resources in archaeology; the largely unpublished reports generated by commercial or "rescue" archaeology, commonly known as "grey literature".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 paper presents the development and evaluation of a Named Entity Recognition system of Dutch archaeological grey literature targeted at extracting mentions of artefacts, archaeological features, materials, places and time entities. The role of domain vocabulary is discussed for the development of a KOS-driven NLP pipeline which is evaluated against a Gold Standard, human-annotated corpus.

Keywords: Named Entity Recognition, Archaeology, Grey Literature, CIDOC-CRM, Knowledge Organization Systems.

1 Introduction

Across Europe, the archaeological domain generates vast quantities of text in form of unpublished fieldwork and specialist reports often referred to as "grey literature" [1]. In the Netherlands, it is estimated that just under 60,000 of such reports have been produced over the last 20 years with a current estimated growth rate of 4,000 reports per year [2]. Access to the valuable information contained in such reports is a known problem. 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. The role of the Natural Language Processing (NLP) has been recognised as vital for the automatic indexing, metadata generation, retrieval and dissemination from integrated online catalogues [3]. The results of Brandsen's [4] study on the uses of Named Entity Recognition (NER) for the development of effective search experiences, agree with earlier findings that

NER applications can enable semantic indexing of archaeological grey literature for the purposes of retrieval and cross searching [5].

The paper discusses the development and evaluation of a NER system of Dutch archaeological grey literature targeted at extracting mentions of artefacts, archaeological features, materials, places and time entities. The system uses a rule-based information extraction technique supported by domain vocabulary, utilising the GATE (General Architecture for Text Engineering) framework [6] and contributing to NLP aims of the European FP7 project ARIADNE [7]. The main motivation of the work is to enable the semantic annotation of archaeological reports with a core set of entities of interest for automating metadata generation. The extracted output of entities is delivered in a structured and interoperable XML format, constituting a document index which can be further analysed and used to identify patterns, trends, and "important" words or terms within text by subsequent applications. Such semantic annotation interoperable outputs have been delivered by previous studies in English to facilitate archaeological information discovery, retrieval, comparison, analysis, and link texts to other types of data [8]. The current study expands the method of automatic metadata generation using NER, previously available in English, in the context of Dutch archaeological reports.

2   Background

The field of Named Entity Recognition has been consistently growing over the past two decades. The early rule-based (handcrafted) systems which provided good performance at a relatively high system engineering cost were succeeded by the machine learning (supervised) systems which allowed for greater scalability and domain adaptation but required human-annotated data for system training [9]. The latest developments in NER explore semi-supervised and unsupervised learning techniques which promise to overcome some of the limitations of supervised ML methods and to provide information extraction results without the prerequisite of an annotated corpus [10]. Our decision to employ a rule-based approach for the development of the Dutch system was based on the absence of an annotated corpus and on the availability of the Rijksdienst vor het Cultureel Erfgoed (RCE) Thesauri which supported the rule matching approach with a breadth of domain vocabulary.

Information Extraction (IE) aims to identify instances of a particular prespecified class of entities, relationships and events in natural language texts, and the extraction of the relevant properties (arguments) of the identified entities, relationships or events. NER is specified as a subtask of IE and the term was first used during the Sixth Message Understanding Conference (MUC-6) [11] to describe the task of extracting instances of people, organizations, geographic location, currency, and percentage expressions from text. Similarly, the task of NER was defined by the 2002 Conference on Computational Natural Language Learning as the extraction of 'phrases that contain the names of persons, organizations, locations, times, and quantities' [12]. However, there is no single definition of NER as the task has kept on expanding and diversifying through the years to include additional entities of interest such as products, events, diseases, to name but a few [13]. In the context of archaeological fieldwork reports the entities that have

extracted the most interest relate to physical object, material, spatial and temporal information [5].

