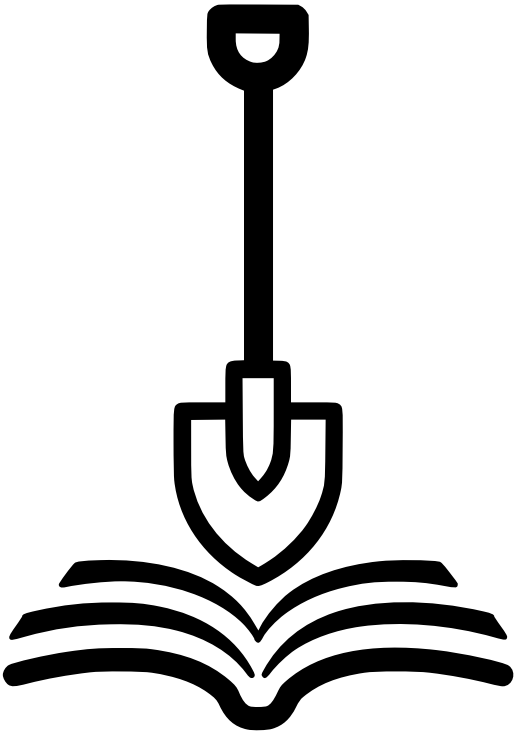
Digging in documents: using text mining to access the hidden knowledge in Dutch archaeological excavation reports

Brandsen, A.

Digging in Documents



Digging in Documents

Using Text Mining to Access the Hidden Knowledge in Dutch Archaeological Excavation Reports

PROEFSCHRIFT

ter verkrijging van

de graad van doctor aan de Universiteit Leiden,

op gezag van rector magnificus prof.dr.ir. H. Bijl,

volgens besluit van het college voor promoties

te verdedigen op dinsdag 15 februari 2021

klokke 11.15 uur

door

Alex Brandsen

Promotoren:	Dr. K. Lambers & Prof.dr. J.C.A. Kolen
Copromotor:	Dr. S. Verberne
Promotiecommissie:	Prof.dr. A.P.J. van den Bosch, Prof.dr. A.N. Brysbaert, Dr. M.G.J. van Erp, Prof.dr. D.R. Fontijn, Prof.dr.ir. W. Kraaij, Prof.dr. J.D. Richards & Prof.dr. H.G.D.G. de Weerdt

Printed by Gildeprint

Lay-out & cover design by Alex Brandsen
AGNES 'spade in book' logo designed by Mike Smethurst
Other vector elements used on cover provided by Freepik.com

All figures in this book are by the author

Acknowledgements

First of all, I would like to thank my grandmother Sietske Haker-Ples, who already knew I was going to be a doctor when I was five years old, and encouraged my curiosity for antiquities by taking me to museums. Also my parents, Silvia Pels and Piet Brandsen, for their never-ending support in my academic endeavours. And to Kayleigh Hines, for putting up with me being busy, stressed and/or absent minded during this research, for supporting me wholeheartedly, and even leaving her home and moving to the Netherlands with me. I would also like to thank Inge van Stokkom, for sending me the job ad for this PhD. Without her, I would have never known about this opportunity and would not be here.

Special thanks go out to my supervision team and promoters: Karsten Lambers, Suzan Verberne, Milco Wansleeben, David Fontijn and Jan Kolen. Each of you helped me tremendously with this research, and your guidance was invaluable. I could not have wished for a better support team.

I am grateful to Martin Koole and Femke Lippok for their excellent collaborations on the research described in Chapter 4 and 8 respectively. And to the entire focus group, for testing and providing feedback: Mette Langbroek, Femke Lippok, Arjan Louwen, Stijn van As, Rik Feiken and Ronald Visser. I would also like to thank Mike Smethurst for designing the AGNES logo.

I would like to thank all of my colleagues at the Faculty of Archaeology and the Data Science Research Programme, specifically Wouter Verschoof-van der Vaart, Anne Dirkson, Hugo de Vos, Daniela Gawehns and Gineke Wiggers, for exchanging ideas, helping me solve problems and inspiring me, as well as the great office banter.

Finally, I would like to thank the various institutes that supported this research: the Faculty of Archaeology and the Data Science Research Programme for funding this PhD position, the Leiden Institute for Advanced Computer Science for access to a web server and their computing cluster, the Leiden ALICE computer cluster for providing the computing power needed for Chapter 7, and the Leiden University Centre for Digital Humanities for funding part of the research described in Chapter 3.

Abstract

The archaeology domain produces large amounts of texts, too much to effectively read or manually search through for research. To alleviate this problem, we created a search system (called AGNES), which combines full text search with entity and geographical search. We first created a manually labelled data set to train a Named Entity Recognition model, which is used to extract entities from text. We also did a user requirement study, and usability evaluation on the system, to make sure it is suitable for archaeological research. In a case study on Early Medieval cremations, we show that using AGNES leads to a knowledge increase when compared to the knowledge of experts, gathered using previously available search engines. This shows that this kind of intelligent search system can help with literature research, find more relevant data, and lead to a better understanding of the past.

Samenvatting (Dutch abstract)

Archeologen produceren grote hoeveelheden teksten, te veel om effectief te kunnen lezen of handmatig te doorzoeken voor onderzoek. Om dit probleem op te lossen hebben we een zoeksysteem ontwikkeld (AGNES), dat zoeken in de volledige tekst van de docmenten combineert met zoeken op entiteiten en zoeken op een kaart. We hebben eerst een handmatig gelabelde dataset gemaakt om een *Named Entity Recognition* model te trainen, dat gebruikt wordt om entiteiten uit tekst te extraheren. We hebben ook een studie gedaan naar de gebruikerseisen en een evaluatie van de *usability* van het systeem, om er zeker van te zijn dat het geschikt is voor archeologisch onderzoek. In een case studie over Vroeg-Middeleeuwse crematies, laten we zien dat het gebruik van AGNES leidt tot een toename van kennis in vergelijking met de kennis van experts, verzameld met behulp van eerder beschikbare zoekmachines. Dit toont aan dat dit soort intelligente zoeksystemen kunnen helpen bij literatuuronderzoek, meer relevante gegevens kunnen vinden, en uiteindelijk kunnen leiden tot een beter beeld van het verleden.

Contents

List of Figures

List of Tables

1 Introduction

"*A library serves no purpose unless someone is using it.*"
Mr. Atoz, Star Trek TOS, s03e23 'All Our Yesterdays'

In the last decade, archaeology has joined other disciplines and has started generating what is known as 'big data': "Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value" (De Mauro *et al.*, 2015, p. 103). The challenge is how to best mine, mix and analyse these incredibly rich data sets. While a lot of research in this area is dedicated to processing structured information – such as spatial information, databases, and so on – less attention has been given to processing unstructured information: the texts describing archaeological research (Bevan, 2015).

Easy access to the information hidden in these texts is a substantial problem for the archaeological field. Making these documents accessible, searchable and analysing them is a time consuming task when done with the existing metadata search systems or by hand, and will generally lack consistency. In addition, in the last fifteen years or so, the amount of archaeological texts in general – and fieldwork reports specifically – have seen an explosive growth. It is practically impossible to keep up with the rate of documents being produced, and the available literature is so extensive that the current search systems are not effective enough for detailed search. Text Mining[1] provides methods for automatically extracting meaningful information from these large data sets, allowing researchers to locate texts relevant to their research questions, as well as being able to identify patterns in the literature (Richards *et al.*, 2015).

This dissertation will describe the use of Text Mining techniques on Dutch language grey literature (further defined in Section 2.2.2): field reports from excavations in the Netherlands, deposited in the Data Archiving and Networked Services (DANS) digital archive. More detailed information on the data set will be provided in chapter 2 and 3.

1.1 Research motivation

The work carried out in this project is motivated by the need of researchers in the archaeological field to be able to efficiently find information related to their research questions in the available literature. This requirement has been well documented around the globe (e.g. Richards *et al.*, 2015; Van den Dries, 2016; Habermehl, 2019) and some studies have investigated different applications of Text Mining in archaeology in English (Vlachidis & Tudhope, 2016; Amrani *et al.*, 2008; Byrne & Klein, 2010), French (Mélanie-becquet *et al.*, 2015) and

[1] Please see the Glossary at the end of this document for a definition of this term, as well as other technical terms and acronyms used throughout this dissertation.

Dutch (Paijmans & Brandsen, 2010; Vlachidis *et al.*, 2017, also see section 2.4).

However no system is currently in place that allows easy access to the full text in the Dutch archaeological corpora (document collections), meaning that relevant and valuable information is not being utilised by researchers. This is a problem from a theoretical point of view, as key information that is currently being overlooked could change archaeological interpretations, but it also devalues the monumental effort that has gone into collecting, digitising, archiving and publicising these documents, as well as the legislation that has been drawn up surrounding the archiving of these documents. The scientific value of these reports will be further discussed in section 2.2.4.

More and more text and data mining tools and techniques have become available over the last years, which provide a way to access and extract information from this wealth of information currently hidden in the text data of these reports. This, combined with the relatively easy access to higher computer processing power available to us now (see section 2.5.2), makes a systematic implementation in Dutch archaeology not only feasible, but also highly desirable.

The end goal of this project is to develop a search engine that combines Text Mining techniques with a full text search, allowing archaeologists to search through archaeological reports stored in the DANS repository (also see section 2.2.3). The search system is named AGNES, and will be further discussed in the following chapters.

1.2 Research questions

When creating a tool for archaeologists, it is important to ensure it can positively impact their work. Unfortunately, many digital tools in archaeology have been created, and subsequently ended up unused in a corner of someone's web server, or even worse, not available online at all, such as the OpenBoek project (Paijmans & Wubben, 2008) and the work by Tudhope *et al.* (2011) (also see section 2.4). To ensure an impact is made, it is required to investigate the user requirements, as well as the effectiveness and usability of the developed system, but also to evaluate AGNES's output using a real-world case study.

Main Research Question: To what extent can a search engine using Text Mining improve archaeological research and aid information discovery in grey literature data sets?

To answer this question, the following subquestions have been formulated:

1. Can we use existing labelled data sets for Named Entity Recognition[2] in the archaeological domain, or do we need to create our own data set? If so, to what extent does the accuracy increase? (Chapter 3)
2. To what extent can we automatically generate time period and site type metadata for Dutch excavation reports? (Chapter 4)
3. Which questions do archaeologists want to ask of this data set, and which user requirements do they have for a search system? (Chapter 5)
4. What user interface features of AGNES are experienced as positive or negative, and how can we optimise the usability of the system for archaeologists? (Chapter 6)
5. To what extent does adding more domain-specific training data to BERT models improve Named Entity Recognition accuracy? (Chapter 7)
6. What is the impact of the developed system on archaeological research? (Chapter 8)

Sub questions 3 & 6 are closely linked, as the impact of AGNES will be evaluated using a case study. The case study research question is defined in chapter 8.

1.3 Research Methodology

The research described in this dissertation is relatively varied, and as such a number of different research methodologies are combined. For the development of the search engine, we apply the Agile principles, which are described below. To evaluate classification methods, we use human labelled data, and to evaluate the system functionality we use a focus group and case study.

1.3.1 Agile Development Principles

To define the agile development principles, and to explain why they are relevant to this project, it is useful to draw an analogy to a non-programming project. Highsmith *et al.* (2002) uses a battle as an example: a commander will plan extensively, but also realise that plans are just the beginning. Probing enemy defences (creating change) and responding to enemy actions (responding to change) are more important. A commander is successful by defeating the enemy (the mission), not by strictly conforming to a plan. Battlefields are uncertain, constantly changing, and turbulent, so planning everything up front simply isn't feasible. A

[2]Further defined in section 2.3.3

similar parallel can be drawn to archaeological excavations: although they are planned to a high degree, initial findings in the field can massively change the approach taken during excavation.

Both the battlefield and excavation scenarios are examples of projects with a relatively clear mission, but the specific requirements to complete that mission are (partly) unknown, volatile, and constantly evolving as change unfolds. Many programming and research projects are of a similar nature, where the full extent of the work is not known from the onset. These projects are dubbed "high-exploration factor projects" (Highsmith *et al.*, 2002, p.4), and this is where agile approaches are most suitable, as they can not simply be completed by plan-driven methods. The more experimental the technology and the more volatile the requirements are, the more agile development improves the chance of success.

This project certainly falls under this category as well: while the mission is clear (to disclose archaeological reports), the specific techniques that should be used to attain this goal and the exact user requirements were not clear at the start of the project. As this makes it a prime candidate for agile development, it has been decided to use this methodology for this project.

As a more formal definition, agile development is a combination of a philosophy and a set of development guidelines. The philosophy encourages user satisfaction, quick, incremental development of systems, minimal development work products, and overall development simplicity. The guidelines describe iterative feature-driven cycles, delivery over analysis and design, and continuous communication between the developer and the end users, often by using user focus groups (Pressman, 2005).

In the Manifesto for Agile Software Development, Beck *et al.* (2001) describe four main ideas to define agile development:

- **Individuals and interactions** over processes and tools
- **Working software** over comprehensive documentation
- **Customer collaboration** over contract negotiation
- **Responding to change** over following a plan

This describes the agile mindset in general terms. In more practical terms, in the software development in this project:

- The first version of AGNES shall be a proof of concept, with only basic capabilities, to act as a starting and discussion point, enabling better feedback
- The system shall be developed in small cycles, each leading to a working prototype that can be demonstrated and assessed

- After each cycle, a user panel will provide feedback on AGNES to integrate into the next development cycle. More information on the user panel can be found in section 1.3.3

While this specifically describes the software development process, the rest of the research is undertaken with the same principles in mind. This makes the research more suited to be published in papers than a monograph, which is further detailed in section 1.5 below.

1.3.2 Machine Learning Assessment with Labelled Data

To develop and evaluate supervised machine learning technologies, we use human-labelled data. In Chapter 3 we describe the problems with the existing labelled data set for Named Entity Recognition, and how we created a new one which led to better results for this task. This labelled data is used to evaluate the work in Chapters 5 and 7. For the document classification task in Chapter 4, we created a labelled data set by converting the manually added metadata to a controlled list of labels. For any of these tasks, we report the recall, precision and F1 score, as further described in section 2.3.4.

1.3.3 System Evaluation by Focus Group

In this project, a focus group of potential users has been assembled, including academic researchers, PhD students, commercial archaeologists, and archaeologists working at the government level. An initial meeting has been organised with the group to synthesise a list of user requirements: the objectives for the system (chapter 5). Based on these requirements, the first iteration of AGNES has been created. The user interface of this version was evaluated, and the feedback was used to improve the system (chapter 6). Finally, we evaluated the quality of results retrieved by AGNES in a case study (chapter 8).

1.3.4 System Evaluation by Case Study

There is an abundance of research questions that could be answered if archaeologists had easy access to the full text of the reports. One example is presented by the 'by-catch opportunity': many excavations focus their research questions on a specific time period (e.g. on a Roman cemetery), but often also reveal objects and features from other periods (e.g. a small cluster of flint objects or a single stone axe from the Stone Age) or other types of contexts (e.g. a single residential

find). These other finds will always be presented in the publication, but will escape the attention of Stone Age archaeologists since they probably would ignore a report titled 'The Roman cemetery at Vlodrop (Limburg)'. However, these individual finds and small clusters do express a very valuable component of human behaviour, often called off-site (or off-settlement) activities (Foley, 1981).

In this project, Femke Lippok – a PhD researcher at Leiden University – has formulated an archaeological research question, which will be used as a case study, or test case, for the system. The first stage is to create a set of baseline results using current approaches, i.e., using existing search systems and the archaeologists' knowledge of their field.

Once this baseline is established, it is possible to analyse and compare this to the results obtained by AGNES qualitatively (comparing the interpretations) as well as quantitatively, via statistical and geographical analyses of the resulting document sets. This way we can demonstrate the increase in knowledge discovery by using the system, as well as the change caused to the archaeological view on a particular research question by being able to integrate more knowledge into research. More information on the case study can be found in chapter 8.

1.4 Contributions

Of course, the main contribution of this research is the AGNES search system, which is currently online and being used by archaeologists for their literature research. Besides AGNES, we also contribute a number of data sets, software and language models that can be used in other research, and we end this section with an overview of scientific contributions.

1.4.1 Software and Data

An application-oriented research project such as this inevitably produces resources that can be used by other researchers. Below is a list of the most prominent data sets and software shared publicly:

1. A manually annotated training data set for Named Entity Recognition (NER) in the archaeology domain (doi.org/10.5281/zenodo.3544543)
2. A training data set for the classification of archaeological reports on time period and subject (doi.org/10.5281/zenodo.3676702)
3. A JavaScript Object Notation (JSON) export of all the entities extracted from our corpus (doi.org/10.17026/dans-zcs-7b72)

4. A trained Conditional Random Fields (CRF) model for Dutch archaeological NER, with code to generate such models (doi.org/10.5281/zenodo.1238860)
5. A BERT model further pre-trained on our corpus, called ArcheoBERTje, and a specific model for NER inference, hosted on HuggingFace for ease of use (huggingface.co/alexbrandsen)

1.4.2 Scientific Contributions

Besides the resources described in the previous section, we make the following unique contributions to the scientific field.

In Chapter 3 we show that annotation of a NER data set with rigorous annotation guidelines, tailored to machine learning, leads to higher performance than a previously available data set. We also argue that for NER data, the pairwise F1 score between annotators is a better indicator of Inter Annotator Agreement than the commonly used Cohen's Kappa.

In the following chapter (4), we investigate the difficult task of multi-label text classification with many classes. We are the first to do this in the archaeology domain, and show that this method can contribute to either faceted search (by filtering documents by topic) or even metadata assignment at the time of deposition in an archive.

In Chapter 5 we show the need for a system such as AGNES, and make a case for the adoption of user requirement solicitation and short development cycles for digital tools in the archaeology domain. We also present our CRF based NER method which outperforms previous rule-based approaches.

The usability evaluation of our search system described in Chapter 6 is (as far as we know) the first of its kind in the archaeology domain, and we contribute to the general discussion of information needs in archaeology. We show the importance of a diverse group of test users, and argue that usability evaluation should be a core part of tool development.

Chapter 7 describes our work on the use of BERT language models for NER in Dutch archaeology. We present the first Dutch domain specific BERT model, which is also the first archaeology specific BERT model. We show that adding language-specific and domain-specific training data to an existing language model (by further pre-training) increases the performance of the model.

Perhaps the most important contribution for archaeologists is described in Chapter 8: in this case study we show that for Early Medieval cremations, using AGNES increased the amount of sites known to experts by 30%. This indicates that this type of search through grey literature can lead to more efficient and

more detailed research.

And finally, in Chapter 9, we contribute to the discussion on development-led archaeology more broadly, and how computational tools might solve existing problems and shape future research.

1.5 Dissertation outline

This dissertation consists of a collection of papers, sandwiched in between the introduction / background chapters and a discussion chapter. A majority of the papers have already been published in – or submitted to – peer-reviewed journals and conference proceedings during the course of the PhD. The papers are not in chronological order of publication, but in the order that makes the most sense for the narrative. Each paper can be read independently from the other chapters.

In the following Chapter (2), we will give an overview of the current state of affairs in Digital Archaeology, grey literature, and the value of excavation reports. We will also introduce Text Mining techniques, so the following chapters can be understood by anyone. Finally, we present previous research on Text Mining in archaeology, and the resources we use for this research.

In Chapter 3 we discuss the difficulties with an existing training data set for Named Entity Recognition, and how we have created a new data set with rigorous guidelines that improves the accuracy (Brandsen *et al.*, 2020).

Chapter 4 describes the work in collaboration with Martin Koole, where we trained a number of models to automatically classify excavation reports in subject and time period categories. This information can then be used for faceted search: allowing users to filter documents based on these categories (Brandsen & Koole, 2021).

The following chapter (5) describes the user requirements solicitation process, where we held a workshop with users to determine what features they would like when searching through excavation reports. The results of this process are the basis for how we developed the search system. We also describe the first version of AGNES (v0.1), and how it was used to elicit more feedback from the users (Brandsen *et al.*, 2019).

In Chapter 6, we evaluate the user interface created based on the input of the previous chapter. We specifically look at how quickly the users learn the interface, and which interface components are experienced as positive and negative. The outcomes have been used to improve the search system (Brandsen *et al.*, 2021b).

Chapter 7 describes how we experimented with different BERT language models to perform Named Entity Recognition, and how adding more domain-specific

training data increases the accuracy (Brandsen *et al.*, 2021a).

The case study is described in Chapter 8, where we worked together with Femke Lippok to investigate Early Medieval cremations. We used AGNES to retrieve relevant excavation reports, and assessed to what extent these documents are new information to the researcher and to what extent her view of this topic changed.

In Chapter 9 we discuss the results, and what these mean in a wider context. We then provide answers to the research questions in the conclusion, and end with proposed future research.

2

Background

"*Those who can imagine anything, can create the impossible.*"
Alan Turing

Creating new digital technologies, or new applications of existing technologies, often requires imagination and creativity. And as we see over time, things that we thought impossible – or even inconceivable – ten or twenty years ago, have become commonplace in research, but also in society as a whole. Think of the internet, mobile phones, and artificial intelligence, all examples of phenomena that even most computer visionaries could only imagine in science fiction. Yet today, these technologies are ubiquitous and have changed science and society profoundly.

In this chapter, we provide a background on digital archaeology and big data, the life cycle and properties of the excavation reports, and we give an introduction to Text Mining techniques. With this background, we aim to make it possible to read and understand the following chapters, which due to publishing page limits, might not have the level of explanation needed for non-experts. This is particularly true of the more technical chapters (4, 6 & 7). At the end of the chapter, we provide an overview of previous research on Text Mining in archaeology, and end with an overview of the resources used for this research.

2.1 Digital Archaeology and Digital Humanities

In general, archaeologists have always been eager to apply and adapt methods from other sciences to their own research, and computer science is no exception. In the last half of the 20th century, the constant technical innovation in computer science meant we had more and more digital tools available to help us research the past, leading to "tool-driven revolutions" (Schmidt & Marwick, 2020, p. 1). In the last twenty years or so, this trend accelerated even faster, and digital technologies are nowadays ubiquitous and pervasive within archaeology (Zubrow, 2006).

However, 'Digital Archaeology' does not have an agreed-upon definition, and many authors have defined the term in various ways. Zubrow (2006) defined it – rather poetically – as: "the use of future technology to understand past behavior", although perhaps 'future technology' is a bit of a misnomer, as we archaeologists tend to use techniques that are already considered 'old' in other fields of science. Averett *et al.* (2016) are a bit more practical: "Digital Archaeology is the use of computerized [...] tools and systems aimed at facilitating the documentation, interpretation, and publication of material culture". However, using this definition makes just about any archaeology 'digital', as practically all research uses databases, spreadsheets, or at least word processors and the internet to write and disseminate work. This is also reflected on by Morgan & Eve, who state that "we

are all digital archaeologists" (Morgan & Eve, 2012, p. 523), and Costopoulos (2016) notes that this has been the case for at least 20 years.

Perhaps a more refined definition could be: research where the use of digital tools is a principal component in the analysis, presentation, and/or dissemination of archaeological data. If we look at the Computer Applications and Quantitative Methods in Archaeology (CAA) conference, the oldest and most influential digital archaeology conference, this definition fits most, if not all of the research presented there. Interestingly, the research presented in this dissertation does not fully fit in this definition, as no data is analysed, presented, or disseminated directly. Perhaps it can be considered 'meta digital archaeology': building tools for archaeologists to do digital archaeology with. The prefix 'computational' is often used in other fields to describe the development of computational tools, so computational archaeology could be a good fit, however in archaeology this term is practically synonymous with digital archaeology.

However digital archaeology is defined, we can not say it is just about making our research simpler and easier. The digital tools we use have had a profound effect on archaeological theory and how we view the past. This is particularly true of visualisation tools and the way we can now easily disseminate information to colleagues (Tanasi, 2020), which accelerate and influence our ideas about material culture.

The field of Digital Humanities is in many ways similar to Digital Archaeology. Both are interdisciplinary, and deal with digital methods and technology to study humanity. However, as Huggett (2012) notes, Digital Archaeology does not feature often in Digital Humanities journals, nor do archaeological publications mention Digital Humanities very much. It seems that Digital Archaeology is not a subfield of Digital Humanities but stays largely separate.

Perhaps the reason for this is that historically, the humanities have focused more on textual data, while archaeology mainly produces tabular and geospatial data. However, overlap can be found in specific areas, for example in 3D visualisations and methods like network analysis we increasingly see humanities and archaeology scholars collaborating and sharing expertise. As in archaeology we do not analyse texts often, research in this area is sparse, but some scholars have experimented with computational approaches (also see section 2.4).

In this dissertation, texts are the main source of data, and as such this study lies perhaps closer to Digital Humanities than most Digital Archaeology research. This is mainly from a methodological point of view, as our texts are secondary sources, while in Digital Humanities, the texts tend to be the primary sources.

2.1.1 Big Data in Archaeology

In recent decades, the biggest change in archaeology has been caused by the impact of the Information Technology revolution, having introduced new digital and statistical methods that have changed much of the way we do archaeology. This revolution greatly affects the way archaeological data are collected, analysed, and disseminated. This makes methods that were previously too complex or time-consuming achievable on standard desktop PCs (Levy, 2014). However, the use of these new digital techniques also creates problems. The amount of data created is many times greater than with non-digital methods, creating what is known as 'big data'; massive volumes of data that are so large, often multiple terabytes, making them difficult to process using traditional database and software techniques (Bloomberg, 2013).

Although Big Data lacks a clear and consistent definition – like many other tech buzz words – it is usually defined with the four V's: volume, velocity, variety, and veracity (De Mauro *et al.*, 2016). Another commonality between different definitions is the idea that the data is so unruly and large that innovative methods and large amounts of computing power are needed to process and analyse the data (Bloomberg, 2013; Gartner Glossary, 2021; Boulton *et al.*, 2012). This shift in scale of analyses is evident in most disciplines, and we can see that the processing of large amounts of data has the potential to produce insights that were previously impossible and unimaginable (Wesson & Cottier, 2014).

