Knowledge dissemination and discovery in Dutch excavation data; using AGNES to uncover hidden information

Alex Brandsen. Faculty of Archaeology & Leiden Centre for Data Science, Leiden University
Suzan Verberne. Leiden Institute of Advanced Computer Science & Leiden Centre for Data Science, Leiden University
Milco Wansleeben. Faculty of Archaeology, Leiden University

In this paper we will discuss the current state of knowledge dissemination for Dutch excavation data, and present AGNES; Archaeological Grey literature Named Entity Search, an online search tool aimed at uncovering the information hidden in excavation reports.

Over 60.000 reports are available online, and this number is growing by around 4.000 a year. The main reason for the existence of these reports is the Malta Convention[1]; a European agreement, aimed at protecting archaeological remains. Every excavation has to be documented and deposited according to Dutch law, which has created a collection of grey literature too vast to comprehend. Many of these reports are threatened to end up in a proverbial graveyard, unread and unknown.

Currently it is only possible to search through the metadata of these documents, mainly via the DANS (Data Archiving and Networking Services) repository. However, these metadata are often of poor and inconsistent quality. Also, an archaeologist will generally want to search more fine-grained, and might be interested in what is known as the 'by-catch opportunity'; i.e. a single Bronze Age find in a Medieval excavation, not mentioned in the metadata. There is a strong need for a better way to search through these documents. Also, archaeologists are eager to use multiple aspects in their searches; an example query might be to find all documents relating to the Iron Age, from a particular geographical area, that mention cremations. This is currently possible via DANS, but it is difficult and inaccurate. There is a strong need for a better way to search through these documents, as documented by e.g. Richards et al [2] and Dries [3].

To effectively index these texts, Named Entity Recognition (NER) is needed to correctly identify and distinguish between entities. Standard approaches to NER, and NER in related fields such as history, are insufficient to deal with the peculiarities of archaeological concepts and the wealth of potential classes. Some of the challenges include non-standard naming, extensive polysemy & synonymy, and complex word formation, including different spellings and concepts including capitals, numbers and symbols. This is particularly true for archaeological time periods, which can be expressed in numerous ways. For example, the following entities all equate to roughly the same time period: Neolithic, Swifterbant culture, Early Neolithic, New Stone Age, 3500 v.Chr, 5000 to 4000 BP and 4915 ± 40 Cal BC.

Some research has already been done on NER in archaeological texts in e.g. English[4,5] and Dutch[6,7], but these are not combined with full-text search, or tend to focus on limited entity types, and not the full breadth of archaeological concepts, which includes artefact, time period, place, material, ground context and monument. This means that currently there is no working system in place for Dutch archaeology.

In this presentation, we will present AGNES v0.2, in which machine learning is used to perform NER. The initial experiments use Conditional Random Fields and a feature set fine-tuned to archaeological concepts. The identified entities are combined with a full-text index to create an effective online search, allowing researchers to answer research questions that are currently impossible to solve.

[1] Council of Europe, "European Convention on the Protection of the Archaeological Heritage (Revised)," 1992.
[2] J. Richards, D. Tudhope, and A. Vlachidis, "Text Mining in Archaeology: Extracting Information from Archaeological Reports," in Mathematics and Archaeology, pp. 240–254, CRC Press, 6 2015.
[3] M. v. d. Dries, "Is everybody happy? User satisfaction after ten years of quality management in European archaeological heritage management," in When Valletta meets Faro. The reality of European archaeology in the 21st century (P. Florjanowicz, ed.), pp. 126–135, 2016.
[4] A. Amrani, V. Abajian, and Y. Kodratoff, "A chain of text-mining to extract information in archaeology," Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008., pp. 1–5, 2008.
[5] K. Byrne and E. Klein, "Automatic extraction of archaeological events from text," Proceedings of Computer Applications and Quantitative Methods in Archaeology, 2010.
[6] H. Paijmans and A. Brandsen, "Searching in archaeological texts: Problems and solutions using an artificial intelligence approach," PalArch's Journal of Vertebrate Palaeontology, vol. 7, no. 2, 2010.
[7] A. Vlachidis, D. Tudhope, M. Wansleeben, J. Azzopardi, K. Green, L. Xia, and H. Wright, "D16.4: Final Report on Natural Language Processing / Resources / Ariadne - Ariadne," tech. rep., ARIADNE, 2017.