A number of projects have employed IE and NER techniques on archaeological literature. An early pilot application was carried out by Amrani, Abajian and Kodratoff [14] which used string matching to extract information from archaeological literature. The OpenBoek project experimented with memory based learning to extract chronological and geographical terms from Dutch archaeological texts [15]. Byrne and Klein [16] also investigated the extraction of information from archaeological literature primarily focusing on extraction of events from unstructured text. The Archaeotools project adopted a machine learning approach to enable access to archaeological grey literature via a faceted classification scheme of What (what subject does the record refer to), Where (where, location, region of interest), When (archaeological date of interest) and Media (form of the record) which combined databases with information extracted from reports in an interesting faceted browser interface [17]. The OPTIMA system applied a rule-based, Knowledge Organization System (KOS) driven approach to semantic indexing of archaeological grey literature [18]. It used named entity recognition, relation extraction, negation detection and word sense disambiguation for associating contextual abstractions with classes of the standard ontology (ISO 21127:2006) CIDOC Conceptual Reference Model (CRM) for cultural heritage together with concepts from English Heritage thesauri and glossaries.

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. Moreover, the traditional model of manual cataloguing and indexing practices has been receiving less attention and priority. For example, prominent European research projects such as the eContentplus explicitly did not fund the development of metadata schemas and the creation of metadata itself [13]. Natural language processing techniques can support automatic generation of rich metadata, providing methods for disclosing information in large text collections whilst enabling semantic search of grey literature across disparate collections and datasets [3][8]. Such approaches compensate full text indexing techniques, enabling retrieval on multiple meanings and allowing researchers to search on concepts taking account for synonymy and polysemy [4].

A significant amount of research effort has been spent on information extraction in English, covering NER as well as higher-level IE tasks such as relation and event extraction. Comparatively less attention has been spent on non-English languages. The performance of non-English IE systems is usually lower and linguistic phenomena impose challenges [19]. Such challenges include lack of whitespace, which complicates word boundary disambiguation; productive compounding, which complicates morphological analysis in German and Dutch; and proper name declension forms in Greek and Slavic languages which complicate named entity recognition [20].

The Dutch NER pipeline discussed in the paper, is challenged by language related issues that directly affected recognition of compound noun forms, place names and time entities. The following sections discuss the methods and techniques used in an attempt to address some of these challenges and to deliver a customised application for the extraction of entities of interest from Dutch archaeological grey literature.

3   Method

This section discusses the stages of developing an NLP pipeline which employed rule-based information extraction techniques and integrated a range of domain vocabulary resources for supporting the task of entity recognition. The final pipeline is the result of an iterative process which involved the definition and evaluation of an earlier pipeline version. The earlier version adapted the domain vocabulary to the NLP task by utilising the SKOSified[1] version of the RCE thesauri[2], designed and developed the IE rules and evaluated the performance of the NER pipeline. The updated version of the Dutch NER pipeline, improved a range of vocabulary issues in connection to coverage, spelling variations and synonyms, refined the gold standard and modified the entity matching rules for better performance.

The employment of rule-based IE and domain vocabulary resources distinguishes our approach from supervised machine learning methods, which heavily 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 the rule-based techniques. Hand-crafted rules invoke input from ontologies and thesauri that provide to the entity recognition rules specific terms of predefined groups, such as person names, organisation names, week days, months etc. In addition, the rules 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.

3.1   Vocabulary GATEfication

The NER pipeline is designed to extract core concepts (entities) of research interest in the context of Dutch archaeological grey literature, such as artefacts (finds or physical objects), features (archaeological context e.g. posthole), materials, monuments types, places (focus on place names such as districts) and time entities (periods and time appellations including numerical appellations e.g. 480 BC). The following RCE thesauri have been selected to support extraction of the above entities; archaeological artifact types, materials, archaeological complex (features) types, locations, archaeological periods and landscape elements of object types thesaurus.

---

[1] Simple Knowledge Organization System (SKOS) is a Semantic Web format and aW3C recommendation designed for representation of thesauri, classification schemes, taxonomies, subject heading systems, or any other type of structured controlled vocabulary https://www.w3.org/2004/02/skos/

[2] http://openskos.org/api/collections/rce:EGT.html

The process of importing the RCE thesauri resources into the GATE framework involved retrieval of the thesauri resources and their serialisation (transformation) to the Ontology Web Language (OWL-Lite) format using automated methods (i.e. XSL templates). The original serialisation of the thesauri can be only partially parsed from the framework, causing the rich thesaurus structure to flatten and in turn restricting the definition of rules that can exploit the broader/narrower semantic relationships. Hence, the transformation of the original resources was necessary for exploiting the hierarchical relationships of the resources, enabling matching on alternative labels and synonyms, and 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 uniform resource identifiers (URIs) while maintaining the original references for individual entries (i.e. *rna:contentItem* and *skos:Concept* and *rdf:about*). The necessity to provide new human-readable URIs for classes is dictated by GATE's behaviour towards exposing class URI to JAPE rules.