Most archaeological data does not have a particularly large volume, as our data sets are often small compared to other disciplines (under 1GB). However, we have seen a shift from the past, where data scarcity was a prevailing issue, to much larger data sets now, relatively speaking. This is partly due to more and more legacy data being made available freely, digitally, and often linked, and partly due to archaeologists 'borrowing' data from other fields, such as remotely sensed data and other sources from the environmental sciences.

The velocity, or the speed at which the data updates, tends to be very slow compared to some other disciplines, so that is another V we generally do not deal with. Although we create thousands of data sets and documents per year, this is not comparable to e.g. social media posts, being created by tens of thousands per second.

Variety is one aspect that almost all archaeological data tends to have: we record data in a multitude of mediums and formats, including databases, photos, geospatial data, texts, and drawings. And the variety between data sets is large as well, as many different formats and standards are used, if a standard is used at all.

Veracity has two aspects: truthfulness and quality. We can assume that most – if not all – archaeological data is truthful, or at least not purposefully false. However, when we talk about quality, and the related concept of completeness, we can see that archaeology does struggle with this V to a large extent, even on small data. At a conceptual level, all archaeological data is incomplete, and fuzzy or inaccurate to various degrees. At a practical level, we see that data can in some cases be low quality due to e.g. recording methods, data formats, or – in the case of this project – Optical Character Recognition (OCR) mistakes causing noise in our texts.

Another way we can look at Big Data is simply that it is too much to handle effectively. The problem of having too much data has been outlined by multiple researchers, with Vince noticing "we are drowning in our own data" (Vince, 1996, p. 1), and Bevan describing this problem as the "data deluge" (Bevan, 2015, p. 1). Certainly when we look at the amount of data being generated in the Netherlands, both as text and in other formats, we can conclude that there is too much to keep on top of.

Other authors have suggested Big Data is less about data that is big, but about the capacity to search and cross-reference large data sets (Boyd & Crawford, 2012) and working with (almost) all available data that can be useful to solve a question (Mayer-Schönberger & Cukier, 2013). This takes a more relative approach, looking at Big Data as All Data. And it is these viewpoints that are more commonly used when dealing with and discussing Big Data in archaeology, as it allows for more choices when exploring data from different angles, and to comprehend aspects we cannot understand using smaller data. Another aspect of Big Data is modelling, applying methods to large quantities of data to infer probabilities and make predictions from patterns in the data (Gattiglia, 2015).

All that being said, there are some examples of data in archaeology that really are large in volume. The most well-known example is remotely sensed data, which can be multiple terabytes, depending on the geographical scale. Now methods to wrangle this Big Data are becoming more accessible, we are seeing a lot of research in this area (Cowley, 2012; Bennett *et al.*, 2014; Traviglia & Torsello, 2017; Trier *et al.*, 2018; Lambers *et al.*, 2019; Verschoof-van der Vaart & Brandsen, 2020; Fiorucci *et al.*, 2020).

The other main source of big data in archaeology are texts, often collected in repositories at a large scale, these collections can easily have large volume and variety, and with thousands of reports being added each year, they have a relatively high level of velocity compared to other archaeological data. However, as also noted by Bevan (2015), much less research is dedicated to analysing this unstructured data.

2.2 Data

In this section, we describe the origins and properties of the data used in this research. We first discuss the legal reason these reports are produced – the Malta convention – and then provide an overview of grey literature, the Findability, Accessibility, Interoperability, Reusability (FAIR) principles and archives, the importance of archaeological reports, and finally, some properties specific to this data set.

2.2.1 Malta Convention

The Malta Convention (also known as the Valletta Treaty) is a European treaty, signed on 16 January 1992. It came into effect on 25 May 1995, and its aim is to protect archaeological remains by making "the conservation and enhancement of the archaeological heritage one of the goals of urban and regional planning policies" (Council of Europe, 1992, Art. 1). The convention was implemented in the Netherlands through the Archaeological Heritage Management Act (*Wet op de archeologische monumentenzorg*) in 2007 (Ministerie van Onderwijs Cultuur en Wetenschap, 2007). Any traces or remains of past human behaviour are considered part of the archaeological heritage. This includes structures, constructions, groups of buildings, developed sites, movable objects, monuments of other kinds as well as their context, whether situated on land or under water. Preferably, preserving these remains is done by keeping them in situ, but when this is not possible, the developer disturbing the ground record pays for the archaeological research. This development-led research is generally performed by commercial archaeology units.

The Malta legislation led to a big increase in the amount of archaeological research being performed, due to the 'developer pays' principle and the obligation to handle archaeological remains with due care in spatial plans, amongst other things. All this archaeological research has created a collection of texts that is too vast to comprehend. The number of reports created in the last 20 years is currently estimated at around 60,000 and is growing by approximately 4,000 per year (Rijksdienst voor het Cultureel Erfgoed, 2019a). Most of these reports are categorised as 'grey literature', and are likely to end up in a proverbial 'graveyard', unread and unknown, unless they are properly archived, disseminated and indexed.

2.2.2 Grey Literature

The term grey literature is used to describe a collection of documents which are not published in the traditional sense of the word, both in hard copy and digitally. In 1997, at the *Third International Conference on Grey Literature*, a definition was agreed by participants: "that which is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers" (Farace & Schoptimefel, 2010). It is not a term used exclusively in the archaeological domain, but stems from the library and information science communities (Falkingham, 2005).

One of the first, and pivotal, discussions of grey literature ironically does not actually use the term grey literature, but is called *Use of Reports Literature* (Auger, 1975). It was only in 1989 that Auger phrased the term in his publication *Information Sources in Grey Literature*. From these early discussions to more recent studies (Roth, 2010), it seems that grey literature is more or less synonymous with reports literature, although it can also include conference proceedings, official documents and theses.

Researchers often perceive grey literature as being of lesser quality than traditionally published 'white' literature, as it is not peer reviewed and does not necessarily have quality control (although often rigorous standards and requirements exist for these documents). This leads to the unfortunate perception that the researchers who publish grey literature are of lesser quality also. Just using the word 'grey' changes our perception of this literature, as it conjures connotations of 'dull', 'drab', and other similarly negative concepts (Roth, 2010).

However, while grey literature does not have the prestige and rigour of more traditional publishing, it does provide "greater speed, greater flexibility and the opportunity to go into considerable detail" (Auger, 1989, p.3). These reports generally contain comprehensive, detailed, and up to date information on research findings. Even when a traditional paper is published as white literature, generally detailed information, techniques, and results are omitted. To gather information of importance, grey literature is often the most direct source of information (Falkingham, 2005).

Grey literature is generally created to disseminate information, not to sell for profit. This means that practically, it does not have the advantage of the publicity and marketing normally associated with commercial texts. Combining this with substandard bibliographical information, low print runs, and nonuniform digital storage means that these documents can be extremely inaccessible (Auger, 1989).

Grey Literature in Archaeology

In the early 1990's, the term grey literature first started appearing in the archaeological world in cultural resource management documents in the United States (Seymour, 2010), although of course the actual grey literature – the research reports – are much older. It was not until 1996 that the term made its way across the Atlantic and appeared in the editorial of the first *Internet Archaeology* (Vince, 1996), which discussed photocopies of full site reports stored in archaeological archives. Since this time, the concept of grey literature has not changed a great deal, but the way we store and access these reports has. While some reports do still get printed and deposited in depots as hard copies, generally these documents are created digitally and stored in e-depots or repositories (see section 2.2.3).

Over the last 20 years or so, we see a dramatic increase in the number of reports being produced by the commercial archaeology sector. However, editing and proof reading are generally undertaken in house, if at all, and quality control remains an issue as no peer review is done on these documents (Falkingham, 2005). This is also partly due to the competition between archaeology units and the reluctance of developers to pay enough money. This lack of funding directly translates into hurried work that perhaps is not always as polished as we would like. Also, there is no incentive for commercial units to go beyond what is required by law, so reports are often the bare minimum as prescribed by regulations.

However, these reports should not be considered of lesser importance than traditional academic output. Archaeology is fundamentally scholarly, whether in a commercial or academic setting. Both types of archaeology use the same methods, are highly demanding of intellectual excellence, use the same theoretical building blocks, and are being conducted by people with the same degrees (Athens, 1993; Seymour, 2010).

In addition, while there is no formal peer review, there are rigorous regulations that these reports must adhere to by law, at least in the Netherlands (Stichting Infrastructuur Kwaliteitsborging Bodembeheer, 2016). So while there might be no review by peers, the quality is fairly uniform due to these regulations.

A recent study undertaken by Wiseman & Ronn (2020) used the Covid-19 pandemic as an opportunity to assess how archaeologists access literature. They asked archaeologists that were furloughed in the United Kingdom to volunteer for an information seeking task on certain subjects. While there is generally a perception that archaeological grey literature is of lesser value than traditionally published material, the volunteers in this project rated grey literature as more useful than monographs, even when the monographs are digitised and searchable.

Their volunteers also note that they would like to do word searches, but are currently unable to, something we are making possible within this project.

In a way, perhaps the field reports are actually more important than some more theoretic or synthesising academic research, as commercial units are much more concerned with documenting intrusive investigations where the archaeology is destroyed. The reports and associated data are invaluable as they are the only remaining evidence we have of excavations, unlike theoretic and synthesising research which is generally repeatable and reproducible.

2.2.3 Repositories

In the Netherlands, archaeological companies who perform research are required by law to deliver a report describing the research (Ministerie van Onderwijs Cultuur en Wetenschap, 2015, Art. 5.6). As stipulated by the *Stichting Infrastructuur Kwaliteitsborging Bodembeheer* (SIKB), a report has to be deposited into an e-depot within two months of completing the project (Stichting Infrastructuur Kwaliteitsborging Bodembeheer, 2016, Art. 2.6.2). While some companies and municipalities are still reluctant to deposit their reports into national e-depots, most reports do end up in one of three of the main e-depots of the Netherlands; the DANS repository, the *Rijksdienst voor het Cultureel Erfgoed* (RCE) Document Management System or the *Koninklijke Bibliotheek* (KB) e-Depot. There is some overlap between the DANS, RCE and KB data sets, and between them it is estimated that they hold around 60 to 70 percent of all Malta reports (Rijksdienst voor het Cultureel Erfgoed, 2019a). DANS has also been working on digitising and archiving older reports from before the Malta legislation.

2.2.4 The Importance of Archaeological Reports

It is crucial that the knowledge gained from development-led archaeological research leads to new insights into the past. These new insights allow more accurate archaeological predictions to be made, on which heritage management policy can then be based. The main way in which information from this research is disseminated is via archaeological reports, and as such, these documents contain a wealth of information.

This information potential was harnessed by a special *Nederlandse organisatie voor Wetenschappelijk Onderzoek* (NWO) grant programme "*Oogst voor Malta*" (Valetta Harvest), in which a limited number of specific thematic and/or regional projects were allowed to address specific archaeological voids. This programme illustrates the new role of academic archaeologists within the field heritage man-

agement. Their added value is to synthesise the results of all these commercial excavations into new archaeological theories and update our views on existing ones; to sketch a coherent picture of the behaviour of people in the past. They combine information across the country, across different types of sites and preservation conditions for specific time frames (Theunissen & Feiken, 2014; Habermehl, 2019).

Another interesting property of these reports is that they are almost a random sample. Before Malta, most excavations were aimed at finding specific types of archaeology, based on prior knowledge or whichever period and region the researcher was interested in. This means that there was a bias in which certain periods and regions were researched more. However, the Malta reports are generated whenever building work is done, which is a much more random pattern, thus giving us a less biased sample to work with. However, there is still a bias that is introduced by the fact that some regions in the Netherlands simply have more building work going on than others. For example, Theunissen & Feiken (2014) mentions that there is a lot of information available about the sandy areas of Noord Brabant, but much less excavations have been executed in the peaty areas of Friesland. Also, some areas simply see a very limited amount of building work – or none at all – such as bodies of water and protected nature areas.

Nevertheless, the current situation gives us information across the entirety of the Netherlands, allowing for broad synthesising research that was previously impossible. And to do this research, we must be able to find, access, and reuse the data generated in these excavations.

2.2.5 The FAIR Principles

It is important that the results of any research are open, meaning it is accessible for free for anyone. This will lead to better science, as checking each others work and further building on it can only improve our research. It is also a question of fairness, most research is indirectly funded by the tax payer, and as such, the results from that research should be available to them. The Open Access and Open Science movements are making headway in making science more open, which is a great development. However, just making the results available is a first step, but not an end goal, as the data needs to be reusable for it to have any effect.

The FAIR (Findability, Accessibility, Interoperability, Reusability) principles are a set of guidelines that aim to improve this data reuse. There is an emphasis on machine readability, as the size of data sets are increasing, and researchers are relying on computational support to deal with the data. The first step in

reusing data is to actually find them, and that is where this project tries to make archaeological reports more FAIR. Specifically, we help with the fourth item of the Findability principle: "F4. (Meta)data are registered or indexed in a searchable resource" (Wilkinson *et al.*, 2016, p. 4). Currently, the metadata is registered in a searchable resource, but the data itself (the text in the reports) is not. The system we describe in this project will make the text data itself also searchable, which hopefully will lead to more data reuse.

2.2.6 Document Properties

Archaeological reports contain a large amount of descriptive details. This includes lengthy descriptions, many illustrations, and tabular data about the discovered finds and their context. These publications often follow a distinctive chapter/section division that has a semantic meaning (by period, by material category, by type), which would ideally be incorporated into the Text Mining. Kintigh (2015) specifically mentions that the scope of natural language statements is often not implicit, but inferred from a hierarchy of chapters and sections. Kintigh uses units within sites as an example, but what we see more often in Dutch reports is e.g., the snippet "we have found an axe" in the section "Neolithic", indicating a Neolithic axe. The section heading might be paragraphs – or even several pages – before the snippet, so there is no direct relation within the vicinity of the text. Apart from the complexity of the text itself, this 'semantic inheritance' makes extracting information or finding relations difficult.

However, these documents differ largely in internal structure from commercial unit to unit. Since no commercial publisher is interested in these large volume books anymore, most archaeological organisations publish these reports in their own internal series. While there are regulations for the content of the reports, the order, structure and format is not prescribed (Stichting Infrastructuur Kwaliteitsborging Bodembeheer, 2016), and as such we see a large variety. This is not a Dutch only problem, as this problem is also noted by Wiseman & Ronn (2020) for reports from the United Kingdom. The inconsistencies make extracting the heading structure a challenging task. A compounding factor is the format the documents are stored in: Portable Document Format (PDF) files are notoriously difficult to extract structured text from, as it is a format geared towards correctly displaying text and any structure that the text might have is lost. When extracting text from PDFs, we can get information about font style and size for example, but nowhere are certain snippets marked as being a heading. We have experimented with a rule-based approach to automatically label chapter and section headings, but due to the noise from PDFs and the different styles between

documents, we found it incredibly difficult to do this with a decent level of accuracy. A machine learning approach for this task might be better suited, but as there is no training data, this is outside the scope of this dissertation. In section 9.9 we discuss this further.

Most grey literature reports are in Dutch, but many archaeologists write in English as well, and we even found some German in our data set. Ideally we would address all three of these languages, but this adds a level of complexity beyond the scope of this project. As such, for now we focus just on Dutch, which will cover the majority of our data set. In a follow up project, we will work on adding English and German as well, which is further described in the Future Research section (9.9).

A small part of the data set are scans from hard copy reports, which have been converted to digital text using OCR. The OCR process introduces some noise – especially on older reports – as it is not a perfect method. However, this should not cause too many problems as it is only used in a minority of reports.

2.3 Introduction of NLP and Information Retrieval Concepts

In this section, an overview is given of relevant concepts that are useful to understand further chapters.

Natural Language Processing (NLP) as a research field explores how computers can be used to understand and manipulate natural language, i.e. speech and written text in human language (as opposed to formal/constructed language such as programming languages) (Chowdhury, 2005). It is a rather broad field on the intersection of linguistics, computer science and artificial intelligence, and is used to process and analyse large amounts of text. It originates in the 1950s, and was originally quite separate from Information Retrieval (IR), but over time, NLP and text IR have converged to some extent (Nadkarni *et al.*, 2011).

2.3.1 Information Retrieval

Information Retrieval can be defined as "a field concerned with the structure, analysis, organisation, storage, searching, and retrieval of information" (Salton, 1968, p. V). The field has made significant advances in the last fifty years, but this definition from 1968 is still appropriate, even though nowadays the focus lies more on the last two items: searching and retrieval of information. The type of information is most often text documents, and since the rise of the internet,

web page search is one of the key areas of research. In comparison to tabular (database) data, text data is unstructured, and the complicated task of computers 'understanding' language to retrieve documents relevant to a user's search goal (or information need) is at the core of IR (Croft *et al.*, 2010).

The concept of information needs is also worth discussing here, as it will be used in some of the following chapters. Talja (1997) mentions that information needs arise when someone finds themselves in a problem situation they can no longer manage with the knowledge that they possess, and as such is the catalyst for information seeking behaviour, i.e. using a search system. More practically, an information need is often regarded as a user's end goal in a specific search session, a description of the information or the answer they are looking for. This can be the same or overlap with the actual query a user enters in a search engine, but not necessarily. Some web search examples of information needs might be "how far can a trebuchet launch a 90kg projectile?", or "find a recipe for hummus".

In archaeology, our information needs are often list-based retrieval questions based on What, Where and When. Some examples are "find all excavations in a twenty kilometre radius around Leiden" or "find all documents about Early Medieval cremations". The first type is common in commercial archaeology, where in desk-based assessments the archaeologist is looking for sites nearby a building development area. The second type is more typical of academic archaeology, where research is often focused on specific time periods, artefacts, and/or contexts.

The information need is strongly related to relevance, a fundamental concept in IR. In short, a document is relevant if it contains the information the user is looking for when entering a query. This sounds relatively simple, but there are many factors that influence whether a user finds a document relevant. Simply returning all documents that contain the exact query entered would lead to poor results in terms of relevance (Croft *et al.*, 2010). This is due to vocabulary mismatch: polysemy (a word having multiple meanings) and synonymy (multiple words with the same meaning), which is further described in section 2.3.3.

Related to relevance is ranking, another important concept in IR. Ranking is a method which aims to rank the retrieved results in such a way that the most relevant documents are at the top of the returned list (Croft *et al.*, 2010). While much research is done on this topic, and many methods are available, we do not focus much on ranking as our user requirement study (chapter 5) revealed that users mostly have information needs where completeness is more important than relevance ranking. In other words, as long as the returned documents contain as many relevant results as possible, archaeologists generally are less concerned with the order of the documents, as they will check all of them anyway, if possible within the time available for analysis.

Professional Search

A lot of research on IR is geared towards general online search, where the users are a large group with very diverse searching goals. This research however, focuses on what is known as professional search (Lancaster & Gallup, 1973). As opposed to general web search, research in professional search addresses and supports the search tasks of professionals in a variety of domains (Russell-Rose *et al.*, 2018), in this case the archaeology domain. This type of search has specific requirements, and is often characterised by the use of specialist search systems, with more complex queries and information needs than generic web search (Verberne *et al.*, 2019).

Specifically for archaeologists, we see that search is often focused on the where, what, and when, in much more detail than web or generic document search. We also notice that archaeologists are more concerned with obtaining as many relevant results as possible, even if this means having some irrelevant documents in the results list. This means we are dealing with a high recall task (see section 2.3.4 for a definition of recall). To deal with the spatial and temporal aspects of common archaeological information needs, we need to apply map-based search and more complex time period search, which is discussed in more detail in Chapters 5 and 7 respectively.

2.3.2 Text Mining and Machine Learning

Text mining is an umbrella term describing a range of techniques that allow software to extract useful information from text collections (Truyens & Van Eecke, 2014; Feldman & Sanger, 2007). These techniques are not new, with the first manual Text Mining processes being done in the 1980s (Peterson & Seligman, 1984) and more automated computer aided Text Mining emerging in the 1990s (e.g. Feldman & Dagan, 1995). Recently, Text Mining has received renewed attention due to the emergence of the 'Big Data' and data mining trend, and Text Mining applications have been steadily increasing in number. Typical Text Mining tasks include text categorisation, text clustering, sentiment analysis, translation, document summarisation and NER (Truyens & Van Eecke, 2014). NER is the task we focus on in this study, and is further described in the next section.

Machine learning is often used to perform Text Mining tasks, as opposed to rule-based approaches that were popular originally. Machine learning can be broadly defined as "computational methods using experience to improve performance or to make accurate predictions" (Mohri *et al.*, 2013, p. 1), where experience refers to past information available to the learning algorithm, also called

training data. This data generally consists of examples that have been labelled by human annotators, from which the algorithm can extract meaningful statistical relations. The success of the prediction process depends on the quality and size of the training data, and the complexity of the task (Mohri *et al.*, 2013).

2.3.3 Named Entity Recognition

NER is the process of finding different categories of named entities (or concepts) in text. Quite often, the categories of entities are persons, organisations, locations, time periods and quantities, as defined in CoNLL-2002, the most used NER benchmark (Tjong Kim Sang, 2002). For archaeology, these entities are not as relevant, with the exception of time periods and locations. In this study, we focus on the following entity types:

- Artefacts
- Time Periods
- Materials
- Contexts
- Locations
- Species

Table 3.1 in the next chapter gives more formal definitions of these entities and some examples.

But why is NER relevant for searching in archaeological texts, and why is a standard free text search not sufficient? In one of the previous sections, we already mentioned polysemy and synonymy, which are the main reason why NER can help us find relevant documents.

Polysemy is the phenomenon of one word having multiple meanings. An example is the word "flint". This can mean the material flint, or a person with the surname Flint. In Dutch archaeology, a good example is "*Swifterbant*", which can mean either an excavation event, a type of pottery, a time period, or a place in The Netherlands. A standard free text search would return results about all of these meanings, but if we know which meaning a user is looking for, and we can detect the meaning in the documents, then we can return more relevant results. We can use NER to disambiguate between these meanings in the documents.

Synonymy is the other way around: a concept that can be described by many different words. An example is the location Den Haag, which can also be written as 's Gravenhage and The Hague. While synonymy occurs in all six entity types described above, it is only a major challenge for time periods. There are countless

ways in which we can describe e.g. the Neolithic, or periods and years within the Neolithic. To name a few examples:

- the Late Stone Age
- 7300 - 4000 BP
- 5300 - 2000 BC
- 4th to 3nd millenium B.C.
- 5693 ± 26 BP (a carbon dating date)
- Funnelbeaker culture
- NEO (a code for the Neolithic)
- 3400 BC

But when an archaeologist searches for the Neolithic, ideally they would want all mentions of a date or period within the Neolithic to be returned, and not just the documents that literally contain the word "Neolithic". If we want to be able to do this, we first need to find all mentions of time periods in the reports, which is where we can use NER. Once we have a list of time periods for each document, we can translate these mentions to year ranges using a thesaurus of time periods and a rule-based approach for dates and years. So we can translate "Funnelbeaker culture" to the year range -4350 to -2700, and "4th to 3nd millenium B.C." into the range -4000 to -2000. Users can then search on specific date ranges, or we can translate their query of "Neolithic" to a year range, and find all mentions of time spans that fall within that range. This way we can find more relevant results in the document collection.

Tokens, Terms and the BIO format

Another concept that warrants explaining in the context of NER are tokens. A token is an instance of a sequence of characters that are grouped together as a useful unit for processing (Manning *et al.*, 2008). Tokens are similar to words, and a token often is a word, but not always. We can illustrate this with the following sentence: "We didn't find any 'Swifterbant' pottery in pit 1, 2 and 3.". When this sentence is converted into tokens, in a process called tokenisation, we find the following tokens, here separated by spaces:

```
We didn ' t find any ‘ Swifterbant ’ pottery in pit 1 , 2 and 3 .
```

As we can see, most of these tokens are indeed words, but punctuation marks have also become individual tokens and "didn't" has been converted to three separate tokens. This tokenisation process is important as it removes noise (such

as the quotes around Swifterbant) and turns sentences into chunks that can be processed further. Also, specifically for NER, predictions are done at a token level. This means that for each of these tokens, a prediction is made.

This is also reflected in the way NER training data and predictions are generally stored, in the Beginning, Inside, Outside (BIO) format (Ramshaw & Marcus, 1999). This format is most commonly used for sequence labelling tasks such as NER. The file format is a simple text file, with each token on one line, followed by a space and the label. Sentence boundaries are denoted by a double line break. An example is shown below:

```
We O
found O
a O
pottery B-ART
shard I-ART
from O
the O
Neolithic B-PER
. O
```

Here we see a sentence where 'pottery' has been labelled as the start of an Artefact entity, 'shard' as inside an Artefact entity, and 'Neolithic' labelled as the start of a Time Period entity. The other tokens are labelled O for Outside an entity.

Related to tokens are terms, which are all of the tokens that are included in a search engine's index. Quite often, not all terms are included in an index, for example, very common words such as 'the', 'and', 'of' etc (also called stop words) are removed as they are not useful for searching. Punctuation is also commonly not indexed.

Also worth mentioning here are Part Of Speech (POS) tags. A Part Of Speech is a category of words that have similar grammatical properties, such as noun, verb and adjective. These POS tags can be used as a feature in NER, and as such are often saved together with the BIO tags in a file.

2.3.4 Evaluation Metrics

Evaluation is important for all NLP techniques, to assess to what extent the method is working. As in this project we are mainly dealing with the evaluation of NER, we will discuss the different evaluation metrics relevant to this technique

		Prediction	
		True	**False**
Label	**True**	tp	fn
	False	fp	tn

Table 2.1: Illustrating the true/false positive/negative categories.

and give examples within this context. Most metrics involve calculations of percentages between correctly and incorrectly classified items. In the case of NER, we predict a label for each token. That predicted label is compared to the true label, and we can then put each prediction in one of the following categories:

- True positive (tp). When a token is part of an entity, and the predicted label is the correct entity.
- True negative (tn). When a token is not part of an entity, and the predicted label is also not part of an entity.
- False negative (fn). When a token is part of an entity, but the predicted label is not part of an entity. More simply put: an entity that has not been recognised by the system.
- False positive (fp). When a token is not part of an entity, but the predicted label is an entity. More simply put: the system recognises an entity where there is none.

These categories are further illustrated in table 2.1. Once we have this information, we can calculate some metrics. The most used measures in machine learning in general are recall, precision and F1 score, and these are almost always used to evaluate NER too.

Recall is a measure that indicates out of all the entities in a text, what percentage have been correctly labelled as an entity. It can also be viewed as the percentage of entities that have been found. It is defined as follows:

$$\text{Recall} = \frac{tp}{tp + fn} \tag{2.1}$$

Precision is a measure that indicates, out of all the labelled entities, what percentage has been assigned the correct label. In essence, this means that it shows that when an algorithm predicts an entity, how often it is right. It is defined as follows:

$$\text{Precision} = \frac{tp}{tp + fp} \tag{2.2}$$

The F1 score (or F measure) combines recall and precision to provide an overall evaluation metric. More specifically, it is the harmonic mean of precision and recall, and is defined as:

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{2.3}$$

For NER, these measures are calculated for each possible label separately. To evaluate a NER algorithm as a whole, the micro average is calculated for all labels combined, with the exception of the O label. This is because the O label is much more prevalent in the data (most tokens are not an entity) and is easy to predict, so including it would unfairly increase the recall, precision and F1 score.

2.4 Previous Research

Although there is limited prior work addressing NLP and IR in the archaeology domain, there are some examples of related research in the literature. Almost all of those studies have focused on grey literature as the source material, presumably because it has the greatest potential for computational techniques.

One of the earliest applications of IR in archaeology was done by Copeland (1983), who did a study on information needs of users of a sites and monuments record. As this was back in 1983, the information was stored on physical 5 by 8 inch record cards, ordered by grid coordinates. Even though the situation was very different to our current situation, the problem is the same: the metadata (grid coordinates) were not good enough for information retrieval, as users want a way to cross reference or search through the data (text on cards). Copeland sent out surveys by post asking archaeology professionals on their opinion on the use of computers for record manipulation, and found that 63% already did, or were hoping to do so in the future, meaning 37% of respondents did not see any value in using computers for this task. Eventually they concluded that "A computer-based recording system gives the potential to relieve problems of lack of space, lost data, inaccuracies in recording and to provide a flexible and efficient retrieval system, therefore relieving staff time for other work" (Copeland, 1983, p. 43), which is basically also the main aim of this project. It seems not much

has changed in the last 40 years in that respect.

At the end of the 20th century, computer systems became increasingly common place, and in the last 20 years a number of projects have used Text Mining techniques on archaeological texts. Amrani *et al.* (2008) created a full workflow allowing experts to extract information from text, but in a quite specialised way on small collections, and is not meant for searching through large corpora. Byrne & Klein (2010) experimented with extracting archaeological events and converting them to Resource Description Framework (RDF) triples, to increase the interconnectivity between data silos.

Going more in the direction of IR, the Archaeotools project used a combination of rules based and machine learning approaches to automatically generate location, time period, and subject metadata for a small selection of a thousand reports, with moderate success. This generated metadata could then be used for searching in a facetted interface (Jeffrey *et al.*, 2009). In the OPTIMA project, Vlachidis (2012) applied purely rules based techniques to perform NER and semantically annotate grey literature reports, by expressing entities in the CIDOC-CRM schema[1]. The output of this research was further built upon in the STAR and STELLAR projects, where Tudhope *et al.* (2011) created a search demonstrator which cross-searches through five excavation databases and a small selection of archaeological reports, two types of data that are normally queried separately.

In a more classical IR setting, Gibbs & Colley (2012) describes a search engine in Australia, allowing for full-text search and facetted browsing of around a thousand grey literature reports. However they do not attempt any NER or information extraction, and the facets are based on manually added metadata.

The Advanced Research Infrastructure for Archaeological Dataset Networking in Europe (ARIADNE) project (Niccolucci & Richards, 2013) aimed to bring together and integrate archaeological research data infrastructures, so that archaeologists can use these varied and fragmented data sets in their research. As part of this project, some experiments were undertaken with NLP on grey literature. The Archaeology Data Service (ADS) in the UK created a prototype web application and Application Programming Interface (API) that performs NER using the CRF algorithm, to automatically create metadata for English reports (Vlachidis *et al.*, 2017).

In her Master's thesis, Talboom (2017) specifically targeted zooarchaeological entities in reports, and used a Bidirectional Long Short Term Memory (Bi-LSTM)

[1] A Conceptual Reference Model (a way to model information) for cultural heritage and museum documentation, as defined by the International Committee for Documentation (CIDOC) (2014)

algorithm to perform NER. This showed promising results, but unfortunately the technique has not been evaluated fully yet. Building on her work, Talks (2019) added more entity types and did an extensive evaluation with users.

All the research described above has been on the English language, and research on Dutch and other languages is much less prevalent. For Dutch, there are two main examples: the OpenBoek project and the experiments on Dutch texts in the above mentioned ARIADNE project.

The OpenBoek project (Paijmans & Wubben, 2008; Paijmans & Brandsen, 2010) aimed to create a full text search engine combined with entity search, on about 2,000 reports. They used Memory Based Learning to automatically label time periods and locations, which were searchable together with the full text in a web application based on the SMART system (Salton, 1971). While the search engine showed promising results, unfortunately this web application has gone offline not too long after the funding for the project ended.

The ARIADNE project – besides the work on English texts described above – also experimented with Dutch and Swedish grey literature. For Dutch, they applied a rules based technique using the General Architecture for Text Engineering (GATE) framework (Cunningham *et al.*, 1995). The rules were mainly based on thesauri, but they found many issues with the thesauri and gold standard, making effective NER with this approach difficult.

Very recently, Fischer *et al.* (2021) experimented with Text Mining and IR as part of their research on urban farming and ruralisation in the Netherlands. They extracted text from a number of PDFs, created a term document matrix and compared this with a list of keywords related to the topic of urban farming, to automatically assess the relevance of a large number of documents for a number of topics.

In a slightly different direction, recent work by Plets *et al.* describes research on Dutch archaeological texts from Belgium, looking at theoretical trends over time. They successfully manage to use Text Mining to find these trends, and chart the decrease in text quality due to developer-led archaeology. Similarly, Jackson *et al.* (2020) used topic modelling techniques on large-scale English data to see if there are patterned ways in which archaeologists write about bone.

Almost no research has been done on multilingual techniques, but Mélanie-becquet *et al.* (2015) present some interesting results for NER on English, German and French documents, although technical details have not been published yet. Another notable study on IR is the work by Eramian *et al.* (2017), who built an image-based retrieval system for biface artifacts.

Overall, we see that most previous research is experimental and exploratory, with many prototypes being developed, but no useable search systems are avail-

able long-term for archaeologists to actually use in their research. This project aims to do exactly that for Dutch grey literature, with longer term support in the form of a follow up project, described in more detail in section 9.9. During this next project, we aim to find a national organisation to host the system for the foreseeable future.

2.5 Resources

Various resources were used in this research, which we describe below.

2.5.1 The DANS Corpus

The corpus we use for this research is a complete download of all PDF files with the 'archaeology' label from the DANS archive, taken in 2017. DANS is an online archive of research data, based in The Hague. They store data from a variety of domains, including archaeology. The majority of the commercially created data sets and reports are deposited in this archive, and as such it is a good document collection for this research. Some academic output is also stored here, but this is a small proportion of the archaeological data.

The total number of files we have at our disposal is 65,083. This includes not just reports, but also appendices, research plans (*Plan van Aanpak*), maps, and some reports are split into multiple PDF files. The total number of unique DANS data set IDs in our collection is 24,029, meaning there are documents about roughly 24k different research projects.

These documents in their PDF form total around 1.5TB of data, but when only the text is extracted, this drops to about 2GB. To give an idea of the amount of text, the full collection contains 658 million tokens across 16.6 million sentences.

2.5.2 Computing Power

Due to recent advancements in computing power, as well as the increased availability and decreased cost, it is now feasible to run more complex code in a relatively short time. This opens up possibilities for the use of advanced machine learning and/or Deep Learning methods which were previously outside the reach of ordinary researchers with no access to a high performance computer cluster. In this project, these recent developments have been used to create a cutting-edge search engine, which should provide better results than previous projects, which

sometimes struggled with the required computing power needed for an ideal solution, often leading to systems where simpler solutions were used simply because the computing power was not available.

To harness this computing power, this project is in association with the Data Science Research Programme (DSRP). We have used both the Leiden Institute of Advanced Computer Science (LIACS) Data Science Lab (DSlab) and the Academic Leiden Interdisciplinary Cluster Environment (ALICE) cluster provided by Leiden University. Initial experiments (Chapter 3, 4, 5) have been run on the DSlab, on a machine with 32 2.4GHz CPU cores and 1.5TB of RAM, and no GPU. Most methods used only a fraction of these resources, and could potentially be run on a desktop PC, although with longer processing times.

The experiments with Deep Learning models (Chapter 7) have been run on the ALICE cluster, generally on a GPU node with 24 2.6GHz CPU cores, 384GB of RAM, and 4 GeForce RTX 2080TI GPUs. These models require significantly more processing power and would utilise all the available resources on the node.

2.5.3 Ontologies

To clear up any possible confusion, when 'ontology' is mentioned in this dissertation, this does not refer to the branch of philosophy, but the information science concept: a representation of concepts in a specific domain (Gruber, 1995). This is similar to a thesaurus or word list, with the most well known Dutch example being the *Archeologisch Basisregister* (ABR) ontology (Brandt *et al.*, 1992).

For NER, it is useful to have ontologies for the categories of entities you are targeting, as whether or not a token occurs in such a word list is an indication that it might be an entity. For Artefacts and Time Periods, we use the aforementioned ABR ontology. This is a hierarchical list of artefacts, time periods and monument types, created and maintained by the RCE. We have slightly adjusted some of the entries to better match natural language, e.g. changing "*bijl, doorboord*" (axe, perforated) to "*doorboorde bijl*" (perforated axe).

Unfortunately, the ABR is not very exhaustive and only contains a basic list of time periods. This is why we decided to use the PeriodO time appellations list (Rabinowitz *et al.*, 2016) for translating Time Periods to year ranges (further described in chapter 7). We also altered this list by adding more time periods, mainly geological time spans (e.g., Holocene) and specific cultures (e.g., Bell Beaker Culture).

For Locations and Species, we are not using any ontologies, as we are focusing more on Artefacts and Time Periods for the time being. For future work on these

entities, we have found suitable ontologies: GeoNames[2] and the Catalogue of Life[3].

2.5.4 Gold Standard

To train NER algorithms, and assess the accuracy of the models, a manually tagged collection of documents is needed. This is called a gold standard, and at the start of the project, the data set created in the ARIADNE project was used (Vlachidis *et al.*, 2017). This data set consists of eight documents, 355k tokens, 20k entities across nine categories. This set has been annotated by hand by highlighting spans in the Microsoft Word word processor.

These highlighted entities have been extracted from the eXtensible Markup Language (XML) of the Word file, and converted to the BIO file format. However, when we started experiments with this data set, we found some inconsistencies and issues in the annotations that might be causing low F1 scores on the NER task. These problems with the data set have also been described by Vlachidis *et al.* (2017). To try and improve our system, we created a new data set, optimally annotated for NER, which we further describe in the next chapter.

[2] www.geonames.org
[3] www.catalogueoflife.org

3

Data Set

"*Analysis complete. Insufficient data to resolve problem.*"
Nomad, Star Trek TOS, s02e03 'The Changeling'

Previously published as: Brandsen, A., Lambers, K., Verberne, S. and Wansleeben, M., 2020. Creating a Data Set for Named Entity Recognition in the Archaeology Domain. In: *Proceedings of The 12th Language Resources and Evaluation Conference*, pp.4573-4577.

In this paper, we present the development of a training data set for Dutch Named Entity Recognition (NER) in the archaeology domain. This data set was created as there is a dire need for semantic search within archaeology, in order to allow archaeologists to find structured information in collections of Dutch excavation reports, currently totalling around 60,000 (658 million words) and growing rapidly. To guide this search task, NER is needed. We created rigorous annotation guidelines in an iterative process, then instructed five archaeology students to annotate a number of documents. The resulting data set contains roughly 31k annotations between six entity types (artefact, time period, place, context, species & material). The Inter Annotator Agreement (IAA) is 0.95, and when we used this data for machine learning, we observed an increase in F1 score from 0.51 to 0.70 in comparison to a machine learning model trained on a data set created in prior work. This indicates that the data is of high quality, and can confidently be used to train NER classifiers.

3.1 Introduction

The archaeology domain, like other scientific fields, produces large amounts of textual data. Specifically, a large amount of excavation reports are available, which are created whenever an excavation is completed, detailing everything that has been found together with an interpretation of the site (Richards *et al.*, 2015). In the Netherlands, this corpus is estimated at 60,000 documents, and is growing by 4000 each year (Rijksdienst voor het Cultureel Erfgoed, 2019a). Most of these reports are created and published by individual commercial archaeology companies after they excavate, in low numbers and not widely shared.

This so-called grey literature is currently underused, even though most scholars agree that the information hidden in these reports is of immense value (Evans, 2015). The systems currently available to explore this corpus are metadata search engines that simply do not offer enough granularity for archaeologists to easily find what they are looking for. An example might be a single find from the Bronze Age which was not included in the temporal metadata as it is too specific. Currently, there is no way of finding this so called 'by-catch'; single finds of a different type than the rest of the excavation. Users of the currently available search engines report they download whole portions of the available data and manually search through PDF files one by one to find the information they are looking for (Brandsen *et al.*, 2019).

Free text search across the entire corpus would already be a vast improvement, however this does not account for polysemy and synonymy, which occur often in

archaeological texts. An example of polysemy could be the time period of the Neolithic, which can also be expressed as the Late Stone Age, 11,000 - 2000 BC, 13,000 BP, etc. And the other way around, there are terms like 'Swifterbant' that can mean a time period, an excavation, a specific type of pottery or a town in the Netherlands. To alleviate this problem, we have applied Named Entity Recognition (NER) to the data set, to automatically extract and distinguish between these entity types. We are building an online search system that allows archaeologists to search through these entities, as well as full text search, using an intuitive interface. The system is called AGNES (**A**rchaeological **G**rey-literature **N**amed **E**ntity **S**earch)[1]. The overall goals of the project, a description of the first version of AGNES , and a user requirement solicitation study can be found in a previous publication (Brandsen *et al.*, 2019).

As we are using machine learning for the Named Entity Recognition, a labelled data set is needed as training data. A Dutch data set created in the ARIADNE project (Vlachidis *et al.*, 2017) was used initially in this project, but after some experiments we found that the data was of insufficient quality, with some entities being annotated incorrectly and some having inconsistent and inaccurate span lengths. For example, often (but not always) a quantifier was included in the span for time periods, e.g. "roughly around 200BC", where the correct entity would be just "200BC". When using this data set as training data for a sequence labelling classifier with Conditional Random Fields (CRF) (Lafferty *et al.*, 2001), we only managed to reach an F1 score of 51% (Brandsen *et al.*, 2019). To see if we could alleviate these problems, we created a new training data set.

The research questions for this paper are:

- How high is the Inter Annotator Agreement, and by proxy, the reliability of the newly created data set?
- To what extent will creating a more rigorous data set yield higher accuracy in Named Entity Recognition?

The training data set is available for download[2] (Brandsen, 2019).

3.2 Related Work

The go-to benchmark for Dutch Named Entity Recognition is the CONLL-2002 shared task, for language independent NER , which includes a Dutch data set.

[1] Which can be found at http://agnessearch.nl
[2] doi.org/10.5281/zenodo.3544544

But this task only looks at common, general-domain entities and is not comparable to our data set (Tjong Kim Sang, 2002).

In the archaeology domain, NER data sets exist in other languages (English and Swedish), created in the ARIADNE project (Vlachidis *et al.*, 2017). To our knowledge, the only directly related data set that deals with both Dutch and archaeological texts is another data set created in the same ARIADNE project, as briefly described in the introduction. As we are going to show in this paper, the data set we have created is of better quality and much larger than the ARIADNE data.

3.3 Data set Collection

From the total available corpus (70k documents), we currently have access to ~60,000 excavation reports and related documents, such as appendices, drawings and maps. These texts have been gathered by DANS (Digital Archiving and Networked Services) in the Netherlands, over the past 20 years. We received the documents from DANS as PDF files, and have used the pdftotext tool (Glyph & Cog LLC, 1996) to convert these to plain text. This data set contains 30,152,318 lines and 657,808,600 words (as counted by the command line tool "wc").

The texts are quite diverse; the dates of publication span decades with the earlier ones having been scanned and OCRd from hardcopies created in the 80s. The other temporal variation is in how old the found artefacts are, ranging from 200,000 BC to the present. Also, the type of research can be very different between reports, some might describe a short desk evaluation of a small area without any fieldwork, while others detail huge excavations over multiple years with detailed analysis by a team of specialists. To get a representative sample across all these ranges, a random sampling strategy would not be ideal, and we instead opted to manually select documents, taking into account the variation described above. We selected a total of 15 documents as annotation candidates (~42,000 tokens).

For the purposes of calculating the IAA and evaluating the annotation guidelines, we manually selected roughly 100 sentences from these documents containing all the entity types (Table 3.1, explained below) and specific difficult cases as validation set, annotated by all annotators.

Entity	Description	Examples
Artefact	An archaeological object found in the ground.	Axe, pot, stake, arrow head, coin
Time Period	A defined (archaeological) period in time.	Middle Ages, Neolithic, 500 BC, 4000 BP
Location	A placename or (part of) an address.	Amsterdam, Steenstraat 1, Lutjebroek
Context	An anthropogenic, definable part of a stratigraphy. Something that can contain Artefacts	Rubbish pit, burial mound, stake hole
Material	The material an Artefact is made of.	Bronze, wood, flint, glass
Species	A species' name (in Latin or Dutch)	Cow, *Corvus Corax*, oak

Table 3.1: Descriptions and examples for each entity type. Examples are translated from Dutch.

3.4 Annotation Setup

As an annotation tool, we used Doccano (Nakayama, 2019), an open source and intuitive system. After comparing the system to other available entity tagging tools, we found this was the easiest to use and most efficient tool for our purposes. The system was set up on a web server, data was uploaded for each user and entity types defined within the system.

3.4.1 Annotation Guidelines

The annotation guidelines were created in an iterative process. A first draft was created, containing general guidelines as well as specific examples of difficult situations. Two archaeologists used the guidelines to annotate around 100 sentences, and these annotations were compared to our own desired annotations to see where problems and inconsistencies were encountered. This information was then used to update the guidelines, after which they were tested again. This led to an IAA (F1 score, further explained in section 3.5.1) of 0.94 between the two testers, which we consider sufficient for this task.

During the annotation process itself, whenever one of the annotators ran into a situation that was unclear, this was added as an example to the guidelines.

The annotation guidelines (in Dutch) can be downloaded as part of the data set (Brandsen, 2019).

3.4.2 Entity Types

Table 3.1 lists the targeted entities and provides a brief explanation of each type with some examples. With the exception of location, these are all uncommon entity types, not occurring in general-domain Named Entity Recognition tasks. The entity types have been chosen based on a user requirement study, where archaeologists indicated which entities they would like to search on.

3.4.3 Annotation Process

To carry out the annotation work, we recruited five Dutch archaeology students at the Bachelor level. We specifically selected students in their second and third year, as some basic knowledge of archaeology is extremely helpful in determining whether a word is a specific entity or not.

The students were asked to annotate a total of 16 hours each, over a two week period, during which they could come and work at times that suited them, a few hours at a time. We opted not to have the students work a whole day on this task, as the annotation process is tedious and monotonous, which makes it hard to keep concentration. Loss in concentration can cause mislabelling, and so having them work for only small amounts of time might help prevent this.

The students were first asked to thoughtfully read the guidelines and ask any questions. During annotation, we were always present to resolve difficult sentences and entities and explain to the students how to handle these. The students reported this to be very helpful, and learned from each other's problems. Most of these issues were relatively rare edge case though, and the original annotation guidelines covered most encountered entities sufficiently.

3.5 Annotated Corpus Statistics and Results

Table 3.2 lists general statistics on the annotated corpus, including number of documents, sentences, tokens, annotations and averages over these categories.

Over a total of 90 hours, the students annotated ~31,000 entities, setting the average annotation rate at 346 per hour, or 5.7 per minute, which is higher than we expected. The previous data set we used contained only around 11,000 annotations, so we almost tripled the amount of available training data. While this seems like a large amount of entities, the amount of tokens seen by annotators is but a fraction (0.066%) of the total number of words in the data set. The breakdown per entity type is shown in Table 3.3.

Documents	15
Sentences	33,505
Avg. sentences per document	2,234
Tokens	439,375
Avg. tokens per sentence	13.1
Annotation spans	31,151
Annotated tokens	42,948
Avg. tokens per annotation	1.38

Table 3.2: Annotated corpus statistics.

Entity Type	**Quantity**
Artefact (ART)	8,987
Time Period (PER)	8,358
Location (LOC)	4,436
Context (CON)	5,302
Material (MAT)	1,225
Species (SPE)	2,843
TOTAL	31,151

Table 3.3: Number of annotations per entity type in the data set

3.5.1 Inter Annotator Agreement

For most tasks, Cohen's Kappa is reported as a measure of IAA, and is considered the standard measure (McHugh, 2012). But for Named Entity Recognition, Kappa is not the most relevant measure, as noted in multiple studies (Hripcsak & Rothschild, 2005; Grouin *et al.*, 2011). This is because Kappa needs the number of negative cases, which isn't known for named entities. There is no known number of items to consider when annotating entities, as they are a sequence of tokens. A solution is to calculate the Kappa on the token level, but this has two associated problems. Firstly, annotators do not annotate words individually, but look at sequences of one or more tokens, so this method does not reflect the annotation task very well. Secondly, the data is extremely unbalanced, with the un-annotated tokens (labelled "O") vastly outnumbering the actual entities, unfairly increasing the Kappa score. A solution is to only calculate the Kappa for tokens where at least one annotator has made an annotation, but this tends to underestimate the IAA. Because of these issues, the pairwise F1 score calculated without the O label is usually seen as a better measure for IAA in Named Entity

Cohen's Kappa on all tokens	0.82
Cohen's Kappa on annotated tokens only	0.67
F1 score	0.95

Table 3.4: Inter-annotator agreement measures on 100 sentence test document. Calculated by doing pairwise comparisons between all combinations of annotators and averaging the results.

Recognition (Deleger *et al.*, 2012). However, as the token level Kappa scores can also provide some insight, we provide all three measures but focus on the F1 score. The scores are provided in Table 3.4. These scores are calculated by averaging the results of pairwise comparisons across all annotators. We also calculated these scores by comparing all the annotators against the annotations we did ourselves, and obtained the same F1 score and slightly lower Kappa (-0.02).

3.5.2 New NER Results

We have used these entities as new training data, using the same CRF model as mentioned in the introduction (Brandsen, 2018), and have seen a large increase in the overall micro F1 score, from 0.51 to 0.70, showing that this data is of better quality than the previously used training data. The difference between this, and the F1 between five human annotators (0.95) indicates that there is also still room for improvement.

In Table 3.5 we show the difference in F1 score per entity type. Most types see a substantial increase, especially Locations, while the Material category sees

	Old	**New**	**Difference**
Artefact	0.51	0.63	+0.12
Time Period	0.57	0.69	+0.12
Location	0.26	0.66	+0.40
Context	0.58	0.84	+0.26
Material	0.54	0.39	-0.15
Species	n/a	0.49	n/a
Overall Micro F1	0.51	0.70	+0.19

Table 3.5: F1 scores for entity types and overall micro F1 compared between the previous and new data set. Species wasn't included in old data set, so we only present the score for the new data set.

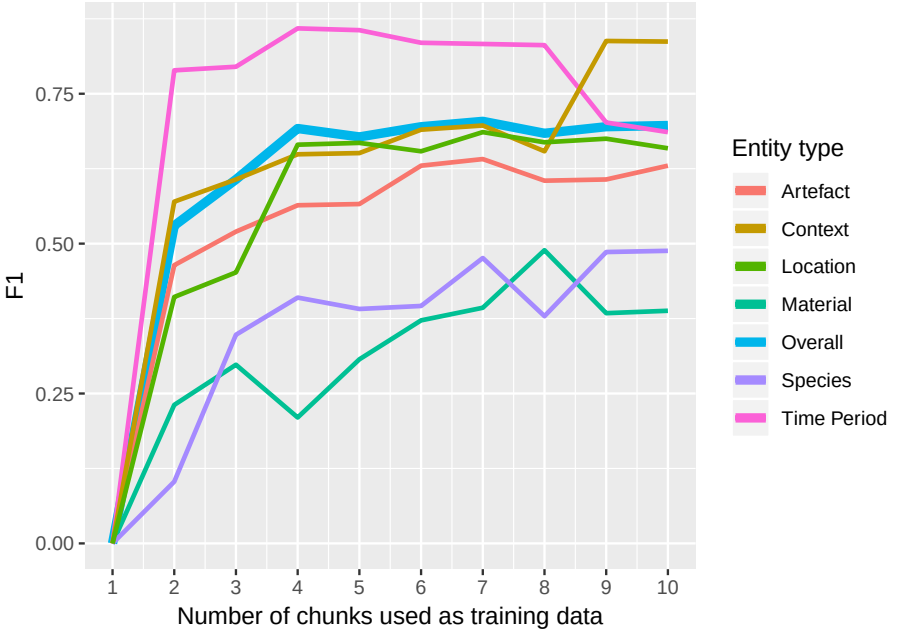


Figure 3.1: CRF F1 score for each entity type per 1/10th chunk of data added to the training set.

a decrease in F1 score. We wondered if this could be explained by the fact that we have much fewer annotations for the Material category, only 1,078 while all other categories have at least double that amount.