GATE enables Ontology Based Information Extraction (OBIE) techniques using OWL-Lite that purely support the aims of information extraction and are not stand alone formal ontologies for logic based purposes. Such ontological structures 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" and their 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 the NER task. Thus, matches derived from an ontological resource enjoy dimensions that could be useful for further information retrieval and interoperability purposes.

Transformation of the RCE thesauri to OWL-Lite was performed with XSL templates which produced 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 instead of the original *skos:Concept* property, 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 *rdfs:subClassOf* structure was used to implement the broader-narrower terms as Parent/Child relationships.

The NER pipeline is composed by several general purpose, domain independent NLP modules and a series of bespoke JAPE transducers which contain the hand-crafted rules that exploit contextual evidence and domain vocabulary. In detail, the pipeline

integrates the modules in the following order; Apache Open NLP[3] (Tokenizer, Sentence Splitter, Dutch part-of-speech tagger), the Snowball stemmer[4] , the GATE Onto-Root Gazetteer[5] and finally a range of JAPE transducers responsible for extracting entities of interest from text. The pipeline runs in a cascading order where each module adds a layer of semantics to the output, hence, the order of the modules is important. The stemmer output is critical for the operation of the OntoRoot gazetteer module which produces lookup annotations that link to the specific concepts or relations from the ontology. The output of the OntoRoot module is then exploited by the NER matching rules which combine lookup and token input, for example the following rules matches in text transitively all instances of the ontology class artefact which are tagged by the part-of-speech tagger as nouns.

```
{Token.category ==N, Lookup.class ==Artefact}
```

In addition, a flat gazetteer containing period related suffixes, such as A.D, B.C, voor Christus was built and used in the definition of JAPE rules targeted at matching numerical dates e.g. *1200 AD, 800 v.Chr*. Similarly, a set of JAPE rules was defined for matching grid references and geographic coordinates (numerical places), such as 216.518/568.889.

3.2   The NER Pipeline

The updated version of the Dutch NER pipeline addressed shortcomings following the review of the earlier NER pipeline. Such imperfections concerned the coverage and suitability of the RCE Thesauri to support the NER task and under performance of JAPE rules in connection to vocabulary and pattern matching. The updated version did not address issues of compound noun extraction and negation detection which were also revealed during the review.

The main effort of vocabulary improvement was focused on resolving overloaded vocabulary entries into individual term components. The RCE thesauri were not necessarily developed with Natural Language Processing in mind and as a result contain entries that are not suitable for automatic and algorithmic term matching due to their multiterm, sometimes descriptive and verbose punctuation structure. For example, the vocabulary entries amulet/talisman and its child entry *amulet/talisman – kruisvormig* (cruciform) do not correspond to the way in which such terms are used in natural language text.

Most likely either amulet or talisman will be found as individual entries and if an adjective is used, such as *kruisvormig* this will follow a grammatically correct syntax form (i.e. *kruisvormig amulet* instead of *amulet kruisvormig*). Vocabulary entries like the above should be enhanced with labels that are closer to what is likely to appear in text rather than containing descriptive and non-natural language descriptions.

---

[3] https://opennlp.apache.org/
[4] https://snowballstem.org/
[5] https://gate.ac.uk/sale/tao/splitch13.html#x18-33400013.8

A set of XSL templates was developed which addressed label patterns that joined synonyms and specialisations together under a single label. For example, the forward slash (/) character joins synonyms as in the case amulet/talisman, the hyphen (-) character adds specialisation as in the case *amulet/talisman – kruisvormig* and the comma (,) character adds a form of periphrastic description which can be treated as an alternative label. The XSL templates incorporated the above patterns to generate the new vocabulary labels where for example amulet/talisman delivers two separate labels (amulet, talisman) and *amulet/talisman – kruisvormig* delivers the labels *kruisvormig* amulet and *kruisvormig talisman*. In most cases, special characters for joining synonyms and expressing specialisations or generalisations are standard across the thesauri and the transformation delivered useful alternative labels. However, there are cases that do not follow the standard use of special characters or are very verbose (e.g. *huisplattegrond:4-schepig - type St.Oedenrode*). Such cases due to their complexity were not matched by the transformation templates and were ignored.