To assess this, we divided the data set into 10 chunks, and retrained the CRF model 10 times, every time adding one more chunk of data. In Figure 3.1 we have plotted the F1 score for individual entity types and the overall micro F1 score for each model. Even though there are some fluctuations, it is evident that after adding a certain percentage of the data, the F1 scores for all the entity types plateau, even for the Material type. This probably indicates that the amount of annotations is sufficient and adding more data won't substantially increase the F1 scores, although redundancy and noise in the data set could also potentially cause similar results. We will investigate this further in future research.

The Species category performs similarly as Material, at 0.49, this could possibly be explained by the fact that Species are written in both Dutch and Latin, but more work needs to be done to see if this is indeed the case. We also performed this analysis but instead of adding 10% of the data each time, we added

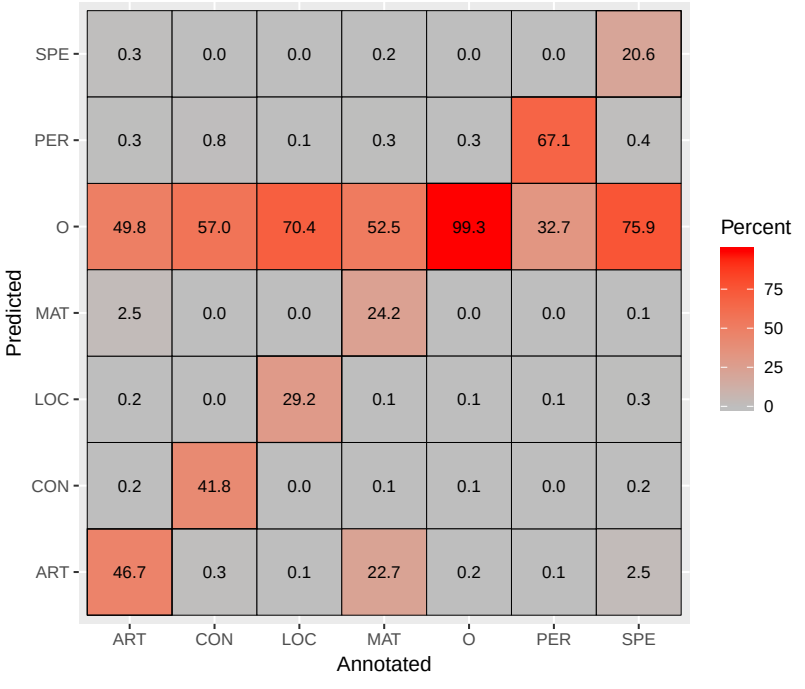


Figure 3.2: Confusion matrix showing percentages for each combination of predicted and annotated entity type.

a new document each time, which showed the same trend.

To see if there is another explanation for the under performance of the Material entity, we plotted a confusion matrix for all the different types, as seen in Figure 3.2. The diagonal and horizontal red lines are expected: the cells on the diagonal is when the algorithm predicts the correct entity, the horizontal red line is when the algorithm mistakes an entity for the O entity, the most common error in Named Entity Recognition. The only significant exception is the cell at the centre-bottom: this shows that in 22.7% of the cases, what has been annotated by humans as a Material, has been predicted by the algorithm to be an Artefact. There is also some confusion the other way around, but at a much lower rate of only 2.5%. Interestingly, from our experience supervising the annotators, this is something humans struggle with as well. The confusion is caused mainly by the words “pottery” and “flint”, which depending on the context can be either a Material (“a flint axe”) or an Artefact (“we found flint”).

3.6 Conclusions

In this paper, we have presented a new corpus for Dutch Named Entity Recognition in the archaeology domain, annotated with six entity types. Many of the entity types are not available in standard corpora.

We trained a CRF model on the data set, as a first experiment to assess the quality of NER with this data. The results with CRF show that using the new data substantially increases accuracy for the NER task compared to an earlier data set. However, we only reach an F1 score of 0.70, while the IAA is 0.95. More research needs to be done to why this is the case and how we can increase the accuracy of the NER model(s).

In our current work we are using the recent advances in transfer learning to our advantage, and apply the BERT (Bidirectional Encoder Representations from Transformers) models to this task (Devlin *et al.*, 2019). We will be using both Google's own multi-lingual model, and a model pretrained on a large Dutch corpus, to see which is more effective.

3.7 Acknowledgements

We would like to thank the Leiden University Centre for Digital Humanities for providing us with a grant to hire students to do the annotation work described in this paper.

4

Text Classification

"*I simply maintain that computers are more efficient than human beings, not better.*"
Mr. Spock, Star Trek TOS, s02e24 'The Ultimate Computer'

Previously published as: Brandsen, A. and Koole, M, 2021. Labelling the past: data set creation and multi-label classification of Dutch archaeological excavation reports. *Language Resources and Evaluation.* DOI: 10.1007/s10579-021-09552-6.

The extraction of information from Dutch archaeological grey literature has recently been investigated by the AGNES project. AGNES aims to disclose relevant information by means of a web search engine, to enable researchers to search through excavation reports. In this paper, we focus on the multi-labelling of archaeological excavation reports with time periods and site types, and provide a manually labelled reference set to this end. We propose a series of approaches, pre-processing methods, and various modifications of the training set to address the often low quality of both texts and labels. We find that despite those issues, our proposed methods lead to promising results.

4.1 Introduction

Over the past decades, the archaeological domain has produced a large quantity of literature in the form of excavation reports, scholarly articles, and books. The Archaeological Grey literature Named Entity Search (AGNES) project (Brandsen *et al.*, 2019) aims to uncover any relevant information from Dutch archaeological excavation reports. Such reports are often grey literature: material that is either unpublished, or published in a non-traditional manner. Information uncovered by AGNES will be made easily accessible through a specifically designed search engine, enabling researchers to search for relevant texts.

In this search engine, certain aspects of documents are used for faceted search, allowing archaeologists to filter search results on site type and time period metadata fields. This information need is further detailed by Brandsen *et al.* (2019). AGNES currently only indexes documents with manually assigned metadata, but in the near future, documents without metadata will be added. To allow for faceted search on these documents as well, we propose to automatically assign metadata. Manual labelling is an unfeasible task due to the amount of texts: there are currently an estimated 70,000 documents and four to five thousand are added each year. Due to this volume, using Text Mining and machine learning techniques becomes a necessity.

In this paper, the labelling of Dutch archaeological excavation reports with time periods and site types[1] will be addressed in the form of a multi-label classification task.

We first create a manually labelled reference set, and then define a collection of pre-processing steps, classification methods, further text formatting and sampling

[1] *Complextype* in Dutch. This can be regarded as a 'subject' field, a site type is what type of past human behaviour has been encountered. Some examples include settlements, churches, graves, etc.

	Error	Correct
1	IJsertijdbewoning	IJzertijdbewoning
2	H et huidige landschapsbeeld	Het huidige landschapsbeeld
3	Time Periods: Gelderland, Ede, Nieuwste Tijd	Time Periods: Nieuwe Tijd

Table 4.1: Examples of noise introduced by (1) OCR mistakes, (2) PDF to text conversion and (3) manual metadata entry in free text fields (locations in time period field). Errors are underlined.

techniques that lead to a multitude of different combinations. We determine which approaches are suitable for this particular type of data, and we discuss how these methods could be further improved.

Although reports are typically freely available in online repositories and archives, processing the documents proves to be rather difficult for four main reasons:

1. Some of the documents are scanned hard copies, and the OCR process introduces noise
2. The documents are only available in PDF format, and conversion to plain text introduces noise
3. The training data labels are derived from the metadata which has been added through a free text field, leading to highly diverse and inaccurate labels
4. There are a large number of target labels (146 site types, 42 time periods) with a strong class imbalance

See table 4.1 for examples of point 1 to 3, and see Figs. 4.2 and 4.3 for point 4.

Besides being useful for faceted search, this machine learning approach can also be helpful for document depositors when they assign metadata to new documents, by suggesting a number of possible labels for the user to choose from. If implemented, this will also lead to more structured metadata in the future, as it prevents free text input on these fields. With these goals in mind, we address the following research questions:

- Which combination(s) of text pre-processing steps, data augmentation/balancing, document pre-selection, and classification method yields the highest F1 scores?
- Are the best combinations the same across the different categories and labels, or do specialised combinations per category lead to better results?

- To what extent can we classify Dutch excavation reports into time periods and site types?

While multi-label classification is a well-studied subject, in this paper we perform this task on a noisy data set in an expert domain, making the process more challenging. Even though the difficulty of the task is high, we achieve decent results: we achieve comparable or better scores when compared to similar studies in other domains (Golub *et al.*, 2020; Kleppe *et al.*, 2019). We also specifically test which pre-processing methods have a positive effect on classification, and provide the created data in an online repository[2].

4.2 Related work

4.2.1 Text mining in the Archaeological Domain

Vlachidis & Tudhope (2012) address the semantic annotation of English archaeological documents, a process similar to our classification task in a multitude of ways. Despite a difference in language, highly similar issues are found in the data set for example. These include the extraction of relevant document sections, scarcity of vocabulary resources, and the construction of a reference set in order to assess the results. Vlachidis & Tudhope (2012) also address the issues of this type of (grey) literature in general. Often, specific archaeological items or names will be mentioned within texts, but hold barely any relevance to the overall topic. Similarly, a variety of terms, such as 'context', 'deposit' and 'cut' yield specific archaeological definitions, but would normally often be seen as common, and therefore not meaningful.

Like our own study, the Archeotools project (Jeffrey *et al.*, 2009) also aimed to automatically generate metadata for faceted search. They focused on 'What', 'Where' and 'When' facets. However, they considered this to be an information extraction task instead of a classification task. As such, they have a slightly different approach based on Named Entity Recognition (NER). The extracted entities are then matched to entries in a English archaeology thesaurus to provide structured metadata. The OPTIMA system by Vlachidis and Tudhope (2016) also focuses on information extraction, but using hand-crafted rules instead of machine learning.

In Dutch, no document classification seems to have been done, but some researchers have experimented with NER, like Paijmans and Brandsen's research

[2]http://doi.org/10.5281/zenodo.3676703

on detecting time periods (Paijmans & Brandsen, 2010), Vlachidis et al. with their work in the ARIADNE project (Vlachidis *et al.*, 2017) and the more recent work by Brandsen et al. (Brandsen *et al.*, 2019, 2020). In the broader context of cultural heritage (also including museums, monuments, etc), Sporleder (2010) gives an overview of the use of Natural Language Processing (NLP) in this domain, but again there is a focus on information extraction, not whole document classification. In an even broader context, Fiorucci et al. provide a summary of – and a critical reflection on – the use of machine learning in the cultural heritage sector, but do not address NLP in any detail (Fiorucci *et al.*, 2020).

4.2.2 Multi-label Text Classification

As already mentioned in the introduction, the classification of Dutch archaeological reports is a multi-label classification problem with many categories and a large class imbalance, as illustrated by Figs. 4.2 and 4.3. These characteristics are not unique to the archaeology domain, and are also often encountered in e.g. the biomedical domain (Laza *et al.*, 2011) and library domain (Golub *et al.*, 2020).

A multi-label classification problem refers to a set of items which can be assigned zero or more labels, according to defined categories. As opposed to binary classification, where an item can have one of two labels, i.e., true or false. Multi-class classification shares the multitude of categories, but here, each item receives one label, rather than zero or more.

Cherman *et al.* (2011) present a case study for multi-label classification with many categories. They propose to transform the n-label problem to n binary relevance problems. One major advantage is that the computational complexity is drastically lowered compared to other multi-label strategies. A disadvantage however, is that relationships between labels cannot be taken into account. In our case, this is not likely to be a problem: though consecutive time periods are naturally more likely to occur together, there are no direct relationships between these periods in terms of archaeological finds. As a matter of fact, time periods are generally defined based on finds, or the material culture of people in the past (Renfrew & Bahn, 2019). Because of this principle, we decided not to introduce a smaller penalty for consecutive periods compared to periods that have a (large) time span between them, i.e., ordinal evaluation. Thus, similarly to the site types, we consider the evaluation of the time periods to be discrete.

To evaluate our methods, we use the F1 score, which is the weighted average – or harmonic mean – of the *precision* and *recall*. Precision is defined as the fraction of positive items that are predicted correctly, and recall is the fraction of

positive items retrieved with respect to all positive items within the set (Powers, 2011). As the harmonic mean over these values, the F1-score is defined as follows:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{4.1}$$

Due to the nature of the task, there is no preference for either recall or precision, and as such we do not use the more recall oriented F2 score, or the more precision oriented F0.5 score (Sasaki, 2007).

With regard to the class imbalance, Joachims (1998) showed the robustness of Support Vector Machines (Support Vector Machine (SVM)), as they provide built-in protection for unbalanced data sets. Another promising approach is the integration of Doc2Vec, a neural network that converts texts into vector representations. In combination with an SVM, Doc2Vec yields high results in terms of *F1 scores* on the task of multi-labelling, for example in ground lease documents (De Romas, 2019).

Finally, a recent state-of-the-art classification technique is the Bidirectional Encoder Representations from Transformers (BERT) architecture (Devlin *et al.*, 2019). This method distinguishes itself from traditional sparse word vectors by learning pre-trained dense language representations from unlabelled data, creating context sensitive embeddings. As such, BERT yields a better contextual understanding of languages, and can lead to improved performance on a lot of NLP tasks.

4.3 Data

In this section, we discuss and analyse the raw data. First, a general description of the data set will be provided based on document titles, content observations and relevant statistical properties. Next, we present the method that has been used in order to extract labels from available metadata, to construct the training and test sets. We then create an overview of the categories extracted from the data and the corresponding labels based on the *Archaeologisch Basis Register* (ABR) notation, further detailed in Sect. 4.3.2. Finally, observations are made regarding the difficulties that the data set might introduce in later stages of the overall research process.

4.3.1 Source Data

We use all documents in the 'archaeology' category in the 2016 version of the Data Archiving and Networked Services (DANS) repository, one of the largest Dutch e-depots. This data set consists of just over 65,000 files, all of which are in PDF format. Examples of included files – based on document titles – are (excavation) reports, publications, separate appendices and figures, letters, and metadata. Although we have not statistically tested the representativeness of this data set, it represents almost all the output of commercial archaeology units from the last 30 years or so, spanning all time periods, site types and different types of reports.

Quite often reports have been split into multiple PDFs, one file for each chapter and appendix is quite common for longer reports. For our research, AGNES already provides a collection in which all files have been converted to both XML and raw text format, which allows for the use of information retrieval and text classification. In this research, we only use the raw text files, which have been created using the pdftotext software (Glyph & Cog LLC, 1996).

We see that the conversion of the PDF files to the required text format introduced a lot of noise. This includes headers, page numbering and various indices appearing at random positions in the text. The main culprits are tables and figures, which are no longer recognisable after conversion. Brandsen *et al.* (2019) estimate that around 15% of all documents are OCRed, a process likely to introduce noise even before the PDF to text conversion. Luckily, this percentage will only decrease, as more and more born digital documents are added over time.

4.3.2 ABR Ontology

The ABR is a Dutch archaeological ontology describing time periods, artefacts, materials and site types, and their corresponding shorthand codes, created and maintained by the RCE (*Rijksdienst voor Cultureel Erfgoed*, the Dutch heritage agency) (Brandt *et al.*, 1992)[3]. The main aim of this ontology is to provide an exhaustive list of terms and definitions for terms commonly used in archaeology as a reference.

Unfortunately, the ontology is not geared towards NLP, as concepts are often defined in ways that do not mirror their use in running text, e.g. the entry for 'perforated axe' is '*bijl, doorboord*' (axe, perforated). Also, synonyms and lemmas/stems are not included, and terms might occur in multiple categories (e.g. 'Iron' as a material, or part of the time period Iron Age). While this does

[3]Available online at https://thesaurus.cultureelerfgoed.nl/

not pose a problem for creating a set of target labels for machine learning (as described in the next section), we are aware that this will cause noise in the term extraction described in Section 4.4.5, where we use entities as features in a classifier.

4.3.3 Definition of Categories

Classification is to be done in two dimensions: time periods and site types. The categories for time periods and site types are based on the ABR ontology. These codes are specifically defined for the description of Dutch archaeological concepts. In general, the ontology will provide us with a thesaurus, linking aforementioned codes, textual representations and corresponding descriptions. Furthermore, the ontology introduces sub-categorisation for both time periods and site types. Tables 4.2a and 4.2b show an overview of the categories we will take into account.

Ideally, we would also like to label the documents on artefacts (objects, e.g. an axe) and materials (e.g. flint), as these categories, combined with site type and time period, are the most used aspects in the information needs of archaeologists (Brandsen *et al.*, 2021b). Unfortunately, this is currently not possible as we do not have training data for these fields, because this information was not recorded for our training set.

4.3.4 Obtaining the document labels from the data

As mentioned briefly in the introduction, the data set has associated metadata for each document, as entered by the document authors at time of deposition in the DANS archive. The metadata entry was originally performed through a free text field, but has since been updated to dropdown boxes with specified ABR codes, and they are not required fields. Instructions for metadata entry are available on a separate page. Due to these factors, we see that the quality is relatively low: many documents are missing metadata, there are large inconsistencies between documents, and we even encountered wrongly entered metadata. To create a training set for document classification, we retrieve the manual metadata and clean it where possible, which is described below.

The retrieval of manually assigned metadata (time periods and site types) for each document is done by means of an XML crawler that uses the DANS Easy API.[4] All fields can have zero or more entries.

[4] easy.dans.knaw.nl/ui/home/

Time periods		
Label	**Category**	**Sub**
paleo	Paleolithic	5
meso	Mesolitic	3
neo	Neolithic	9
brons	Bronze Age	5
ijz	Iron Age	3
rom	Roman Time	9
xme	Middle Ages	8
nt	Modern	3

(a) An overview of the eight time period categories.

Site types		
Label	**Category**	**Sub**
xxx	Unknown	1
cthd	Cult / sanctuary	8
bewv	Habitation / settlement	32
apvv	Agricultural production	12
wrak	Shipwreck	3
idnh	Industry	21
sv	Shipping	8
gw	Resource extraction	9
bgr	Grave field	1
bgv	Burial (general)	17
infr	Infrastructure	25

(b) An overview of the eleven site type categories.

Table 4.2: Overview of the included labels, full names and the number of subcategories for each main category in time periods and site types. Category names are translated from Dutch.

4.3.5 Exploration of the Extracted Labels

We encountered several issues with the retrieved metadata values. First, there are over 1200 and 2600 unique metadata values retrieved via the XML crawler for the time periods and the site types respectively. Some of these metadata values are valid, but as stated in Sect. 4.3.3, we will only include a predefined selection of labels. Many other metadata values are simply not documented in the ABR ontology, instead being variations or older versions of actual labels, erroneously spelled labels, or completely irrelevant: for example names of cities instead of time periods. This reoccurring issue is because metadata was originally entered in a free text field where mistakes can be easily made. In Sect. 4.3.6 we describe how we processed the extracted metadata values into the set of predefined ABR labels set which we can use for classifier training.

Overall, more than 24,000 files do not have any metadata for the included time periods, and 29,500 files have no site type metadata (see Figure 4.1).

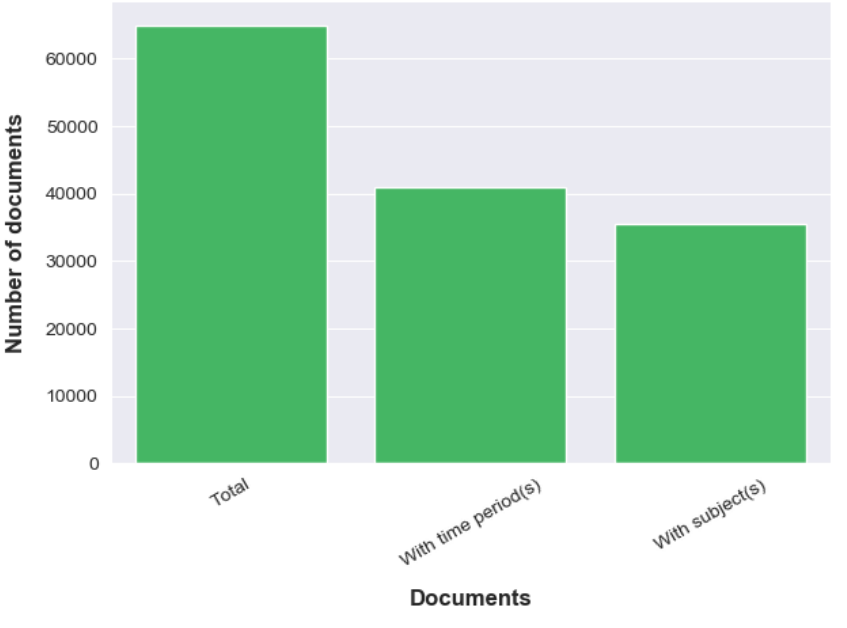


Figure 4.1: The number of documents and available metadata values.

4.3.6 Pre-processing the metadata

In order to introduce consistency, we convert all metadata values to a single, general format that only includes valid labels in the form of ABR codes. However, for time periods alone, over 1200 unique metadata values first have to be mapped onto the 45 labels (or 53 including main categories) we actually take into account. This process was done automatically where possible, but still required manual inspection and decision making regarding unclear metadata. This means that some unwanted labels are assigned to files, further affecting the classification process. In combination with the presence of erroneously assigned labels – those of correct ABR format, but simply not reflecting the content of the document – the training set will inevitably contain an unknown percentage of incorrect labels.

This will most likely harm the performance of the models to some extent, but without manually labelling a large amount of documents as a training set, it would be impossible to overcome this problem. For the test set, we do create a manually labelled set (see Sect. 4.4.4), so we can evaluate the performance even with a noisy training set.

For the site types, there were approximately 2,600 unique values in the re-

trieved metadata. Due to the high number of included categories – 11 main, 146 in total – we opted to only map labels in outdated ABR notation to current ones, and check for textual formats and their plural forms. Here, no further exhaustive manual labelling was done as the amount of metadata values and target labels is too large, making manual labelling too time consuming. Similarly to labelling the time periods, valid ABR codes might be erroneously assigned to documents, again decreasing the reliability of the training set.

After parsing the metadata for time periods to a valid ABR based format, we define the following rules to assign additional categories as to further introduce consistency in terms of time span:

- Whenever a file is labelled by a category of the lowest hierarchical level, then all parental categories will be assigned as well. For example, when a file is only labelled by *lmea* (Late Medieval A), then this file will be given additional labels *lme* (Late Medieval) and *xme* (Medieval – main category).
- When a file is only labelled by an intermediate level category, for example *lme*, its parental category will be assigned, *xme*, and its child categories, *lmea* and *lmeb*.
- When a file is labelled only by any main category, then *all* child categories from all hierarchical lower levels will be assigned as well.

We are aware that the last two rules are based on the following assumption: when someone labels a document as a top level time span (e.g. Medieval), they mean that items from the entirety of the Medieval period have been found, so from early to late Medieval. However, in some cases this will not hold true, as archaeologists often find items that can only be broadly defined as e.g. Medieval, and it is not clear from which of the sub-periods the item originates. Again, this will introduce some noise in the labels, as we cannot with certainty predict which sub-periods are actually present, but we still feel this is the most consistent way to generate our labelled data set.

For site types, there are only two levels of hierarchy. We will therefore limit the addition of categories to only main categories in cases where these are not yet included when only a sub-category is provided. When only a main category is present however, we will not assign any additional sub-categories, as the exact site type(s) cannot be derived.

After this process, we end up with an average of 8.1 labels per document (median: 4, max: 53) for time periods, and an average of 1.65 (median: 0, max: 18) for site types. Table 4.3 shows some examples of manually assigned metadata, and which labels were extracted after the pre-processing steps described above.

Assigned Metadata	Extracted Labels	Type of Conversion
ABR:NT	nt, nta, ntb, ntc	Sub-categories added
Late Middeleeuwen en Nieuwe Tijd	xme, lme, lmea, lmeb, nt, nta, ntb, ntc	Free text to label codes, with sub-categories
Gelderland; Ede;	None	Wrong metadata (location), no label assigned
Dijken, rivierduinen, prospectie, terpen	infr, infr.dij, bewv, bewv.tw	Free text to label codes, with main categories; only two out of four terms are valid ABR codes

Table 4.3: Examples showing the conversion of free text metadata entries to structured label codes.

4.4 Methods

In this section, we describe how we pre-processed the documents, modified the training set, constructed a manually labelled reference set, and selected the classification models.

4.4.1 Document Pre-processing

In order to prepare the textual data for classification tasks, we define several pre-processing methods, some of which are specifically targeting characteristics of observed noise, such as an abundance of punctuation or other non-alphabetical marks. Pre-processing steps include:

1. Lower-casing
2. Removal of all punctuation marks
3. Removal of abundant spacing
4. Removal of digits
5. Removal of all non-alphabetical marks
6. Stemming by means of NLTK's Snowball Stemmer[5] for Dutch words
7. Removal of tokens with a length equal to or less than three
8. Removal of stop words

[5] https://www.nltk.org/api/nltk.stem.html

We define ten combinations of these pre-processing steps, to find which aspects of the noise prove to be of most influence. For clarity, we will refer to each step by its corresponding number as defined in the list above. Some steps are mutually exclusive (i.e. 2 and 5), so we only use the following possible combinations: 128, 158, 13568, 135678, 1237, 1236, 156, 1567, 123, and 134.

It should be noted that these pre-processed texts are not suitable for all classification methods (further discussed in Section 4.4.5). Some only require lowercasing, while others require no pre-processing at all.

4.4.2 Document Filtering

We remove all documents that have fewer than 1000 utf-8 characters. Files shorter than 1000 characters rarely contain proper text, but are appendices with only numbers, or OCRed maps resulting in a file with nonsensical characters.

In addition, we remove non-relevant documents from the data set. This relevance is based on certain terms occurring in the title, indicating it is a specific type of non-relevant document. We define two lists, the first consists of a few general terms: *notulen* (minutes), *bijlage* (appendix) and *meta* (metadata). The second list is more extensive, and includes several types of reports (RAP), working methods (PVA), requirements definitions (PVE), referential research IDs (OMN) and the aforementioned general terms. A complete overview can be found in Appendix B. For upcoming experiments, we refer to the first list consisting of general terms as *genList*, and the extensive list as *totList*.

It should be noted that while these documents are removed from our training and test set, this should not affect the usefulness of the methods on new data. Short documents that do contain useful information can still be labelled by the classifier. The document types in the genList and totList that we here exclude are most often grouped in a DANS data set with associated ID, together with the main report. When this main report has been classified, we can propagate the labels to all documents in that data set, ensuring useful metadata for all related files.

4.4.3 Balancing the Training Set

As can be seen in Figures 4.2 and 4.3, the distribution of the labels among categories is rather skewed. Some categories are not represented very well, leading to an imbalanced data set. As this might induce bias to some classifier types, we introduce two methods that may negate this. The first is balancing of the training set through under-sampling, i.e., reducing the number of documents of a class

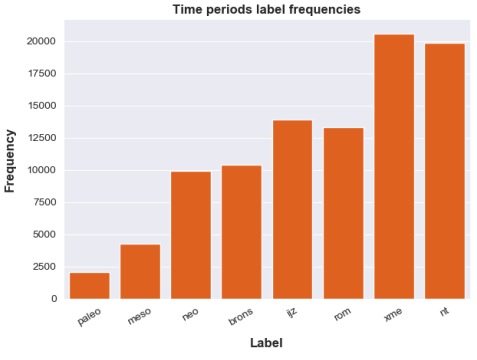


Figure 4.2: An overview of the frequencies of the eight time period categories. X axis labels as per table 4.2a.

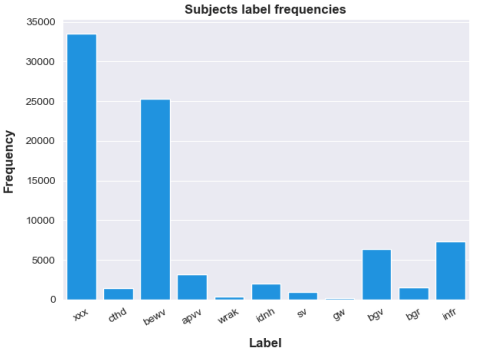


Figure 4.3: An overview of the frequencies of the eleven site type categories. X axis labels as per table 4.2b.

until it equals that of the class with the lowest representation. Under-sampling has been proven to be a reliable method for addressing the imbalance of a data set regarding the distribution of present labels (Branco *et al.*, 2015; Mohammed *et al.*, 2020).

Another option, which primarily aims to create more valid training samples, is increasing the representation of all labels through augmentation. Here, we enlarge the training set by including files multiple times, but applying a synonym mapping function to the duplicate files to avoid bias on certain terms, while still maintaining context as much as possible. We adapt the Easy Data Augmentation (EDA) method proposed by Wei & Zou (2020). Synonyms are chosen at random with the use of the Open Dutch WordNet (Postma *et al.*, 2016) synonym thesaurus. The augmentation should be applied to the complete corpus in order to introduce a large variety of terms, rather than merely the archaeological tokens captured within the texts. We therefore decided to make use of a thesaurus that meets this requirement, not limiting ourselves to a domain specific, in this case an archaeological, thesaurus. Contrary to the EDA method, we insert synonyms for all words longer than five characters – as opposed to a specific number of tokens based on sentence length. This is because the sentence length is in many cases simply impossible to properly determine due to noise in the text. This could potentially lead to too much semantic change in the text for it to be useful, but we found this process can lead to higher performance in some cases (as further described in the Sect. 4.5).

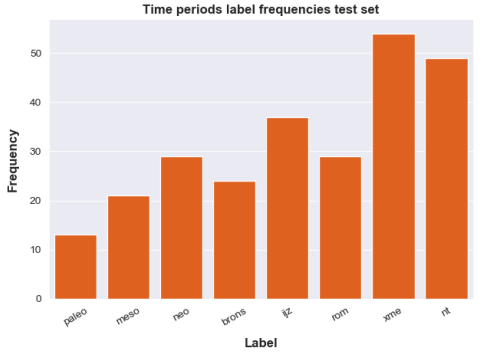


Figure 4.4: An overview of the frequencies of the eight categories for time period classification, as captured within our reference set.

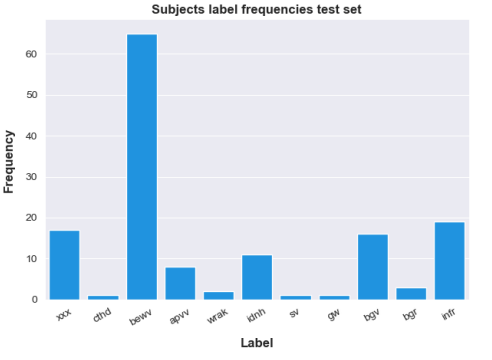


Figure 4.5: An overview of the frequencies of the eleven categories for site type classification, as captured within our reference set.

4.4.4 Construction of a Manually Labelled Reference Set

Because of how we constructed the labels from the data, it would be impossible to ensure that all files within a randomly sampled sub-set hold only correct labels. This means that even our test set would include an unknown percentage of incorrectly labelled documents. Naturally, this is undesirable, as no valid conclusions can be drawn from a flawed test set.

In order to deal with this issue, we created a manually labelled reference test set (Brandsen, 2020), of which we are certain that it consists of correctly labelled documents only. As manual labelling is very time consuming, this test set consists of 'only' 100 files. Figures 4.4 and 4.5 show the frequencies for each of the categories captured within the classification of time periods and site types, respectively. The average number of labels per document is 13.9 for time periods (median: 11, max: 53) and 2.79 for site types (median:2, max:13).

The distributions of the test set are similar to those of the training set, as were shown in Figures 4.2 and 4.3. The only exception is the category of label *xxx* (unknown) for the site types. This is because all files in our test set are labelled by at least one time period, and many files labelled by *xxx*, i.e., reports about sites with no finds, are not labelled by any time period. A complete overview that includes all the frequencies of all main and sub-categories can be found in Appendix C.

4.4.5 Classification Methods

We compare three methods for the classification of time periods and site types: a (naive) baseline, binary relevance, and direct multi-labelling. All methods will be trained and optimised using a train and development set, and finally evaluated on the held-out test set consisting of the manually labelled reference set mentioned above.

Baseline For the baseline, we introduce the rather intuitive method of merely checking whether the label or its corresponding textual version is present within the text, and assign labels accordingly. The minimum occurrence for such tokens in the text is set to two, as often lists of ABR codes are present as period or site type lists. Naturally, these are uninformative to our research.

Binary Relevance We translate the multi-label task to a series of binary classification tasks, one for each category, and train a Linear Support Vector Machine (SVM) classifier for each category. We compare four feature extraction methods:

- A bag-of-words model with TF-IDF weighting;
- A Doc2Vec model (De Romas, 2019) for each individual binary classification task. The model has a vector size of 100, a window of 5, an initial learning rate of 0.025, a minimum learning rate of 2.5e-3, and a minimum count of 5 (ignores all tokens with a frequency lower than 5). We let the model train for 5 epochs.
- Using entities as features. Besides applying pre-processing methods, we also investigate the effects when classification is done solely on extracted named entities, again using a bag-of-words model with TF-IDF weighting. We extracted entities based on the ABR ontology. Here, we extract all terms (time periods, site types, corresponding abbreviations, etc.) contained in the ontology from the text, and use this as our input.
- Same as above, but we perform the extraction of entities by means of spaCy (Honnibal & Montani, 2017), using its pre-trained Dutch model[6]. Here, we select entities from any of the following types[7]: *FAC* (groupings), *NORP* (structures) and *DATE* (dates or periods).

For the third method, we are aware that the problems with the ABR ontology as described in Sect. 4.3.2 will cause noise to some extent. Specifically, as no

[6] https://spacy.io/models/nl
[7] https://spacy.io/api/annotation#named-entities

synonyms are available in the ontology, and we do not use lemmatisation or stemming, extracting terms from the text is going to have a low recall. Also, only time period names are included in the ABR, so actual dates (e.g. '1000 BCE') will not be extracted. Despite these issues, we still considered this worthwhile to experiment with, as this method can be improved by using more advanced NER methods if promising results are achieved.

Direct Multi-labelling Finally, we make use of BERT, a state-of-the-art classification model. We incorporate the Simple Transformers library[8] for faster training and evaluation. Using the pre-trained bert-base-multilingual-cased model (Devlin *et al.*, 2019), we use the following default parameter settings to evaluate the method: a train batch size of 4, gradient accumulation steps of 1, a learning rate of 3e-5, and a max sequence length of 256 due to memory constraints. The model will be trained over 3 epochs.

Initially, we limit the classification task to only the top level categories, and will use the results to determine which setting works best for any particular category.

4.4.6 Selection round

In summary, we have six approaches (baseline, four binary, one direct multiclass classification), ten pre-processing combinations, the option of augmenting as well as balancing the training set, and filtering files based on document title. Exploring all applicable different settings on each of these approaches will most likely lead to an abundance of scores that are far from optimal, and not very interesting. We therefore first run each of the approaches on the raw (no pre-processed versions) of the documents, and determine how each method performs with respect to the baseline and one another. To limit the aforementioned parameter exploration, we will continue with the two best performing approaches for the time periods and site types, based on the F1 score.

One aspect that should be taken into account is that BERT in particular should theoretically already be performing closer to optimal compared to the binary translation approaches, as pre-processing is not required for this method.

[8] https://pypi.org/project/simpletransformers/

Accuracy metrics time periods			
Approach	**Prec.**	**Rec.**	**F1**
Baseline	0.500	0.318	0.358
TF-IDF	0.848	0.621	0.703
D2V	0.747	0.500	0.577
ONT	0.854	0.506	0.602
SCY	0.795	0.484	0.565
BERT	0.745	0.519	0.585

Accuracy metrics site types			
Approach	**Prec.**	**Rec.**	**F1**
Baseline	0.161	0.622	0.232
TF-IDF	0.633	0.355	0.408
D2V	0.313	0.282	0.254
ONT	0.434	0.270	0.259
SCY	0.272	0.140	0.121
BERT	0.225	0.151	0.146

Table 4.4: Overview of the scores for each method. Abbreviations refer to the following: TF-IDF (Sklearn, linear SVM with TF-IDF weights), D2V (Sklearn, linear SVM with Doc2Vec vectors), ONT (Sklearn, linear SVM classification based on ontology extracted entities) and SCY (Sklearn, linear SVM classification based on spaCy retrieved entities).

4.5 Results

In this section, we present our results. We first determine how each approach performs on the data set with no modifications, and then select the top two performing approaches for further research. We then investigate the effects of different parameter settings, determine the best possible method per category, and finally perform the classification task on all categories.

4.5.1 Selection Round

We have a baseline and five approaches we will evaluate first. The obtained precision, recall and F1 scores can be seen in Table 4.4. All scores are the macro average over all categories within the corresponding field. For TF-IDF, D2V, ONT and SCY (acronyms explained in the table caption), a linear support vector classifier was used. For BERT, we used the pre-trained bert-base-multilingual-cased model[9]. The two best performing approaches are highlighted in green.

For the time periods, the baseline F1 score of 0.358 is substantially outperformed by the other five approaches. Even without pre-processing, the four binary classification approaches, TF-IDF, D2V, ONT and SCY already lead to decent results. As highlighted, TF-IDF and ONT score the highest, the former by a noticeable amount. BERT unfortunately does not yield very promising results, particularly so as this approach does not require any prior pre-processing on the

[9]https://huggingface.co/transformers/pretrained_models.html

Dev Rank	Test Rank	PP	Aug	Precision	Recall	F1
1	3	1237	0	0.873	0.639	0.719
2	8	134	0	0.856	0.602	0.681
3	6	134	2	0.869	0.597	0.692
4	7	123	2	0.865	0.602	0.684
5	5	158	0	0.857	0.635	0.709
6	1	128	2	0.873	**0.674**	**0.752**
7	4	123	0	**0.880**	0.631	0.711
8	2	128	0	0.879	0.652	0.730
9	10	1237	2	0.874	0.568	0.658
10	9	1236	0	0.863	0.605	0.680

Table 4.5: Overview of the top ten F1 scores for time period classification. PP = numerical values referring to pre-processing steps as described in Section 4.4.1, Aug = number of augments of the training set.

Dev Rank	Test Rank	PP	Aug	Precision	Recall	F1
1	7	123	2	0.626	0.360	0.410
2	4	13568	2	0.637	0.464	0.496
3	3	128	0	0.601	0.462	0.498
4	9	1236	0	0.542	0.347	0.379
5	10	134	2	0.539	0.330	0.366
6	1	158	2	0.640	**0.499**	**0.542**
7	2	128	2	0.702	0.469	0.510
8	8	123	0	0.538	0.345	0.390
9	6	1237	2	**0.715**	0.442	0.482
10	5	1237	0	0.609	0.447	0.484

Table 4.6: Overview of the top ten F1 scores for site types classification. PP = numerical values referring to pre-processing steps as described in Section 4.4.1, Aug = number of augments of the training set.

texts.

For the site types, we find that the baseline performs better than both SCY and BERT, the latter two yielding an F1 score of less than 0.15. Again, TF-IDF and ONT give the best results, though only by a very small, almost negligible margin when comparing ONT to D2V. Nevertheless, we continue with TF-IDF and ONT for both time periods and site types, and will now look at pre-processing optimisation.

4.5.2 Pre-processing Optimalisation

We applied a brute force approach, trying out all 176 combinations of pre-processing steps, balancing/augmenting the training set, and further pruning the training set based on document titles.

The performance metrics were determined by averaging the F1 scores over three separate evaluation rounds. During each round, the training set was split into a 4:1 ratio, retaining a suitable training set size and introducing a smaller development set.

Tables 4.5 and 4.6 show the top ten performing settings, ordered by obtained F1 scores on the development set, but showing the performance metrics on the held out test set. The second column, labelled *Test Rank* indicates which ranking the top ten performance settings on the development set achieve when these same settings are applied to the test set. The ranking captured within the *Test Rank* column thus reflects the ordering of the *F1 scores*, which are shown in the rightmost column. The top ten combinations all use the bag-of-words model with TF-IDF weighting, classifier Linear SVC, no balancing and the GenList document pruning list, so these are not mentioned in the tables.

The results show that rather short combinations consisting of only three or four pre-processing steps lead to the overall highest results in combination with the SVM classifier. Steps 1, 2 and 3 occur almost everywhere. These are lower-casing, removing punctuation marks and removing abundant white space, which are expected to help with classification as these are commonly used.

Augmentation of the training set does not necessarily seem to have a positive effect on the classification process as it only leads to higher F1 scores with certain pre-processing combinations. Finally, we can make the observation that filtering files based on terms included in genList also leads to better performance for both time periods and site types, whereas totList does not appear in any of the top ten rankings.

Despite these scores being the average over three runs, the balancing and augmentation is a rather randomised process. It is therefore possible that a lot

'bad' or 'good' files are filtered out, i.e., files that have (un)informative content. This would mean that the performance metrics could vary slightly when the experiments are to be repeated, perhaps resulting in a different ranking.

Lastly, the development and test ranking orders provide some interesting insight into how representative the defined development sets were compared to the reference set. We can see that for both time periods and site types, the best performing settings on the test set are found at rank six for the development set. As the optimal development and test F1 scores differ quite heavily from one another, the quality of the development sets do not match that of the test set. This was to be expected, as the training set, and therefore the development sets contain an unknown percentage of wrong labels.

4.5.3 Best Methods per Category

The above section shows which approach and parameter settings lead to the highest average F1 scores, and here we investigate if we can achieve a higher average F1 score by combining the best approaches and settings for each individual category. The results for time periods and site types are shown in Tables 4.7 and 4.8, respectively.

For time periods, combining the best method per individual category leads to an average F1 score of 0.710, which is a slight decrease compared to the 0.719 of the settings with the best F1 average over all categories. This again can be explained by the quality of the development sets: by using the optimal parameter settings for a category obtained on the development set, it unfortunately does not imply that these settings are (close to) optimal on the test set. This phenomenon is similar to that observed in the previous section, were the best parameter settings for the test set ranked sixth on the development set. For the site types, the opposite shows, as we find an average increase of 0.133 compared to the highest scoring settings on the development set. Moreover, the F1 score of 0.542 – the result of optimal settings for the test set – is met. It has to be noted that we find F1 scores of 0.0. These categories are barely represented within our test set, and for these it is difficult to determine the quality of the classification process: a recall of 0.0 is frequent.

The MultNB classifier does not appear in the top ten. We expected to see that balancing the training set would have a positive effect on the classification process for this classifier, but this is not reflected by our results. However, it is interesting to see that balancing the training set has a positive effect on the classification process of SVM for *neo* and *ijz*, despite the theoretical unbalanced data set 'protection'. Again, this can be explained by the random influence on

Category	PP	Aug	Bal	List	Precision	Recall	F1 score
paleo	123	2	No	Gen	1.0	0.385	0.555
meso	134	2	No	Gen	1.0	0.550	0.710
neo	123	2	Yes	Gen	0.653	0.630	0.642
brons	158	2	No	Gen	0.714	0.435	0.541
ijz	134	0	Yes	Gen	0.828	0.828	0.828
rom	128	0	No	Gen	0.952	0.741	0.833
xme	1236	0	No	Gen	0.764	0.823	0.792
nt	134	0	No	Gen	0.722	0.848	0.780
average	-	-	-	-	0.829	0.655	0.710

Table 4.7: Overview of the best methods per individual category for time period classification and the overall average of these best methods. Column names yield the meaning as provided in the previous section.

Category	PP	Aug	Bal	List	Precision	Recall	F1 score
xxx	1237	0	No	Tot	0.342	0.765	0.473
cthd	156	2	No	Gen	1.0	1.0	1.0
bewv	123	0	No	Gen	0.810	0.557	0.660
apvv	128	0	No	Gen	0.667	0.286	0.400
wrak	13568	0	No	Tot	1.0	0.5	0.667
idnh	123	2	No	Gen	0.800	0.444	0.571
sv	1236	2	No	Gen	1.0	1.0	1.0
gw	134	2	No	Gen	0.0	0.0	0.0
bgv	156	2	No	Gen	0.875	0.538	0.667
bgr	128	2	No	Gen	0.0	0.0	0.0
infr	1237	2	No	Gen	0.875	0.389	0.538
average	-	-	-	-	0.669	0.498	0.543

Table 4.8: Overview of the best methods per individual category for site type classification and the overall average of these best methods. Column names yield the meaning as provided in the previous section.

the balancing and augmenting process, as 'bad' files get filtered out.

We have determined which settings work best for each main category, the next step is to perform the classification task on all sub-categories by using the settings per corresponding main category. As not all sub-categories for site types are present within our test set, we will only focus on those that were. The full classification results can be seen in Table 4.9 and 4.10.

For any set of sub-categories, we expected to find a lower average F1 score than the corresponding main category, as there are most likely less distinctive terms between sub-categories. This indeed seems to be case for the majority of the categories, but a few exceptions for both time periods and site types are present. We note that in some cases for the site types, F1 scores of 1.0 are found. These (sub-)categories are only represented once. Nevertheless, it does imply that the classifier returns a perfect prediction on our test set. We also find numerous F1 scores of 0.0, which as mentioned earlier is the result of frequent recall values of 0.0.

Such scores are not very indicative of the quality of the classification process itself, but rather implies an insufficient amount of labelled data for that category. For completeness however, we decided not to omit these results from aforementioned tables.

To further illustrate the relation between the frequency of a label in the training set and the achieved F1 scores, we plotted these in Figures 4.6 and 4.7. We can see that – as expected – the higher the frequency of the label is, the higher the performance, as illustrated by the trend lines. We also note that the trend lines are not flattening out, which indicates that adding more training data might be beneficial for all categories, not just the less frequent ones.

4.6 Conclusion

In this paper, we have described our approach for the multi-labelling of Dutch archaeological excavation reports for time periods and site types. In this section we answer our research questions and propose future work.

> *Which combination(s) of text pre-processing steps, data augmentation/balancing, document pre-selection and classification method yields the highest F1 scores?*

We tested many combinations of pre-processing steps, and found that lowercasing, removing punctuation marks and trimming white space are most valuable on average, which is expected as these steps are used widely in text classification

All time periods categories: obtained F1 scores overview									
Label	F1	Label	F1	Label	F1	Label	F1	Label	F1
paleo	0.555	neov	0.591	bronsm	0.583	romvb	0.700	vmec	0.439
paleov	0.600	neova	0.605	bronsma	0.522	romm	0.833	vmed	0.439
paleom	0.667	neovb	0.667	bronsmb	0.640	romma	0.809	lme	0.800
paleol	0.500	neom	0.619	bronsl	0.500	rommb	0.833	lmea	0.756
paleola	0.500	neoma	0.537	**ijz**	0.828	roml	0.780	lmeb	0.787
paleolb	0.500	neomb	0.585	ijzv	0.750	romla	0.800	**nt**	0.780
meso	0.710	neol	0.681	ijzm	0.644	romlb	0.810	nta	0.738
mesov	0.455	neola	0.667	ijzl	0.719	**xme**	0.792	ntb	0.764
mesom	0.500	neolb	0.696	**rom**	0.833	vme	0.455	ntc	0.689
mesol	0.571	**brons**	0.541	romv	0.700	vmea	0.450		
neo	0.742	bronsv	0.483	romva	0.700	vmeb	0.450		

Table 4.9: An overview of the F1 scores for all main and sub-categories for time period classification. Main categories are denoted in bold.

All site type categories: obtained F1 scores overview									
Label	F1	Label	F1	Label	F1	Label	F1	Label	F1
cthd	1.0	bewv.hp	0.000	idnh.hkb	0.0	bgv.x	0.0	**bgr**	0.0
cthd.klo	1.0	bewv.bext	0.667	idnh.ll	0.0	bgv.gvc	0.667	bgr.gvic	0.0
bewv	0.857	**apvv**	0.400	idnh.m	0.0	bgv.gvi	0.0	**infr**	0.571
bewv.x	0.756	apvv.x	0.0	idnh.pb	0.0	bgv.gvx	0.5	infr.x	0.0
bewv.vx	0.0	apvv.cf	0.0	idnh.vb	0.0	bgv.kh	0.0	infr.weg	0.0
bewv.vlp	0.0	apvv.la	0.333	idnh.mb	0.0	bgv.ghv	0.500	infr.per	0.800
bewv.kwb	0.0	**wrak**	0.667	**sv**	1.0	bgv.cjbp	0.0	infr.kan	0.667
bewv.ht	0.0	wrak.schip	0.667	sv.x	1.0	bgv.uv	0.400	infr.brug	0.667
bewv.vic	0.0	**idnh**	0.800	**gw**	0.0	bgv.gx	0.0	infr.dij	0.889
bewv.sk	0.667	idnh.x	0.286	gw.vw	0.0	bgv.vg	0.0	**xxx**	0.473
bewv.rv	1.0	idnh.tn	0.0	**bgv**	0.667	bgv.dier	0.0		

Table 4.10: An overview of the F1 scores for the main and sub-categories for site type classification. Sub-categories not present within the reference test set are not included. Again, main categories are denoted in bold.

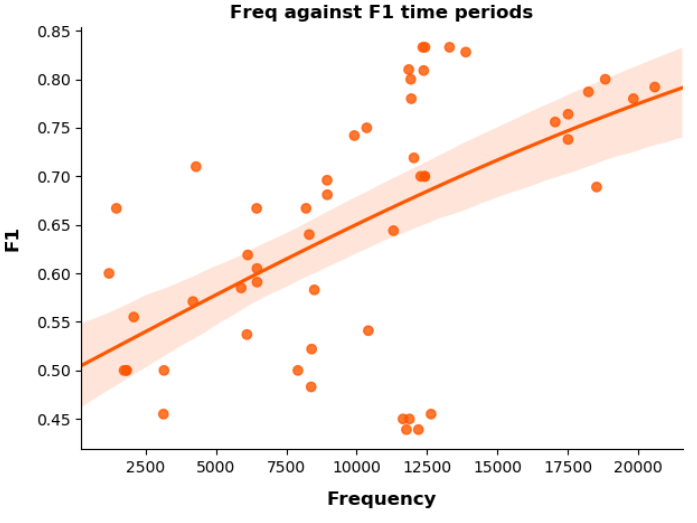


Figure 4.6: Plot of the frequency of time period labels and the associated F1 score for that label. A trend line has been added to illustrate the correlation (Pearson's $r = 0.56$).

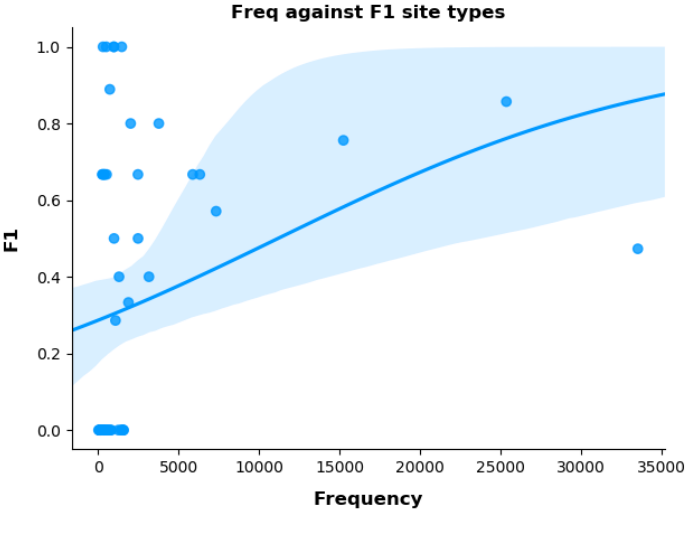


Figure 4.7: Plot of the frequency of subject labels and the associated F1 score for that label. A trend line has been added to illustrate the correlation (Pearson's $r = 0.28$).

problems. Balancing the data set did not lead to better results, and augmentation helped in only some cases, so we can not draw any conclusions on this. Pruning the data set by using the standard filename list proved to be most effective. As for the classification method, using a linear SVM proved to be optimal. In addition, we found that classification on extracted entities by means of the ontology did not yield very promising results.