The updated NER pipeline incorporated improvements to the JAPE matching rules which addressed vocabulary use and matching coverage. A new set of rules was introduced for matching grid references of places and new rules were also included for exploiting input from the Landscape elements of the Objecttypen thesaurus. Rules were also improved for matching the gazetteer lists in connection to dates enabling matching of date range such as *tussen 1600 en 1900* (between 1600 and 1900). In addition, the restriction that any match of a Place entity must commence with an upper-case letter has been lifted, to include matching for place names commencing with s', such as *'s-Heerenberg, and 's-Graveland* which is quite common in Dutch.

Fig. 1. Example semantic annotations in the context of a Dutch archaeological report. Annotation types are highlighted and annotation attributes are shown in respective fields.

The Dutch NER pipeline has been deployed into the GATE Cloud[6] where it is freely available for accessing through a web interface and a dedicated API. Example semantic annotations of archaeological entities of interest include, Time Appellation, Physical Object and Place (see Fig.1) such as, *Swifterbant* which a town in province of Flevoland, the object *Trechterbeker* (Funnel beaker), the time appellation *Steentijd* (Stone Age) and the materials *Houtskool* (charcoal), *Vuursteen* (flint), *Zandsteen* (sandstone). In addition, a range of attributes are assigned on each individual annotation that carry pieces of information about the origin of a term (contributing thesaurus), unique reference (URI) and a corresponding terminological reference to the RCE thesaurus which is uniquely identified by URI.

4   Evaluation

The system performance was benchmarked via a Gold Standard (GS) set of manual annotations defined for the purposes of the ARIADNE project by a group of Dutch archaeologists (Leiden University). The Gold Standard refers to a set of human annotated documents which represents the desirable result and is used for comparison with system produced automatic annotations. Results are reported on the measurement of Precision and Recall and their weighted average F-measure, established as standard measurement units for measuring the performance of IE by the second Machine Understanding Conference, MUC 2. The Gold Standard consisted of 7 long (some are up to 300 pages) grey literature reports containing approximately 10,000 annotated instances of several entity types, including Archaeological Context (Feature), Artefact, Event, Material, Method, Monument, Place, Period and Person. The entities Event, Method, and Person were not in the scope of the NER pipeline and are not included in the evaluation.

The manually annotated GS was very helpful for the purposes of an early evaluation task, revealing several issues with regards to vocabulary coverage and suggesting potential rule matching strategies for a range of different entities. However, it was regarded to be a bit long and quite repetitive for the evaluation of a rule-based system and more appropriate for training a Machine Learning system. In the case of rule-based evaluation, since training is not required, we need a representative manually annotated corpus that covers as many as possible different cases of annotation without repeating the same annotation cases again and again. For example, a place name ("Veemarktterrein") which was not initially included in the rule-based system vocabular is frequently found in text, affecting recall over long documents. This is rather straight forward to rectify the system by including the missing term in the vocabulary.

The early NER pipeline was evaluated against the GS delivering the performance figures as seen on Table 1. The overall score of Precision and Recall were encouraging, reaching 57% and 61% respectively and delivering an F-Measure score of 59%.

Table 1. System Performance of the Early NER system for a range of entities

Entity	Recall	Precision	F-Measure
Arch.Context	0.87	0.63	0.73
Artefact	0.36	0.50	0.42
Material	0.50	0.62	0.55
Monument	0.36	0.45	0.40
Place	0.60	0.65	0.63
Time Period	0.72	0.70	0.71
All(Total)	**0.57**	**0.61**	**0.59**

The least performing entity is the Monument both in terms of Precision (36%) and Recall (45%), followed closely by the Artefact entity which shares the same Precision score and slightly better (50%) Recall. The pipeline delivers slightly better scores for the Material and Place Entities with Precision scores between 50%-60% with Recall scores between 62%-65% respectively. The best performing entity is Archaeological Context which enjoys a Precision score of 87% and Recall 63%, followed by the Period entity which scores 72% Precision and 71% Recall. The contribution of vocabularies is critical to the performance of the pipeline with respect to the discussed entities. Clearly, Precision can be harmed by using too many terms from the available vocabulary which do not fail within the scope of the targeted entity. At the same time, Recall is affected by using too few terms from the vocabulary, hence, matching rules should be also improved for allowing an optimum use of the available vocabulary.