> *Are the best combinations the same across the different categories and labels, or do specialised combinations per category yield better results?*

We investigated whether optimising the methods per (sub-)category leads to higher performance. We found that the optimal parameter settings per individual category for the time periods actually lead to a lower averaged F1 score when compared to the top performing setting over all categories at once. For site types the F1 score is the same. It suggests that for these kinds of classification problems, using the same parameters for all the categories is not only better, but also much simpler as only one model needs to be trained, instead of a model for each category.

> *To what extent can we classify excavation reports into time periods and site types?*

Our overall aim was to test how well we could classify excavation reports, and we found that despite the frequent low quality of both texts and labels, our classification models lead to decent quality when compared to similar studies. For the classification in eight time periods, we obtained an F1 score of 0.752 with settings that were found to be optimal on the held-out test set. These included only a few pre-processing steps, no balancing, and a small selection for filtering documents based on their titles. For the classification in eleven site type categories, we obtained an F1 score of 0.542 with highly similar settings, except for a single different text pre-processing step (removal of non-alphabetical marks instead of removal of punctuation marks) and the augmentation of the training set.

One caveat to these results is that there is a large deviation in the results obtained with different partitions of the data, with the top ten highest scoring partitions of the development set leading to F1 scores on the test set ranging from 0.68 to 0.75 for time period classification and from 0.36 to 0.54 for site type classification.

We expected to see that the average F1 scores over a set of sub-categories would be lower than that of the corresponding main category. This was indeed

the case apart from a few exceptions. We argued that this phenomenon is caused by a smaller number of distinctive terms for sub-categories when compared to solely main categories.

As predicted, the limited input sequence of 256 for BERT led to quite disappointing results, considering this method is regarded as a state-of-the-art approach for multi-label classification tasks. In particular for the site type classification, performance metric scores for BERT were almost bottom tier.

4.6.1 Future Work

There are several aspects that could prove interesting for follow-up research. At the moment, we are dealing with a data set that has manually assigned metadata for the entire collection. This means our methods are not tested on unlabelled, or partially labelled data. It would be interesting to research this, to see to what extent the usefulness of the metadata increases. We plan to do this research when we receive reports without metadata in a follow-up project.

As we were particularly concerned about the effect the quality of the labels and the texts would have on the classification process, we put more emphasis on the application of exploratory parameter settings based on statistics on observations, rather than using all the five approaches. It could prove to be interesting to apply the parameter settings to each of these, and eventually perform hyper-parameter optimisation. Ideally, we would like to create a manually labelled training set to increase the quality of the data, and determine how this affects the performance of our methods. Due to time constraints we have not been able to do so in yet. If this proves too time-consuming, an alternative might be k-fold validation to average out the difference in label quality across the training set.

Initially, we opted for NER based classification by means of a specifically designed NER tool for archaeological named entities. Unfortunately, this tool had not been fully developed yet, and could not be used. SpaCy based NER classification already lead to promising results – scoring second highest for both time periods and site types – despite a lack of entity categories that were specific to our type of documents. If such categories were to be extracted however, classification on such entities might lead to even better results.

A third aspect that could be addressed is that of balancing: we might be able to determine which files are often included in a training set that leads to lower performance. This would arguably imply that such files are either uninformative, or have erroneous labels. Removing these documents will most likely lead to higher overall performance.

Furthermore, there is the option of optimising the BERT approach. Cur-

rently we only use the first 256 tokens of a text due to memory and framework constraints. Distinctive and characteristic terms for categories could therefore be missing in data used for either training or eventual classification, leading to lower performance. Increasing the token limit, or potentially classifying smaller segments, might give us better results.

Finally, an expansion of the test set could be introduced in order to enhance the representation of the categories. This in particular applies to the categories of site types. As discussed in Sect. 4.5.3, we find an F1 score for numerous site type categories to be equal to 0.0 or 1.0. Because of the low representation of these categories, such scores are not meaningful, and therefore do not properly reflect on the quality of the classification process.

5

User Requirement Solicitation

"*Improve a mechanical device and you may double productivity. But improve man, you gain a thousandfold.*"
Khan Noonian Singh, Star Trek TOS, s01e22 'Space Seed'

Previously published as: Brandsen, A., Lambers, K., Verberne, S. and Wansleeben, M., 2019. 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), pp.21-30. DOI: 10.5334/jcaa.33

In this paper, we present the first results of applying Named Entity Recognition and Information Retrieval techniques to tackle the problem of unused grey literature in archaeology, specifically Dutch excavation reports. We used Conditional Random Fields to identify entities, with an average accuracy of 56%. This is a baseline result, and we identified many possibilities for improvement. These entities were indexed in ElasticSearch and a user interface was developed on top of the index. This proof of concept was used in user requirement solicitation and evaluation with a group of end users. Feedback from this group indicated that there is a dire need for such a system, and that the first results are promising.

5.1 Introduction

The archaeological world creates huge amounts of text in different formats, from books and scholarly articles to unpublished fieldwork reports. These reports are also known as grey literature. Easy access to the information hidden in these texts is a substantial problem for the archaeological field. Making these documents searchable and analysing them is a time consuming task when done by hand, and will often lack consistency. Text Mining and Information Retrieval (IR) provide methods for disclosing information in large text collections, allowing researchers to locate (parts of) texts relevant to their research questions, as well as being able to identify patterns of past behaviour in these reports (Richards *et al.*, 2015).

The Malta convention (or Valletta Treaty) is a European treaty, signed on 16 January 1992. It came into effect on 25 May 1995, and its aim is to protect archaeological remains by making "the conservation and enhancement of the archaeological heritage one of the goals of urban and regional planning policies" (Council of Europe, 1992, Art. 1). The convention was implemented in the Netherlands via the Archaeological Heritage Management Act in 2007 (Ministerie van Onderwijs Cultuur en Wetenschap, 2007). Preferably, preserving these remains is done by keeping them in situ, but when this is not possible, the developer disturbing the ground record is required by law to pay for the archaeological research. This research is generally performed by commercial archaeology units.

This archaeological research has created a collection of texts that is too large to be completely read by humans. The amount of reports created in the last 20 years is currently estimated at just under 60,000, and is growing by approximately 4000 per year (Rijksdienst voor het Cultureel Erfgoed, 2019a). Most of these reports are categorised as 'grey literature' (Evans, 2015), and are likely to end up in a proverbial 'graveyard', unread and unknown, unless they are properly

archived, indexed and disclosed.

In the Netherlands, the SIKB (*Stichting Infrastructuur Kwaliteitsborging Bodembeheer*) creates and maintains the standards of activities relating to soil management. As stipulated in their BRL 4000 guidelines, a report has to be deposited into an e-depot within 2 months of completing the project (Stichting Infrastructuur Kwaliteitsborging Bodembeheer, 2016, Art. 2.6.2). While some companies and municipalities are still reluctant to deposit their reports into national e-depots (instead opting to deposit in small local depots) most reports and the associated metadata do end up in one of three of the main e-depots of the Netherlands; the Data Archiving and Networked Services (DANS) repository, the Document Management System of the *Rijksdienst voor Cultureel Erfgoed* (RCE) or the *Koninklijke Bibliotheek* (KB) e-Depot. There is considerable overlap between the DANS, RCE and KB data sets, and altogether it is estimated they hold around 70 percent of all so-called Malta reports. This means that a large portion of the reports is currently available, and access to the files is not a major problem at the moment.

This paper describes the work carried out in the first year of a PhD project. This project is in association with both the Faculty of Archaeology and the Data Science Research Programme (DSRP) at the University of Leiden, combining archaeological knowledge with the technical skills available in the Data Science department.

The work carried out in this project is motivated by the need from researchers in the archaeological field to be able to efficiently and effectively find information related to their research questions in the available grey literature. This requirement has been well documented in previous work (e.g. Richards *et al.*, 2015; Van den Dries, 2016) and some studies have investigated different applications of Text Mining from archaeological reports in English (Vlachidis & Tudhope, 2016; Amrani *et al.*, 2008; Byrne & Klein, 2010) and Dutch (Paijmans & Brandsen, 2010; Vlachidis *et al.*, 2017).

However no system is currently available that allows full-text access to at least a major part of the Dutch archaeological corpus, or document collection. As a result, relevant and valuable information is not being utilised by some researchers, mainly those who are not experts in their field yet. Information like a single Bronze Age find in a otherwise Medieval site is unlikely to be mentioned in the metadata, and is thus nearly impossible to find. This is a problem from a theoretical point of view, as key information could be overlooked at the moment, information that could change archaeological interpretations. It also devalues the monumental effort that has gone into collecting, digitising, archiving and publishing these documents, as well as the legislation that has been drawn up surrounding the archiving of these documents.

More and more Text Mining, data mining and IR tools and techniques have become available over the last years, which could potentially provide a way to access and extract information from this wealth of data currently hidden in these reports. This, combined with the relatively easy access to higher computer processing power, makes a systematic implementation of Text Mining techniques for Dutch archaeological reports not only desirable, but also feasible.

In this project we are developing AGNES (**A**rchaeological **G**rey-literature **N**amed **E**ntity **S**earch), a search system that aims to make archaeological grey literature more accessible and searchable by applying IR techniques to this big data set.

The goals of this paper are (1) to give an overview of previous work on Text Mining in archaeology, (2) to show the need for a search system by interviewing the user group, (3) soliciting user requirements for such a system, (4) presenting the results of the initial experiments with Named Entity Recognition (NER) and (5) presenting the indexing and front end software of the developed system.

5.2 Prior work

Some experiments have been carried out in Text Mining and NER in archaeology, across multiple countries and languages. In English, one of the earliest contributions is the work by Amrani *et al.* (2008), which helped experts to extract information from archaeological literature. Byrne & Klein (2010) also investigated the extraction of information, but focused solely on event information. The OPTIMA system, described by Vlachidis (2012), used a rules-based approach to semantic indexing, including NER. Another notable project is Archaeotools in the UK, which combined databases with information extracted from reports in an interesting faceted browser interface (Jeffrey *et al.*, 2009). A more recent paper is that by Kintigh (2015), which provides a detailed overview of the problems and possible solutions, but does not include the development of a search system.

For Dutch language reports, most of the previous research has been carried out by Paijmans with several collaborators, including extracting monument names from free text fields (Paijmans & Brandsen, 2009) and the OpenBoek system, which used memory-based learning to perform NER (Paijmans & Wubben, 2008; Paijmans & Brandsen, 2010). Like the work by Byrne & Klein (2010), this project focused mainly on time periods, but also applied some rules-based NER to detect place names. The OpenBoek system included an online search interface during the Continuous Access To Cultural Heritage (CATCH) project, but unfortunately this isn't available anymore.

A notable contribution is that by Mélanie-becquet *et al.* (2015), who ran a pilot study on texts from a part of France, dealing with the Iron Age till the Medieval period. They performed NER and other techniques similar to some of the previously discussed projects, but they did this multi-lingually, including French, German and English. Unfortunately, the technical details of their work don't seem to be published yet.

More specifically, this project builds upon the Text Mining experiments performed by researchers of the University of South-Wales in the European ARIADNE project between 2013 and 2017. They applied a rules-based technique to the problem, utilising the GATE framework[1]. Leiden University participated in this project and a limited number of eight Dutch reports were analysed and compared to manually tagged 'gold-standard' documents as a proof of concept, next to English, Swedish and German reports. In the same project, the ADS (Archaeological Data Service) in the UK applied machine learning techniques to English grey literature, and developed an API that can automatically create metadata based on entered text (Vlachidis *et al.*, 2017).

The contributions of this paper compared to previous work are twofold: (1) this system includes a user study which hasn't previously been undertaken; and (2) it combines the results of the NER with a full-text index in an effective search interface, instead of just focusing on the NER.

More broadly, this project is in cooperation with the DSRP, which gives us access to a high computing power cluster, allowing for the use of more computationally expensive techniques on bigger document sets. The length of this project is also of importance; most previous experiments were often performed over a short amount of time, making it difficult to create a finished system, while this project takes place over four years with the specific aim of creating a user-friendly web application.

5.3 Introducing AGNES

AGNES is an acronym that stands for **A**rchaeological **G**rey-literature **N**amed **E**ntity **S**earch, and is the name of the search system currently under development in this project, including both the front end of the web application, as well as the indexing software responsible for finding and indexing archaeological concepts. The current version of the system (v0.2) is available at `http://agnessearch.nl/index.php/search/agnesv02`.[2]

[1] See also https://gate.ac.uk

[2] Please note, free registration is needed to access the system.

	Synonomy		**Polysemy**
Main Term	*Neolithic*	**Main Term**	*Swifterbant*
Synonyms:	Late Stone Age	**Meanings:**	Time Period
	3000 BC		Excavation
	5000BP		Pottery Type
	4th Millenium BC		Location

Table 5.1: Synonymy and Polysemy examples

5.3.1 Named Entity Recognition

A standard full-text index, allowing researchers to search through all of the text instead of just the metadata, would already be an improvement on the current situation. However, such a full-text search would not account for synonymy and polysemy; multiple words that have the same meaning and one word having multiple meanings, respectively. See table 5.1 for two non-exhaustive examples, where a full-text search would either not return all results, or return possibly wrong results. This is why NER is needed to accurately index these documents.

Named Entity Recognition is a method that aims to identify and classify specific entities in natural language, also known as unstructured written text (Marrero *et al.*, 2013). In the case of this project, the entities are archaeological concepts, and the natural language are excavation reports. To give an example, in the following sentence the entities are underlined: "We found pottery dating from the Neolithic inside a rubbish pit", an artefact, a time period and a feature, respectively.

In the current version of the system, we used Conditional Random Fields (CRF). This is a form of machine learning specifically designed to label sequence data (Lafferty *et al.*, 2001), a common choice for NER tasks as words in a sen-

Figure 5.1: AGNES Logo

tence are sequential. We implemented the scikit-learn Python package (Pedregosa *et al.*, 2011), using the default algorithm (gradient descent using the L-BFGS method). The input for this algorithm were manually tagged Dutch reports (also known as a 'gold standard') created in the ARIADNE project (Vlachidis *et al.*, 2017), specifically selected to be a good sample of the corpus. In total, this training set consists of roughly 500,000 words, containing 11,000 tagged entities. Some issues with these documents are discussed later in this section.

These .docx files were tokenised and Part Of Speech (POS) tagged[3] using Frog (Van den Bosch *et al.*, 2007) and then converted to the FoLiA XML format (Van Gompel & Reynaert, 2013). Subsequently, the documents were converted to the format scikit-learn requires; a list of tokens including the token's POS and category (or concept) tag. At the moment, only three archaeological categories are used: artefact, time period and material, although more categories will be added in later versions. For each token, the following features were extracted for the word itself, as well as the word before and after the current one:

- Word in lowercase
- Word starts with uppercase character
- Word is all uppercase
- Word is all numbers
- Part of speech tag
- Word exists in materials wordlist
- Word exists in periods wordlist
- Word exists in artefacts wordlist
- Word is beginning or end of sentence

This is a fairly simple list, and is purely meant to provide a baseline result. As such, it was expected that the accuracy of the NER would not be very high.

To evaluate the results of the NER, a leave-one-out eight fold cross validation was done, meaning that the algorithm is run eight times, each time using seven of the documents as a training set, and using one document to test the model. It rotates through all eight possible combinations, and then calculates an average of the accuracy of the model. The total averaged accuracy (F1 score) is 56%, with the results for the different categories presented in table 5.2. As can be seen from this table, the average precision is fairly high at 71%, but the recall is much lower at only 48%.

[3]Tokenisation is the process of converting a character sequence (text) to individual tokens (words and punctuation). POS tagging is assigning a grammatical part of speech to each token, such as noun, verb, and so on.

	Precision	Recall	F1-Score
Artefact	0.76	0.40	0.53
Time Period	0.65	0.58	0.61
Material	0.72	0.46	0.56
Average	0.71	0.48	0.56

Table 5.2: Precision, recall and F1-scores for the 3 targeted entities, on a scale of 0 to 1.

When assessing the results of the NER, it was discovered that there are some issues with the gold standard documents which could affect the accuracy. It seems that some tagging decisions were made that mean that entities are expanded to the left or right. For example, wherever the word "before" or "after" occurs before a time period, these words are included in the tag, while ideally these shouldn't be included as they aren't part of the time period itself. If the NER then fails to classify these prefixes as the entity, the recall will be lower than the precision, which can also be seen in our results.

The artefact, time period and material wordlists that were taken from the *Archeologisch Basis Register* (ABR), a thesaurus for Dutch archaeology maintained by the RCE. It contains phrases that are written in such a way that they do not match the way we would find these phrases in natural language. For example, the entry for "*doorboorde bijl*" (perforated axe) is "*bijl, doorboord*" in the thesaurus, making it difficult to match the two. These two issues will be further discussed in section 5.5.

The code described in this section is available at https://doi.org/10.5281/zenodo.1238861.

5.3.2 Indexing & front end

For this version, 100 randomly selected reports from the DANS repository were selected to be indexed. For each page in these documents, the trained CRF model is used to extract the named entities. These are combined with the full text of the page and converted into a JSON structure, which can then be indexed directly by ElasticSearch (Gormley & Tong, 2015), an open source search engine running on a web server. ElasticSearch uses JSON over Hypertext Transfer Protocol (HTTP) to index and retrieve information, making it very easy to integrate with other systems. The other advantage of using ElasticSearch is that it includes a number of features by default that are very useful for these kinds of search systems, including a result ranking system.

To query the index, a front end has been developed. As a framework for the web application, the free and open source content management system Concrete5 was used (Concrete5, 2018).

To create a query, the user can use a query builder (Sorel, 2018) that allows for boolean AND / OR logic. You can specify exactly which entity you are looking for in each part of the query, or select a general full-text search. This allows for complex queries such as

```
artefact:scraper AND (period:neolithic
OR period:mesolithic) AND fulltext:burnt
```

which returns results on burnt scrapers from the neo- or mesolithic.

This query is then converted to a JSON format, so the ElasticSearch index can be queried using the ElasticSearch-PHP client (Tong, 2018), resulting in a list of matching results. It is useful to rank and sort these results by relevance, so the documents that are most likely to be relevant to a query are at the top of the list. To do this, ElasticSearch calculates a score for each result, which is based on the 'weight' of each query term that appears in that document. This weight is determined by three factors: term frequency, inverse document frequency and field length norm (ElasticSearch, 2018).

Once the results are displayed, the user can view a snippet of the text surrounding the keywords, preview the page of the report or go directly to the DANS repository to download the document. No PDFs are made available on the AGNES server to deal with the copyright of these files. A graphical representation of the full workflow of AGNES can be found in figure 5.2, which also displays the split between pre-processing of the documents on a high-performance cluster, and the indexing and querying that takes place on a standard web server.

5.4 User study

Part of this research includes a user study, to ensure the needs of the potential users are met. The focus group, as well as the methods and results of the first workshop, are detailed below.

5.4.1 Definition of target audience

To be able to make an effective search system, it is required to define the expected users of the system. As the main goal of this system is to make information available for research, the main expected user is a researcher working in Dutch

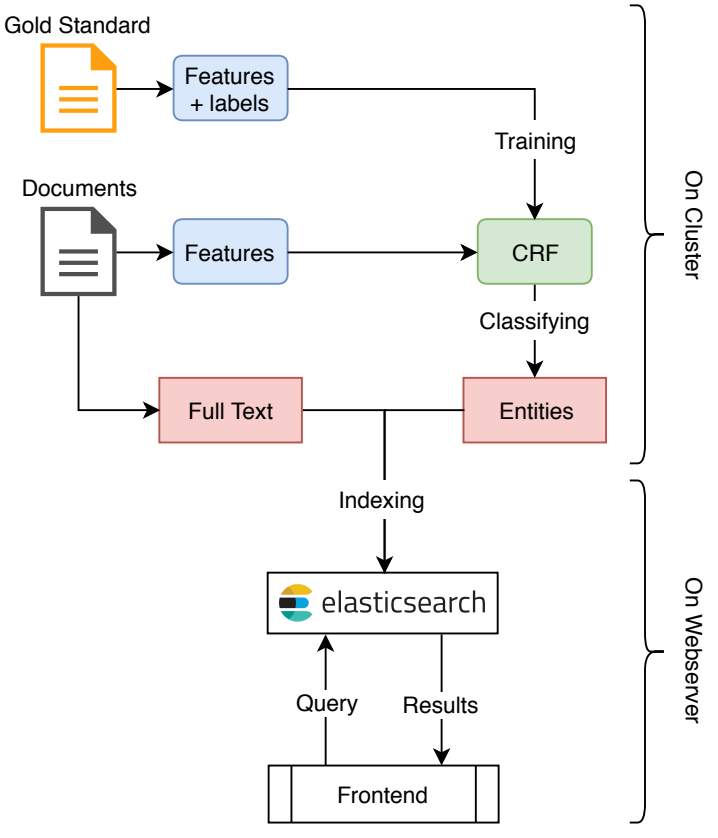


Figure 5.2: AGNES Workflow

archaeology. In the Netherlands, these researchers can be in a variety of organisational levels, including academia, commercial archaeology, regional/national government.

One of the main user groups expected to use this system are academics and people in higher education. However, this group is not homogeneous, as e.g. a professor will have much more in-depth knowledge and will already be aware of most of the literature and field reports related to their field, in stark comparison to e.g. a bachelor student or PhD who will still be exploring the literature and information available. Because of this difference in knowledge, these users will ask different questions of the data set and in different ways. However, regardless of their knowledge level it is expected that academic researchers will generally be asking thematic questions of the data set; questions about a particular time period, artifact type, context and/or location.

Another main user group is researchers in Dutch commercial archaeology. While this group will also be interested in the documents, it is likely that they will mainly want to use the system to find all information about a particular geographic area. This is because the main use of these reports for commercial archaeologists is to create desk assessments (*bureau onderzoeken*) and archaeological assessment/expectation maps (*archeologische verwachtingskaart*) about a specific area, generally because the area surrounds a potential building site. As some maps are also created by period, combined queries of place and time are also expected. There are three types of commercial archaeology, each are expected to have slightly different needs and requirements. These three types are inventarisation (investigating existing research), exploration or prospection (e.g. surveys and coring) and excavation (generally after the previous two types have been completed).

A third expected user group is municipal and regional (or provincial) archaeologists. Regarding their requirements, these will most probably fall in between academic and commercial archaeologists. While generally they will research a certain timespan in a particular area, it is likely that they will also want to research broader themes. However, generally they will be aware of all the available literature already, so perhaps a search system is less useful for this group.

Researchers at the RCE are a fourth user group, and will probably have similar needs to municipal archaeologists, except they are working on a country wide geographical scale. These researchers will commonly work on nation-wide synthesising research, combining the information from a large number of reports into a larger picture.

Outside of the archaeological sphere, it is possible that the system will also be used by historians researching more recent periods such as the Middle Ages, where there's an overlap between archaeology and history. It is expected these scholars will have similar requirements to archaeological academics.

Lastly, it is possible that this system might be used by amateur archaeologists, amateur historians, metal detectorists and other enthusiasts, for a variety of reasons.

5.4.2 Focus Group

In order to collect the requirements of archaeologists in the Netherlands, a voluntary focus group was set up. This group's function at the start of the project is to provide their needs and wishes for a system like this, while in further stages of the project they can provide feedback on the developed features. The size and make up of this group is fluid, and can be changed during the project to fit with

Group	Situation	Count
Academia	PhD Student	3
Academia	Assistant Professor	1
Academia	Lecturer	1
Commercial Archaeology	Excavation	1
Commercial Archaeology	Prospection	1
Government	Municipal	1
Government	National	1

Table 5.3: Overview of participants in focus group per category

the current goals and/or address issues of representativeness.

This group has been selected to be as representative as possible for the Dutch archaeological landscape, taking into account the target audience definition from section 5.4.1. The group consists of 5 academics, 2 commercial professionals and 2 archaeologists working on different levels in government. See table 5.3 for a more detailed break down of the participants.

No amateur researchers were selected for the focus group, mainly because they are not an intended user of the system, but also because their approaches to research are so wide ranging, it would be virtually impossible to assemble a representative group of people.

5.4.3 Prototype for discussion

From personal experience in commercial software development, as well as experiences from IR researchers in other fields (e.g. Verberne *et al.*, 2016), it seems that users in general, but the humanities specifically, find it difficult to express their requirements, oftentimes resulting in broad requirements that are too vague to interpret and implement. This can be further compounded by a lack of understanding of what is technically possible, leading to overly optimistic or very cautious expectations. We therefore first created a prototype with limited functionality (as discussed in section 5.3) as a starting point for discussions, in order to elicit feedback that is more detailed and can be implemented properly.

5.4.4 Workshops

The focus group will gather once a year during the project, for a total of 4 workshops. The initial workshop has been conducted, with the main aim of soliciting the requirements of the users. Later workshops will focus more on

assessing the system and its results. Minutes will be taken at each session to record the comments and feedback of the group, and these will be made public after anonymisation.

The first workshop started with an introduction to the problem, as well as some background information on IR and NER (see also section 5.3.1). The group was then asked what their current search behaviour is, and what problems they encounter, before being shown a prototype of the system (v0.2) and asked to provide feedback on both the functionality and the relevance of the results.

Finally, specific user requirements were discussed. A suggested list of features was provided to the participants, who then discussed amongst themselves in groups of 2 which features they would find most useful, on a scale of 0 to 3 with 0 being not useful or relevant at all, and 3 being very useful and high priority. The participants were also asked to think of features not currently on the list.

5.4.5 Results

From comments of the group, it was clear that the grey literature problem is very familiar to everyone present. Feedback on their current search behaviour showed that most people use the DANS search functionality[4] and find it not sufficient for their search needs, with most people having to manually search through individual documents to find information. Some participants, instead of using DANS, usually ask experts in the field to provide them with references, and the Archis[5] system is used to a lesser degree, again mainly because the search functionality is not sufficient. Some people explained that they create their own literature lists with keywords to be able to find materials previously accessed.

Initial feedback on the prototype indicates that the users find the returned results relevant to their queries, however much improvement is needed on the front end, as further discussed in the next paragraph.

The results from the feature elicitation were interesting; unanimously, everyone agreed that indexing by chapter and section would be more useful than indexing by page or document, and that this should be high priority. Another high priority feature across the board was to implement searching by drawing a polygon on a map as well as plotting results on a map, an indication that archaeologists have a strong need for geographical search. Another interesting result is that in general, everyone preferred to get many results with some irrelevant documents, than to get a smaller set of documents that are all relevant, with

[4] Found at http://easy.dans.knaw.nl

[5] Archis is a national database of archaeological sites in the Netherlands, maintained by the RCE, in Dutch. Located at https://archis.cultureelerfgoed.nl

Feature	Average
Search on map - plot results on map	2.78
Search on map - draw polygon	2.56
High recall over high precision	2.56
Search on map - morphology / expectation overlay	2.44
Index by chapter / section	2.33
Facets - time / artefact / place	2.22
Facets - research type	2.11
Personalise - alert if new docs in saved search	2.11
Related documents - by area	1.89
Facets - timeline	1.78
Personalise - save search	1.78
Related documents - by time	1.78
Ordering - by relevance	1.78
Personalise - mark documents as 'seen'	1.78
Ordering - by distance	1.67
Related documents - by artefact	1.67
Related documents - general	1.56
Plot terms in document	1.56
Ordering - by date added	1.11

Table 5.4: Features and average scores (0-3) across focus group (n = 9), sorted by average score, descending.

the risk of missing some documents. This means that the recall of the system is more important than the precision, which needs to be taken into account in assessing the results of the NER as well as the overall system assessment. For a full overview of the averaged result for each feature, please see table 5.4. In this table, facets mean the option for users to refine results by selecting categories, as often found on online shopping websites.

5.5 Future Work

The work discussed in this paper is the result of the first year of a 4 year project. Each year, a new version will be developed, tested, and assessed by the focus group.

The first issue that needs to be resolved is the gold standard. It seems that

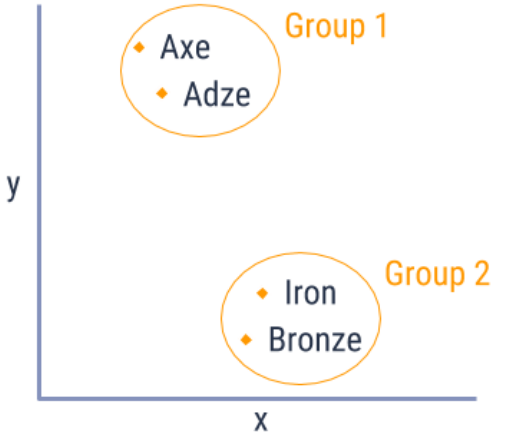


Figure 5.3: 2D representation of clustered word embeddings.

entities have been tagged sub-optimally for the NER task, and it is expected that improving the gold standard will increase the accuracy of the model. We intent to enlist the help of a group of archaeology students to re-tag these documents, and possibly tag new documents as well. We will have multiple people tag the same documents as well, so we can calculate the inter-rater agreement; a measure of how well this task can be accurately completed by humans, and ultimately, an upper limit for the accuracy of any machine learning model.

The other problem that will be addressed are the ABR wordlists. We are currently in discussion with the RCE, who manage these lists, to see if it's possible to add a new field for either the lemma of the word or to include multiple spellings of a word. After these two tasks have been completed, we will train the model again to see what difference these adjustments make.

Once that baseline has been established, we will integrate word embeddings as features, specifically word2vec (Mikolov *et al.*, 2013) and fasttext (Bojanowski *et al.*, 2016). These are both unsupervised machine learning techniques, that place words into a high-dimensional vector space based on their context in the text. The words can then be clustered using e.g. k-means clustering, with the idea that each cluster is a distinct 'type' of word. See figure 5.3 for a 2 dimensional (instead of high-dimensional) representation of this concept, where group 1 contains artefact types and group 2 contains materials. The advantage over using a word list is that related concepts not in the list, as well as misspellings of the concept, will also generally get assigned to the same cluster. Hopefully, this will increase the accuracy of the NER.

Regarding new features, according to the focus group the map functionality is the most required, including searching on a map and displaying results on a map. We are in the early stages of implementing this functionality and will hopefully present this in a future paper. Integration with common GIS systems is another avenue of research. Another feature with high priority is to index the documents by chapter or section, instead of by page as is currently the case.

To further evaluate the system, we will apply future versions to archaeological case studies. The plan is to find a specific archaeological information need, e.g. find all Iron Age cremations in the Netherlands and their geographical positions. We will then compare the results from AGNES with what experts currently know about this topic, and see if a significant increase in knowledge can be detected, probably by calculating the difference and overlap in numbers.

Currently, the system is focused on reports in Dutch, but as this problem is prevalent across the world, we will attempt to make the system multi-lingual, or at least provide ways of easily adapting the system to other languages.

Finally, one of the goals of the project is to expand the corpus from just the DANS documents to also including documents from the RCE and the KB, and creating a pipeline or API that allows for new documents added to these 3 repositories to be automatically added to the index.

5.6 Conclusions

From the user study, it is clear that a system such as AGNES is highly desirable for Dutch archaeology. The features assigned highest priority by the focus group are fairly uniform, which makes planning a road map of features straight forward. The first tentative feedback from the focus group is that results in AGNES are relevant to the queries, but more needs to be done to improve the functionality of the system.

From a technical viewpoint, the NER using CRF and a basic feature list resulted in an overall accuracy of 56%; fairly low, but partly explained by the problems with the gold standard and word lists. Fixing these problems, as well as introducing word embeddings as features, should increase the accuracy.

Overall, it seems that AGNES can address the problem of grey literature in Dutch archaeology, although this needs to be evaluated more thoroughly by comparing the results to expert knowledge. The systems developed should easily be adapted to other languages and areas as well. We are hopeful that AGNES will help archaeologists to answer their research questions more effectively and efficiently, leading to a more coherent narrative of the past.

6

Usability Evaluation

"To start, press any key. *Well where's the 'any' key!?*"
Homer Simpson, The Simpsons, s07e07, 'King-Size Homer'

Previously published as: Brandsen, A., Lambers, K., Verberne, S. and Wansleeben, M., 2021. Usability Evaluation for Online Professional Search in the Dutch Archaeology Domain. *ArXiv preprint.* ArXiv: 2103.04437

This paper presents AGNES, the first Information Retrieval (IR) system for archaeological grey literature, allowing full-text search of these long archaeological documents. This search system has a web interface that allows archaeology professionals and scholars to search through a collection of over 60,000 Dutch excavation reports, totalling 361 million words. We conducted a user study for the evaluation of AGNES's search interface, with a small but diverse user group. The evaluation was done by screen capturing and a think aloud protocol, combined with a user interface feedback questionnaire. The evaluation covered both controlled use (completion of a pre-defined task) as well as free use (completion of a freely chosen task). The free use allows us to study the information needs of archaeologists, as well as their interactions with the search system. We conclude that: (1) the information needs of archaeologists are typically recall-oriented, often requiring a list of items as answer; (2) the users prefer the use of free-text queries over metadata filters, confirming the value of a free-text search system; (3) the compilation of a diverse user group contributed to the collection of diverse issues as feedback for improving the system. We are currently refining AGNES's user interface and improving its precision for archaeological entities, so that AGNES will help archaeologists to answer their research questions more effectively and efficiently, leading to a more coherent narrative of the past.

6.1 Introduction

Archaeologists create large amounts of texts. Besides scholarly publications, another large source of documents are unpublished technical fieldwork reports. These reports are required to be produced by law whenever an excavation is performed (Council of Europe, 1992). They are generally not published in the traditional sense, and end up in various repositories, either in hard copy or digital format. The information in these reports is often needed, and described as 'crucial' and 'essential' by European archaeologists in a user study in the ARIADNE project (Selhofer & Geser, 2014). A recent report by Habermehl (2019) states that the accessibility, findability, and searchability of research output is essential for synthesising research.

In the Netherlands, the amount of reports created in the last twenty years is currently estimated at around 60,000, and is growing by approximately 4000 per year (Rijksdienst voor het Cultureel Erfgoed, 2019a). Most of these reports are categorised as 'grey literature' (Evans, 2015), and are likely to end up in a proverbial 'graveyard', unread and unknown, unless they are properly archived, indexed, and disclosed.

Easy access to this information is a major problem for the archaeological field, as there is currently no free-text search system available for archaeological reports. Searching through these documents and analysing them is a time consuming task when done by hand, and will often lack consistency (Brandsen *et al.*, 2019). A full-text index of archaeological documents, with a user interface, would allow researchers to locate (parts of) texts relevant to their research questions.

Some studies have investigated applications of Natural Language Processing (NLP) in heritage collections in general (Van Hooland *et al.*, 2015), but also from archaeological reports specifically, both in English (Vlachidis *et al.*, 2017; Amrani *et al.*, 2008; Byrne & Klein, 2010) and Dutch (Paijmans & Brandsen, 2010; Vlachidis *et al.*, 2017). However no IR system is currently available that allows full-text access to the documents held in Dutch archives (Habermehl, 2019). As a result, relevant and valuable information is not being utilised at the moment.

In this paper we present the AGNES search system that allows users to harness IR and NLP techniques to search for relevant archaeological literature. To ensure that the needs of the potential users and stakeholders are met, a focus group of archaeologists has been involved in the development and evaluation of the system. It is important that the usability of a system such as this is evaluated properly, as previous research indicates that there is a strong relationship between the usability and uptake of search systems (Dudek *et al.*, 2007).

Archaeology is an archive-heavy discipline in the digital humanities. Much of the archaeological data and finds reside in repositories. Yet to the best of our knowledge, no detailed research has been done into the information needs of archaeologists, nor of the usability of online tools for archaeology.

The following research questions are addressed in this paper:

1. What type of information needs do archaeologists have?
2. What are their query strategies?
3. How satisfied are the users with the usability of the AGNES system?

The contributions of this paper in comparison with previous work is that this is (to our knowledge) the first full text search system and the first usability evaluation of such as system in the archaeology domain. We also investigate archaeologists' information needs, their query strategies, and evaluate the usability of our search system for answering their information needs.

The structure of the rest of this paper is as follows: Section 6.2 provides an overview of related and prior work; Section 6.3 is a short introduction to the current version of our system; Section 6.4 presents the set up of the user study with the results presented in Section 6.5, followed by a discussion in Section 6.6. Section 6.7 describes our conclusions and future work.

6.2 Background

6.2.1 Access to archaeological data

In Dutch archaeology, a number of professional search systems are currently used to access excavation reports. The main two are EASY (DANS, 2019) maintained by DANS (Data Archiving and Networked Services) and Archis (Rijksdienst voor het Cultureel Erfgoed, 2019b) by the State Service for Heritage (RCE). The Dutch National Library (KB) also makes a limited amount of reports available via a standard library portal, but this system is used to a much lesser extent, due to the small amount of texts and the search interface not being geared towards archaeology. None of these systems support full text search, a highly desirable feature we have included in AGNES.

This kind of search through archaeological reports is a form of professional search, which implies that the developed search interface is used by a specific group of professionals, as opposed to web search engines designed for the general public (e.g. Google). Professional search often has very specific user needs that go beyond the needs of the general public.

In the ARIADNE project (Niccolucci & Richards, 2013), interviews and an online questionnaire were used to assess the current state of archaeological data access across Europe. Regarding problems encountered while searching for data, 'most comments related to the accessibility of data. Data appeared as difficult to find, not available online, and if online difficult to access' (Selhofer & Geser, 2014, p. 63). Also, 93% of respondents indicated that a portal enabling innovative and more powerful search mechanisms would be 'very helpful' or 'rather helpful' (Selhofer & Geser, 2014, p. 63).

More specifically for the Netherlands, Hessing *et al.* (2013) did an evaluation of the (then) current archaeological search systems in 2013. They found that the Archis system did allow for geographical search, but due to free text fields in the metadata forms, it is difficult to find the relevant items and make sure the results are exhaustive. A more recent report by Habermehl (2019) shows this is still the case: they state that the current search systems are not useful enough.

Since the research by Hessing *et al.* a new version of Archis has been released (3.0) which allows search across all metadata fields and the plotting of results on a map; something very important to archaeologists as all their research has a strong geographical component. It also allows searching in a specific area plotted on a map, but this cannot be combined with text search in the metadata, only faceted search.

The EASY system also offers text search, but again only on metadata. At the

time of Hessing *et al.*'s report, there was no mapping functionality, but due to this study this has since been added, and results can now be displayed on map. None of the systems offer full text search of the documents themselves, only of (combinations of) metadata. While metadata can be more specific and precise than full text (depending on who created the metadata), it is often incomplete and prone to errors, which makes a full text search highly desirable.

6.2.2 Feedback on existing systems from our user group

Research done early in the AGNES project confirms the findings above. In the initial user requirement solicitation workshop, we asked our user group about their current search behaviour. This showed that most researchers use the DANS search functionality and find it not sufficient for their search needs, with most people having to manually search through individual documents to find information. The Archis system is used to a lesser degree, again mainly because the search functionality is not sufficient and is experienced as being difficult to use. Specifically, not being able to search through all the text, and no proper integration of a map (including searching specific areas) were noted as currently missing. Multiple participants explained that they create their own literature lists with keywords to be able to find materials previously accessed (Brandsen *et al.*, 2019).

We also performed user requirement solicitation, and the user group had a clear need for geographic search, plotting results on a map, and faceted search (Brandsen *et al.*, 2019). These kinds of features are rarely needed in open-domain web search. Specifically the combination of these three features with full text search is highly desired, but not currently offered by any of the search portals we are aware of in the Netherlands and abroad.

6.2.3 Related work in usability studies

Usability studies assess the extent to which a system is easy and efficient to use, and how well users can reach their goals. In other words, usability is the overall usefulness of a product Rosenzweig (2015).

A common evaluation method in usability studies is to have users from the target audience use the software, and ask them to give feedback on the system. In usability studies for IR systems, the most used evaluation protocol is to provide the users with a number of information problems and ask them to solve these problems using the search system at hand. A questionnaire is used after the process to assess their satisfaction (Spink, 2002; Behnert & Lewandowski, 2017; Rico *et al.*, 2019).

Besides asking for feedback after the session, another commonly used method for getting feedback during the use of the software is the Think Aloud Protocol, as originally proposed by Lewis (1982), and more recently applied by e.g. Gerjets *et al.* (2011) and Hinostroza *et al.* (2018). Research by Van Waes (2000) shows that the combination of thinking aloud and recording the user behaviour is a useful observation method to collect data about the searching process, both on usability and cognitive aspects, which is confirmed by e.g. Verberne *et al.* (2016) and Kirkpatrick (2018).

In digital humanities studies, usability evaluation of tools and services is seen as a key part of the research (Bulatovic *et al.*, 2016), and is published and discussed in detail (e.g. Steiner *et al.*, 2014; Bartalesi *et al.*, 2016; Hu, 2018). In archaeology specifically, usability studies are less routinely performed (or at least not often published), and seem to be limited to the fields of virtual reality and digital museums (Karoulis *et al.*, 2006; Pescarin *et al.*, 2014). One recent study by Huurdeman & Piccoli (2020) investigates search interface features for 3D content in a digital heritage context.

Giving that there are key limitations to the currently available archaeological IR systems, and usability studies are rare in the archaeology domain, we think it is vital to research and publish the usability of the system we are creating.

6.3 AGNES

In the current project, we are developing AGNES, an IR system that makes Dutch archaeological grey literature more accessible and searchable. The AGNES index currently contains roughly 60,000 documents, totalling 361 million words. The PDF documents are stored in the DANS repository.

AGNES consists of three parts: software for recognising archaeological concepts (named entities), an indexing system that stores these entities and the full text, and a web application front end that can search through this index.

Named entities are terms that refer to important concepts from the real world (Marrero *et al.*, 2013). In the context of this project, the entities are archaeological concepts, mentioned in excavation reports. To give an example, in the following sentence the entities are bold: 'The **burial mound** yielded some **scrapers** from the **Neolithic**', a context, an artefact, and a time period, respectively. The example illustrates that entities can consist of multiple words. Two particular challenges of entity recognition are that a single term can refer to multiple entity types (e.g. 'Swifterbant' can be either a location, a time period, or a type of pottery), and that multiple terms can refer to the same entity (e.g. 'Neolithic'

and 'New Stone Age'). For more technical information on the NER process, as well as the methods used, see Brandsen *et al.* (2020).

In AGNES, archaeological entities are recognised and labelled during the indexing of the documents. In version 0.3 of AGNES, all 60,000 reports from the DANS repository were indexed. For each page in these documents, the named entities are extracted and combined with the full text of the page and indexed directly by ElasticSearch (Gormley & Tong, 2015). We are currently indexing at the page and document level, but in future we will index at the chapter/section level. This is more suitable to most information needs, as researchers will want to find e.g. all sections that mention 'axe' and 'neolithic', even if they are mentioned on different pages. This was also seen in the user study, as detailed in the next section.

We developed a front end to query the index. The searcher can use a query builder (Sorel, 2018) that allows for boolean AND / OR logic. They can specify exactly which entity they are looking for in each part of the query, or select a general full-text search. This visual interface allows for the creation of queries such as the following pseudo-query[1]:

```
artefact:axe AND (period:neolithic OR period:mesolithic)
AND fulltext:burnt
```

which returns results on axes from the Neolithic or Mesolithic where the word 'burnt' is also mentioned on the same page. It is also possible to refine the query by using facets (filtering for specific metadata values, such as time period or document type) or by drawing a polygon on a map, performing a geographical search.

The query is then sent to ElasticSearch, which returns a list of matching results. Once the results are displayed, the user can view a snippet of the text surrounding the keywords, preview the page of the report or go directly to the DANS repository to download the document. No PDFs are made available on the AGNES server in order to respect the copyright of these files.

See Fig. 6.1 for a screenshot of the AGNES User Interface (UI). This version is the one that has been evaluated in the current study, and is available at `http://agnessearch.nl/v03`.[2]

[1] Note that this is not what the user types in, but an easy to read representation of the query that's generated by the system

[2] Please note, free registration is needed to access the system.

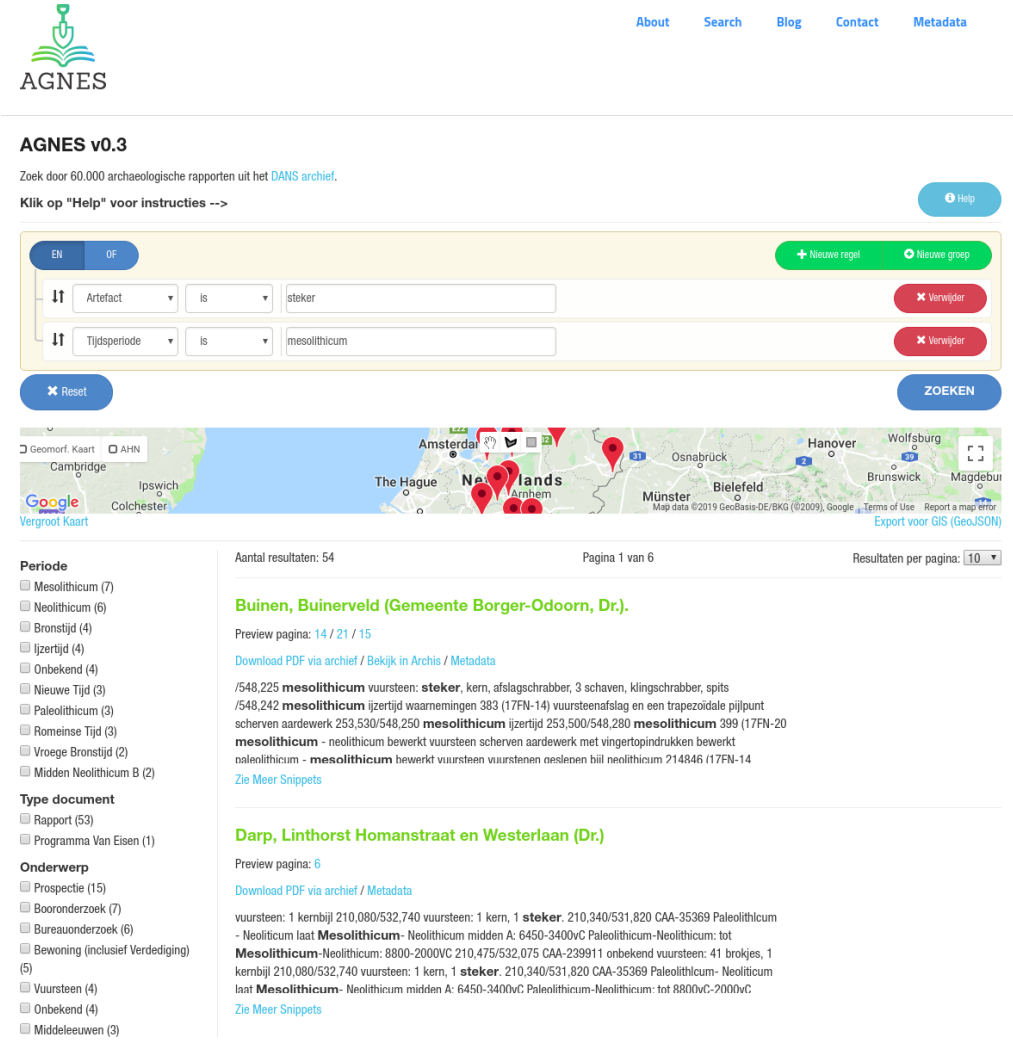


Figure 6.1: Screenshot of AGNES version 0.3. Pictured here is a query for 'artefact:axe AND (period:neolithic OR period:mesolithic) AND full-text:burnt', with the results on a map and in a list underneath (with snippets). On the left we can see the facets, used to filter results on period, type of document, and subject.

6.4 User Study Setup

A focus group is a small but diverse group of people who's reactions are studied to generalise to a larger population. Focus groups are often used for data collection, and have been studied and described in detail in literature (Thomsett-Scott, 2006; Barbour, 2018). Specifically, they are useful for gathering qualitative data quickly and cheaply, as well as gathering data on attitudes, values, and opinions (Cohen *et al.*, 2002). This very much aligns with the purpose of this study; to collect users' opinions on the currently available systems, their requirements for a new system and their assessments of developed features.

6.4.1 Workshops in the AGNES project

For the user study, we followed a user-centred approach, consisting of pre-assessment (user requirement solicitation), mid-term evaluation (feedback on early system versions), and post-assessment (user trial). A user-centred evaluation approach focuses on examining the behaviour and preferences of users, and their interaction with the system. The main purposes of this type of evaluation are to find problems and assessing the quality of a system (Dejong & Schellens, 1997), exactly what we set out to do.

Four workshops are held during the AGNES project, once per year. The first workshop had the aim of eliciting the requirements of the users, and the second workshop aimed to evaluate the user interface. Later workshops will focus more on assessing the quality of the results. Minutes are taken at each session to record the comments and feedback of the group, and these will be made public after anonymisation.

6.4.2 Compilation of the focus group

We compiled a focus group of archaeologists. The size and compilation of this group is fluid, and can be changed during the project to fit with the current goals and/or address issues of representativeness.

This group has been selected in such a way that it includes every category of the target audience as defined in Brandsen *et al.* (2019). At the current stage of research, the group consists of six academics, two commercial professionals, and two archaeologists working on different levels in government. See Table 6.1 for a more detailed break down of the participants.

Regarding the size of the focus group, Nielsen & Landauer (1993) show that the number of additional usability problems encountered when adding more users

Category	Situation	Count
Academia	PhD Student	4
Academia	Assistant Professor	1
Academia	Lecturer	1
Commercial Archaeology	Excavation	1
Commercial Archaeology	Prospection	1
Government	Municipal	1
Government	National	1

Table 6.1: Overview of participants in usability evaluation per category

quickly decreases beyond five users. Thus, the current size of the focus group should be more adequate in this regard.

6.4.3 Design and procedure

The evaluations were performed on a one-to-one basis. Users only got an introduction to what the system was, but no specific instructions on how to use the system. We placed the users in a quiet office with the system running on a laptop, and asked each participant to use the system to perform three predefined tasks, as well as at least three of their own self-defined information needs. The predefined tasks are the following (translated from Dutch):

1. Where in the Netherlands can we find globular jars (*kogelpotten*) in fire pits?
2. Find all literature relating to Neolithic scrapers found south of the river Meuse.
3. Find all Roman pottery found in a settlement.

The first task is intended to introduce the user to the query builder, as well as viewing the results on the map, as this is needed to answer the question. The aim of the second task is to use the geographical search, using the map to draw a polygon around the area. The last task is aimed to force the user to use the facets, by selecting the 'settlement' facet.

To better understand the user behaviour, we asked the participants to use the Think Aloud Protocol, as introduced in Section 6.2.3. Specifically, we asked the participants to say what they think, see, expect, do, feel, and motivate their actions. At the end of the session we also asked the user a number of questions which can be found below.

1. Which elements of the system worked well?
2. Which elements of the system did not work well?
3. Was anything unclear?
4. Is there any functionality that is missing, in your view?
5. What is your opinion on the facets?
6. What is your opinion on the map functionality?
7. Is there anything else you would like to add to this evaluation?

We did not include any quantitative evaluation questions[3] in the questionnaire, as satisfaction with a system was shown to be directly proportional to the quality of the results (Verberne *et al.*, 2016), and as such is not a good measure for usability.

To record the sessions, we used screencasts[4] to record the user behaviour on the screen, together with statistics on the queries recorded by the system itself. We also used sound recordings to capture the thoughts of the participants, together with notes made by the researcher (first author) sitting next to the user. A table containing all queries with related statistics is available in the online data repository for this study[5].