The updated version of the NER pipeline utilised an extended and NLP friendlier vocabulary resource, which has addressed various labelling pattern issues as discussed in section 3. In addition, the pipeline incorporated new improved hand-crafted rules aimed at improving performance and strengthening the matching accuracy. For example, a rule aimed at matching instances of the artefact class, which previously included two conditions, it was strengthen to include five separate conditions as seen below.

```
{!NotLookup, !Context, !Physical_Thing, Token.category ==
N, Lookup.class == [rceArtefact n=ArcheoArtefactTypes]}
```

The rule matches all instances of the RCE Artefact Types class, excluding from matching; certain areas of the resource previously annotated as NotLookup, Archeaological Context, and Monuments (Physical Thing) whilst requiring each match to conform to Noun token-category. The updated version delivers improved results (table 2).

Table 2. System Performance of the Early NER system for a range of entities

Entity	Recall	Precision	F-Measure
Arch.Context	0.90	0.63	0.74
Artefact	0.53	0.60	0.56
Material	0.55	0.62	0.58
Monument	0.64	0.65	0.65
Place	0.61	0.72	0.66
Time Period	0.80	0.76	0.77
All(Total)	**0.67**	**0.68**	**0.67**

The Precision of the NER pipeline is improved by 10% whereas Recall also is also improved by 7%, reaching overall 67% Precision and 68% Recall. Most significantly, the performance of the pipeline has been considerably improved for the Artefact and Monument entities types, with Precision score nearly doubling and increasing from 36% to 54% and 64% respectively. The performance of the pipeline is comparable with the results of the AGNES system [4] which employed a supervised ML approach for the recognition of Artefact, Time Period and Material entities in Dutch archaeological grey literature. In particular, both systems have reported similar scores in respect to recognition of Artefact entity types which presents certain domain challenges. Overall, the ML system delivered a higher Precision 71% score but a lower Recall 48% score, resulting to an F-Measure 56% compared to 66% delivered by the rule-based, KOS-driven system. The above comparison is only indicative and highlights the challenges imposed by the archaeological domain in NER. A full-scale comparative study across the rule-based, ML (supervised and unsupervised) methods of NER would be appropriate for delivering safer conclusions as each method has its own merits and limitations. It is evident that the RCE Thesauri proved a valuable resource to drive the NER effort, providing a significant vocabulary breadth which benefited the Recall rates of the system. At the same time, the hand-crafted rules as improved during the iterative process allowed for a maximum use of the available vocabulary whilst imposing conditions which protected the overall Precision rates of the system.

5 Conclusion and Future Work

A major development of the NER rule-based and KOS-driven approach has been the generalisation of the previous rule based techniques [18] to Dutch archaeological grey literature. The work faced challenges in connection to a different set of vocabularies available via the RCE Thesaurus and also in connection to differences in language characteristics. The NER techniques were focused on the general archaeological entities of Archaeological Context, Artefact, Material, Monument, Place, and Time Period and the method proved capable of extracting entities of interest with relative success. The RCE Thesauri proved to be a valuable resource in support of the NER task, 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 addition, such domain vocabulary resources in many cases have been defined with Information Science principles in mind, and not to support NLP operations. The Gold Standard (GS) evaluation revealed performance drawbacks influenced by structural, labelling and coverage issues of the vocabulary. The results of the evaluation phase lead to resolving the overloaded vocabulary entries into individual term components. Therefore, labelling adjustment and enrichment techniques are necessary for making the vocabulary resources "friendlier" to the NER as discussed in section 3.

The updated version of the NER pipeline improved many cases of underspecified rules that were identified during evaluation of the earlier and also identified some new cases of underperformance. Future improvements include extraction of compound noun forms which appear in Dutch regularly, joining period terms with objects, object terms with material, material terms with archaeological contexts etc. A way forward of tackling such cases could be is to employ partial matching over words instead of the whole word matching currently performed by the NER system. Partial matching is possible but should be planned and executed carefully due to the significant amount of noise that might be generated. A future system should also be able to address negated entities (e.g. *geen vondsten/finds*) which provide facts of no evidence, i.e. a comment in the report that no evidence has been found for a potential finding and thus should not be annotated. Last but not least, the current NER pipeline imposes a noun-validation restriction which excluded adjectives from matching. However, the Gold Standard revealed many material entities of adjectival form, such as *bronzen* (bronze), *stenen* (stone), etc. The restriction can be lifted but careful planning is required in order to conclude whether such cases are annotated as individual material entities or as moderators of object or monument entities.

Acknowledgments

References

1. TN. Evans. A Reassessment of Archaeological Grey Literature: Semantics and paradoxes. Internet Archaeology, 40, 2015.
2. Rijksdienst vvor het Cultureel Erfgoed. 2019. Archis Invoer. Available at: https://archis.cultureelerfgoed.nl (Last accessed 05 May (2019).
3. J. Richards, D. Tudhope, A. Vlachidis. Text Mining in Archaeology: Extracting Information from Archaeological Reports. In: Barcelo, JA and Bogdanovic, I (eds.), Mathematics and Archaeology: 240–254. Boca Raton: CRC Press. (2015).
4. A. Brandsen, K. Lambers, S. Verberne, M. Wansleeben. User Requirement Solicitation for an Information Retrieval System Applied to Dutch Grey Literature in the Archaeology Domain. Journal of Computer Applications in Archaeology, 2(1):21-30, 2019

5. A.Vlachidis, D. Tudhope. A knowledge- based approach to Information Extraction for semantic interoperability in the archaeology domain. Journal of the Association for Information Science and Technology, 67(5): 1138–1152, (2016)

6. H.Cunningham, V. Tablan, A. Roberts, K. Bontcheva. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLoS Comput Biol 9(2): e1002854. (2013)

7. C. Meghini, R. Scopigno, J. Richards, H.Wright, G. Geser, S. Cuy, J. Fihn, B. Fanini, H. Hollander, F. Niccolucci, A. Felicetti. ARIADNE: a research infrastructure for archaeology. Journal on Computing and Cultural Heritage (JOCCH), 10(3):18, (2017)

8. D.Tudhope, K. May, C.Binding, A. Vlachidis. Connecting archaeological data and grey literature via semantic cross search. Internet archaeology, 30, (2011)

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

10. J.I. Toledo, M. Carbonell, A. Fornés, A. J. Lladós. Information extraction from historical handwritten document images with a context-aware neural model. Pattern Recognition (86):27-36, 2019

11. R.Grishman, B. Sundheim. Message Understanding Conference-6: a brief history, 16th International Conference on Computational Lingusitics,:466–471, (1996)

12. E.F Tjong KimSang.Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition, in Proceedings of CoNLL-2002:155–158, (2002)

13. S.Van Hooland, M. De Wilde, R. Verborgh, T. Steiner, R. Van de Walle. Exploring entity recognition and disambiguation for cultural heritage collections. Digital Scholarship in the Humanities, 30(2): 262-279, (2013)

14. A.Amrani, V.Abajian, and Y.Kodratoff, Y. A chain of text-mining to extract information in archaeology. In: Annual IEEE Computer Conference, International Conference on Information and Communication Technologies: From Theory to Applications, and ICTTA (eds.), 3rd International Conference on Information and Communication Technologies: from Theory to Applications, 7–11 April (2008)

15. H.Paijmans, S. Wubben. 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)

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

17. S. Jeffrey, J. Richards, F. Ciravegna, S.Waller, S. Chapman, Z. Zhang. The Archaeotools project: Faceted classification and natural language processing in an archaeological context. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 367(1897): 2507–19. 2009.

18. A.Vlachidis. Semantic indexing via knowledge organization systems: Applying the CIDOC-CRM to archaeological grey literature (Doctoral dissertation, University of Glamorgan). 2012

19. J. Piskorski, J. R. Yangarber. Information extraction: Past, present and future. In Multi-source, multilingual information extraction and summarization : 23-49, Springer, Berlin, Heidelberg, 2013

20. J. Piskorski, K. Wieloch, M.Sydow. On knowledge-poor methods for person name matching and lemmatization for highly inflectional languages, Inf. Retr. 12(3): 275–299, 2009