The answers to the questions, as well as the user's thoughts during searching, were transcribed and translated to English, and the resulting qualitative data were processed using grounded theory techniques (Charmaz, 2006), which entails coding statements and grouping those codes into categories, to allow for a quantitative approach on the data.

We also analysed the screencasts afterwards, and recorded all the query (re-)formulations in a pseudo-query format, together with the time spent on each query and how many results were returned.

6.5 Analysis and Results

To address our research questions, we performed both quantitative and qualitative analyses of the results of the usability evaluations. These are further detailed in the following subsections.

A total of 148 queries were observed and recorded during the evaluation sessions, for a total of sixty-four information needs, making for an average of 2.3 queries per task. A query is defined here as a combination of search terms entered

[3] E.g. How would you rate this system on a scale from one to ten?

[4] Using the Loom Chromium plugin (https://www.loom.com/)

[5] https://doi.org/10.5281/zenodo.4064076, also contains a list of all usability issues and a list of user needs mentioned in later sections.

into the system, an information need as a defined question the researcher wants to answer. The minimum number of query elements is one, as expected, and the maximum is ten, with an average of 2.4. Here, an element of a query is one AND/OR statement, so for example the pseudo query `[artefact : scraper] AND [period : neolithic]` contains two elements.

6.5.1 Information Needs

Based on work on question taxonomies by Voorhees (2001) and Hermjakob *et al.* (2000), we can distinguish three main types of questions; (1) closed questions with a yes or no answer, (2) factoid questions where more than a yes/no answer is required, and (3) list questions, where a list of results is the intended end goal. Other research in the humanities such as Verberne *et al.* (2016) suggest that humanities scholars generally have a mix of all three, with a preference for factoid questions.

In our Think Aloud sessions, we asked the users to also state the question they wanted to answer, and noted this down. We noticed that almost all the questions asked by the users are list questions, e.g. the three tasks mentioned in Table 6.2.

Find all amber from the Middle Neolithic

Query	**Type**
[material:amber] AND [period:middle neolithic]	
[material:amber] AND [period:neolithic]	Generalisation
[material:amber]	Generalisation

Find all beakers from graves in the late Neolithic

Query	**Type**
[period:late neolithic] AND [other:grave] AND [artefact:beaker]	
[period:late neolithic] AND [other:grave] AND [artefact:beaker] AND [filter:prehistory]	Specification
[period:late neolithic] AND [other:grave]	Generalisation
[period:late neolithic] AND [other:grave] AND [filter:neolitic]	Specification

Find all coprolites from the Swifterbant period

Query	**Type**
[period:swifterbant] AND [artefact:coprolite]	
[other:swifterbant] AND [artefact:coprolite]	Parallel / reformulation

Table 6.2: Three examples of user generated tasks and their associated queries and query reformulations (translated from Dutch).

This intuitively makes sense for archaeologists, as research most often entails making a list of all known occurrences of a particular topic and then performing some sort of analysis on this list. In our user requirements study, the users also indicated a preference for high recall over high precision, they much prefer getting all the relevant results with some noise, than to miss some results and have only relevant results Brandsen *et al.* (2019).

6.5.2 Query Strategies and Effectiveness

We analysed the query reformulation strategies in this data, the process of altering a query to be narrower (specification, making the query longer), broader (generalisation, making the query shorter) or replacing one or more terms by other terms without making the query broader or narrower (parallel movement / reformulation) Rieh (2006). Interestingly, there is no trend to be found across all users between specification and generalisation, with both types of query reformulations occurring almost equally (twenty-five and twenty-four times, respectively). We do note that some users have a tendency to start broad and narrow down, while others do the opposite, but this seems to be a personal preference and not a preference for particular user categories. The full data is available via Zenodo[6], and there are three examples in Table 6.2.

While the users let us know in the feedback that they liked the faceted search (see Section 6.5.3 below), when we look at the queries we see that they don't use the facets very often. Out of 148 queries, only 23 include the use of facets (15.5%).

If we look at the use of Boolean expressions, only a small number of queries (9.5%) use the advanced features of the query builder, i.e. have an OR or group operator. It seems that archaeologists are either not trained to think in Boolean expressions, or simply do not have information needs that require them, which is in contrast with other professional search groups (Russell-Rose *et al.*, 2018). This in turn leads to the conclusion that the query builder might be overkill for such a system, seeing as more than 90% of the queries could have just been typed in a text field.

Query Effectiveness

It is difficult to directly measure the effectiveness of user queries, partly because the users themselves are not always sure that they have found the complete answer to the question. As a proxy for query effectiveness, we therefore

[6] https://doi.org/10.5281/zenodo.4064076

make a comparison of the user-formulated query to a reference query of which we are sure that it returns the complete set of relevant items. The reference queries consist of query terms combined with metadata filters (facets). For example, for the task 'Find all Roman pottery found in a settlement' we formulated the query `[artefact : pottery] AND [period : roman] AND [facet - site type : settlement]`. We then counted how often the users succeed in formulating the reference query. Although the users might have found the answer with a different query, this gives us an approximation of the session success.

All the users managed to formulate the same query for task 1 and 2, in 1.6 and 1.2 query reformulations on average, respectively. This means they ended up using the interface in the same way as we intended. Task 3 was more difficult, as only two out of ten participants executed a matching query. The difference in query stemmed from the confusion around the facets; we intended for the users to use the facets to filter on 'settlement', but six users used 'settlement' in the actual query instead and opted not to use the facets. While the facets are more exact and also handle synonyms, entering 'settlement' in the query still produced relevant results. So even though the query was not exactly the same as the intended query, we would argue that this task was still completed by the entire user group.

For the self formulated information needs, we could not determine the query effectiveness as we don't have any reference queries. Instead, we asked the users to only stop editing the query when they were satisfied with the results, and for only a couple of information needs the user indicated they were not satisfied. However, this resulted from inaccurate Named Entity Recognition (NER) and/or documents they expected to be in the system not being present, not from the interface being difficult to use. As a quantitative approach is not possible here, we further evaluate the system using a qualitative approach in the next section.

6.5.3 Evaluation and User Satisfaction

Comments per User Group

If we look at the number of usability issues raised per user category (commercial, academic or government), we find that roughly 58% of them (eighteen out of thirty-one) are raised by one user category only. This indicates that it is important to create a user group that is as diverse as possible, being representative of the target population, as otherwise certain issues will simply not be found.

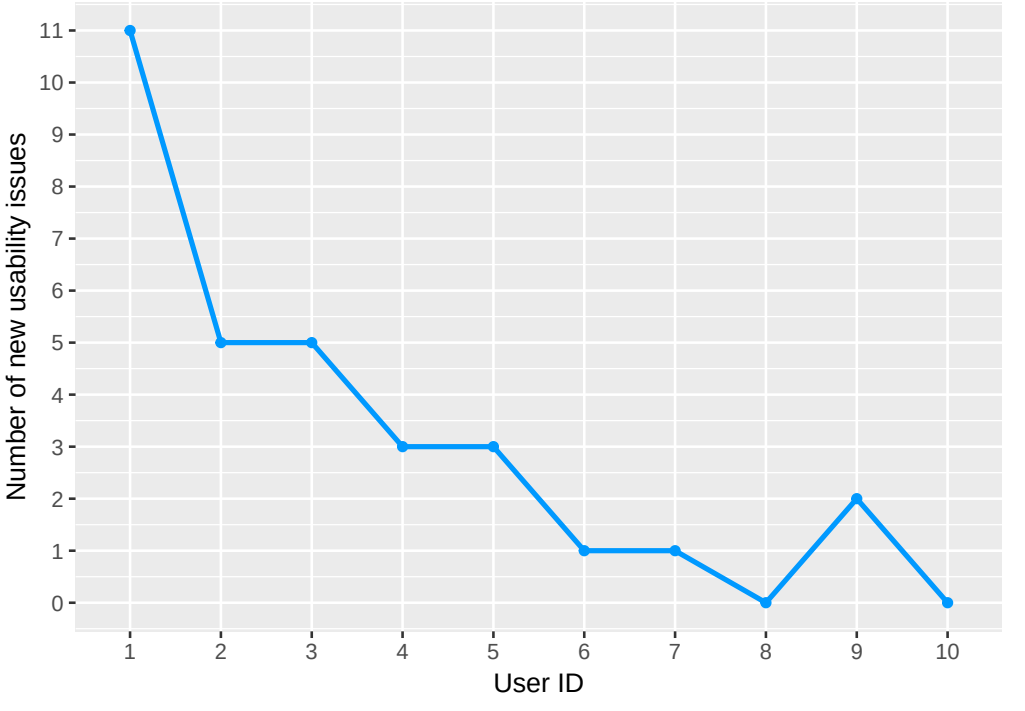


Figure 6.2: Line plot showing the number of new issues raised for each user

Cumulative New Issues per User

The users mentioned a total of sixty-eight usability issues, averaging 6.8 per user. Where two or more users mentioned the same issue, we grouped it, which leads to a total of thirty-one unique issues. Fig. 6.2 shows the number of new usability issues found for each user that is added to the evaluation. We can see that after the fifth user evaluation, new users tend to not identify many new issues, confirming prior work on usability studies. The exception is user 9 with two new issues, who is the only commercial excavation archaeologist in our user group. This again underlines the necessity for a diverse user group.

Positive and Negative elements

From the answers to the questionnaire after each session, we got the impression that overall, the users find the system fairly easy to use and clear. The map functionality is mentioned by everyone, and mentioned often, something that

was expected by the results of the user requirement solicitation. In Fig. 6.3 we have plotted a word cloud of all feedback, after translation from Dutch to English. We lowercased all text, removed punctuation, removed stopwords (NLTK list), and then plotted only words which occurred more than once.

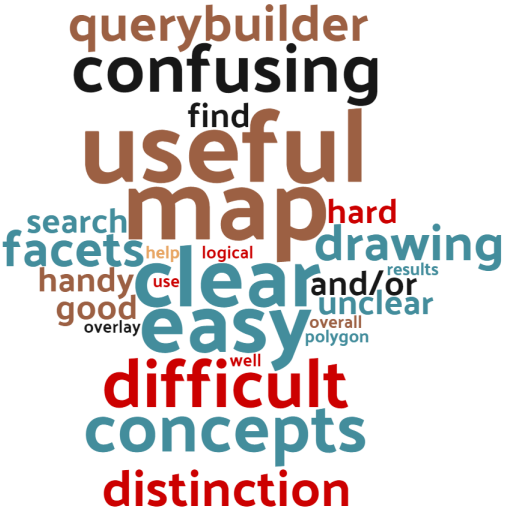


Figure 6.3: Word cloud of all feedback given, both positive and negative (translated from Dutch to English, 'ahn' is the height model of the Netherlands)

We can see that the words 'clear' and 'easy' are often used, as well as the map, confirming the impression we got from the sessions. Also we see the words 'difficult' and 'unclear' used often, these are more in relation to negative aspects of the system.

In table 6.3 we show the most frequent words for the positive and negative feedback fields, respectively, where we have removed all stop words, verbs and opinion-bearing words (such as 'clear', 'hard'). Again we only include words mentioned more than once.

On the negative side we can see that choosing which concept to search for, intuitiveness, the 'help' button, the facets, and the AND/OR toggle buttons are elements that are commonly experienced as negative at the moment. These issues and features will be dealt with in the next version of the system. We also see that the map, query builder, facets, and overall usability are often experienced as positive.

One of the other observations made during the evaluation is that none of the

Positive		*Negative*	
Freq.	**Word**	**Freq.**	**Word**
9	map	6	concepts
4	querybuilder	5	facets
3	facets	4	and/or
3	usability	3	intuitiveness
2	drawing	2	help
2	overall		

Table 6.3: Feedback split into positive and negative, with for each word how often it occurs in that context. Words only mentioned once are not included.

users use, or even see, the 'Help' button, which we did expect them to. This led to some preventable confusion about the system, as some questions the users had were actually explained in the help section. As a solution, we will include in-context help in the next version; pop ups that appear when hovering on certain elements to further explain the system.

Time Spent per Query

As mentioned before, we observed sixty-four research questions, with a total of 148 queries, so 2.3 query reformulations on average. For each initial query and query reformulation, we recorded the time taken to (re-)formulate the query and the number of elements in the query, among other information. We use the time per element instead of time per query to account for the difference in length of query between users, this way we can easily compare them. In Fig. 6.4 we plotted the time per element against the succession of queries attempted by a user. Here we see a clear downward trend (average between users shown in black). As the users had to do at least three of their own tasks, but could continue with more if they wanted, means that we have less data between query 6 and 9. However the trend is already clear between query 1 and 6.

This trend means that as the users perform more queries, the time taken per query element decreases rapidly, indicating that the system is easy to learn.

6.6 Discussion

Gibbs & Owens (2012) talk about how the typical humanities user is often neglected in the design of tools, and how tools' visibility can be increased by good

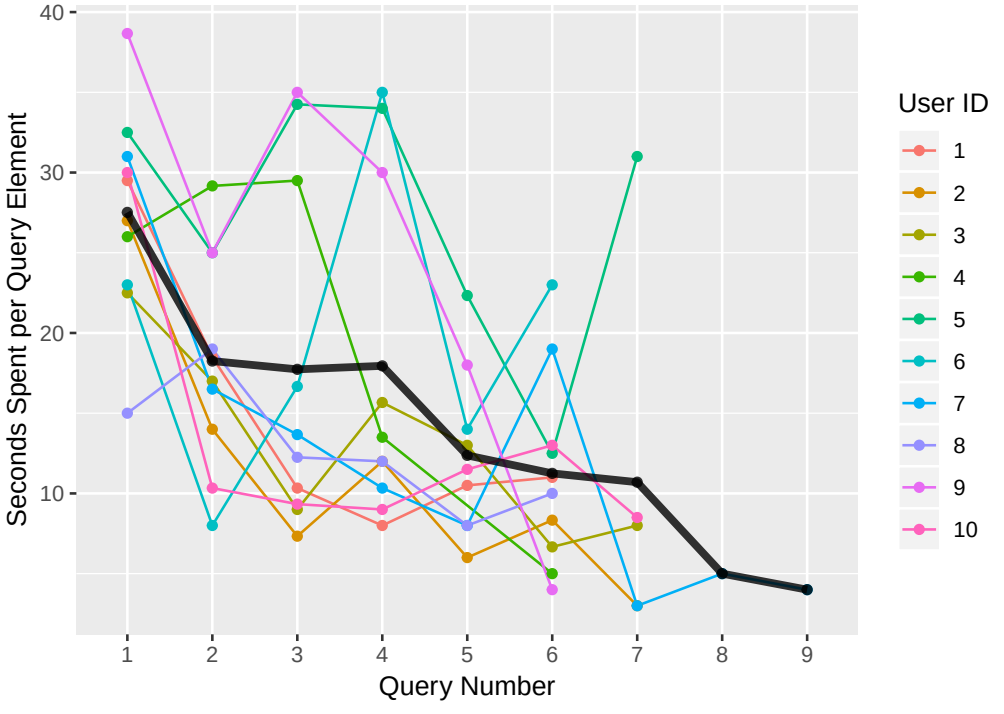


Figure 6.4: Line plot showing for each user, how much time they spent formulating one element of a query, for each new query they attempted. The black line is the average over all the users.

attention to the usability. More recent work by Bulatovic *et al.* (2016) agrees, and states that digital humanities tools often suffer from poor user experiences, mainly caused by the lack of resources spent on usability research.

As mentioned in the introduction, it is important to evaluate usability, as previous research indicates that usability and uptake of search systems is strongly correlated (Dudek *et al.*, 2007). At the same time, it is difficult to evaluate usability independently from the quality of the results, as users will perceive a system as not being usable if the results they get are of low quality. In this work, we found that it is important to brief the test users before hand to manage their expectations, and design the tasks and questionnaire to specifically target usability features that can be evaluated whether the results are good or bad.

Bulatovic *et al.* (2016) also mention that early iterative cycles of testing should be implemented in these kinds of projects, to avoid common usability problems.

This is what we are doing in this project, and we hope this will see more uptake in the digital humanities, as it seems usability evaluation is something done at the end of most projects as an afterthought, if done at all. In 2012, Gibbs & Owens called for a shift to user-centred design techniques, and luckily we do see that most of the more recent studies take this approach (e.g. Hinrichs *et al.*, 2015; Van Zundert, 2016; Esmailpour *et al.*, 2019).

We think that for tools to be used by humanities scholars, the user interface needs significant investment in the design that needs to be integrated into the project budget and timeline. At a more broader level, we agree with Koolen *et al.* (2018, p. 20) that digital tools 'always require critical reflection on how they mediate between researchers and their materials of study', something we will investigate further in future research.

Specifically for the archaeology domain, usability is evaluated and published even less than in the digital humanities as a whole. Seeing as there are key limitations to the currently available systems, and usability studies are rare in the archaeology domain, we think it is vital to research and publish the usability of the system we are creating. More generally, we believe that the research presented here is not only of value to the system itself, but also to other researchers building online tools; perhaps the findings are not generalisable to other applications due to the small sample size, but can at the very least serve as inspiration. When more archaeologists publish their usability studies, we can together make more useful, meaningful tools.

6.7 Conclusions

In this paper, we have investigated how Dutch archaeologists prefer to use online search, what features they deem positive and negative, how well our UI performs, and have assessed which analyses are useful for usability studies of this and similar systems. Here we will answer our research questions.

1. What type of information needs do archaeologists have? From previous studies and our own user requirement solicitation study, we see that Dutch archaeologists are mainly interested in geographic search, plotting results on a map, and faceted search.

We see that Dutch archaeologists have a clear preference for high-recall list-type questions when doing research. A difference between archaeologists and other professional search domains (Russell-Rose *et al.*, 2018) is the lack of preference for Boolean expressions, our user group barely used them, nor told us they

wanted them.

2. What query strategies do archaeologists use? We did not find any preference on query reformulation: specification and generalisation occurred roughly equally. We also noted that all users were able to create the reference queries for the predefined tasks, indicating the UI is being used as we intended. Regarding facets, we see that while users report these as being helpful, they do not use them very often, occurring in only 15% of the queries.

3. How satisfied are the users with the usability of the AGNES system? By analysing the feedback during the system evaluation, we found that users found the UI easy to use, clear, and useful. They specifically found the map features and query builder to be good features of the system. When we visualised the feedback, we see that the query builder, map features, facets, and snippets are experienced as positive. Some negative features include the help button, uncertainty about the mechanism behind the facets and concepts in the query builder, and the overall intuitiveness.

We see that the time taken per query element decreases fairly rapidly when users perform more queries, which indicates the system is easy to learn.

Using a relatively small user group of ten participants was expected to be enough to find and address usability issues, and this proved to be correct; we found that as the number of users increased beyond five, the number of issues highlighted dropped rapidly.

The importance of a diverse user group has been shown, as we found that roughly two thirds of issues were only raised by one of the user groups. Interestingly, if this is combined with the previous conclusion, this might mean that the ideal size of a user group might be five users per user category, instead of five in total.

In conclusion, it seems that AGNES can address the problem of accessing grey literature in Dutch archaeology, although this needs to be evaluated more thoroughly by comparing the results found with the use of AGNES to the prior knowledge of the topic, i.e. lists of occurrences of certain types of artefacts archaeologists have compiled manually. We are hopeful that AGNES will help archaeologists to answer their research questions more effectively and efficiently, leading to a more coherent narrative of the past.

6.7.1 Future Work

The work discussed in this paper is the result of the second year of a four year project. Each year, a new version of AGNES is developed, tested, and evaluated by the focus group. The first two workshops dealt with user requirement solicitation and evaluation of the interface, for the next workshop we will evaluate the quality of the results returned.

Further work is needed to refine the User Interface, all the issues and suggestions raised by the user group will be dealt with in the next version of the system. This should make it easier to focus purely on evaluating the results in the next workshop.

At the moment, we only evaluated the system using ten users. We believe that a quantitative study using statistics generated by the system could be useful in finding usability issues, as well as seeing patterns in usage. To this end we will make the next version of the system public and invite a large group of archaeologists to use the system. This should give us a much larger user group, although this is a more superficial analysis and loses some of the depth of evaluations done on the current group with the one-on-one approach.

Some recent work by Russell-Rose & Shokraneh (2020) suggests that traditional query builders like the one used in this project might not be ideal, and a more visual layout of a query provides a more direct mapping to the underlying semantics, and makes it more transparent. This is something we'd like to experiment with in future versions of AGNES, especially since our user group didn't seem to need the query builder for boolean expressions.

7

Using BERT for Named Entity Recognition

"*Look at all the exciting new discoveries, look at all the knowledge here.*"
Bert, Sesame street ep. 1621, 'Bert and Ernie in a Pyramid'

Accepted for publication as: Brandsen, A., Verberne, S., Lambers, K., & Wansleeben, M., 2021. Can BERT Dig It? – Named Entity Recognition for Information Retrieval in the Archaeology Domain. *Journal on Computing and Cultural Heritage*

The amount of archaeological literature is growing rapidly. Until recently, these data were only accessible through metadata search. We implemented a text retrieval engine for a large archaeological text collection (∼ 658 Million words). In archaeological IR, domain-specific entities such as locations, time periods, and artefacts, play a central role. This motivated the development of a Named Entity Recognition (NER) model to annotate the full collection with archaeological named entities. In this paper, we present ArcheoBERTje, a BERT model pre-trained on Dutch archaeological texts. We compare the model's quality and output on a Named Entity Recognition task to a generic multilingual model and a generic Dutch model. We also investigate ensemble methods for combining multiple BERT models, and combining the best BERT model with a domain thesaurus using Conditional Random Fields (CRF). We find that ArcheoBERTje outperforms both the multilingual and Dutch model significantly with a smaller standard deviation between runs, reaching an average F1 score of 0.735. The model also outperforms ensemble methods combining the three models. Combining ArcheoBERTje predictions and explicit domain knowledge from the thesaurus did not increase the F1 score. We quantitatively and qualitatively analyse the differences between the vocabulary and output of the BERT models on the full collection and provide some valuable insights in the effect of fine-tuning for specific domains. Our results indicate that for a highly specific text domain such as archaeology, further pre-training on domain-specific data increases the model's quality on NER by a much larger margin than shown for other domains in the literature, and that domain-specific pre-training makes the addition of domain knowledge from a thesaurus unnecessary.

7.1 Introduction

Like in other domains, archaeologists produce large amounts of text about their research. Besides research leading to scholarly output, commercial archaeology companies survey and excavate areas before developers build there and might destroy the archaeological remains. For each of these investigations, a report is written and stored in a repository. In the Netherlands, more than 4,000 of these documents are produced every year (Rijksdienst voor het Cultureel Erfgoed, 2019a), with the total currently estimated at 70,000. These documents are used to some extent by both academic and commercial archaeologists to do further research.

Currently, this so-called 'grey literature' is underused, as the available search tools only offer metadata search, making searching through these reports time

consuming and inaccurate (Habermehl, 2019). A strong need for better search tools has been well documented in prior work (Van den Dries, 2016; Habermehl, 2019; Richards *et al.*, 2015; Brandsen *et al.*, 2019), as the information in the full text of the reports can be of great value. Archaeological information needs are often recall-oriented list questions, consisting of a combination of What, Where and When aspects, e.g. "Find all cremations from the Early Middle Ages in the Netherlands" (Brandsen *et al.*, 2019). These are difficult to satisfy as the previously available search interfaces only offer search on the title, a short description, and sometimes information about the dating and type of archaeology encountered (stored in metadata fields), but the latter two are often missing or incorrectly assigned. Archaeologists want to search in more detail, and are often interested in the so-called 'by-catch': a single find unlike the rest of an excavation. For example, on an excavation yielding mainly Bronze Age material, a single Medieval cremation most likely will not be mentioned in the metadata, making it difficult to retrieve without manually searching through all the PDFs.

To address these needs, we implemented a text retrieval engine for a large collection of archaeological reports in the Netherlands. The retrieval collection contains an export (obtained in 2017) of every PDF file in the DANS repository[1] with the label 'Archaeology'. This totals over 60 thousand documents and 658 Million tokens.

A full text search would alleviate a lot of the current challenges archaeologists face in their search of information, but as Habermehl (2019) mentions, even in the relatively structured metadata, both synonymy and polysemy are a challenge, which is likely to be even worse in the free text in the body of the documents.

- Synonymy is a challenge because it leads to a lower recall: as there are numerous ways to write concepts relevant to archaeology, a search for one of these variants will not return the others. Specifically time periods have many synonyms. For example, the 'Early Middle Ages' can also be expressed as the 'Early Medieval Period', or 'Merovingian Period', or as dates that fall within the period, such as '600 CE' and '1400 BP'.
- Polysemy on the other hand, causes precision to be lower because one word can have multiple meanings, causing irrelevant meanings to appear in the search results. A good archaeological example is *Swifterbant*, which is a location, a type of pottery, an excavation event, and a time period. This problem of polysemy causes query ambiguity, as a full-text search engine does not know which meaning the user is looking for in their query, and then also does not know which meaning to retrieve from the corpus.

[1] https://easy.dans.knaw.nl/ui/home

Automatic query expansion is often used to combat problems with synonymy, either by using thesauri or embeddings to add synonyms and similar terms to a query and increase the recall (Soto *et al.*, 2008; Carpineto & Romano, 2012). Unfortunately in the case of time periods, this is difficult, as some time periods span thousands or millions of years, and adding each year with multiple variations (AD, BC, CE, BCE, BP) would result in an extremely large query. Polysemy is usually addressed in web search engines by diversifying search results or query suggestions (Capannini *et al.*, 2011; Song *et al.*, 2011): for each possible meaning of the ambiguous query, at least one relevant result is shown. For our specific domain, this is not possible because we do not have the large amount of user traffic that generic web search engines have, to be able to learn the different relevant results for any query term.

Instead, we opt for Named Entity Recognition (NER) to automatically detect archaeological entities in the corpus, and then allow archaeologists to find these using an entity-based query interface, combined with a full text search. The entity search attempts to solve the polysemy problem, as the user specifies – in the structured query interface – which meaning of a word they are looking for, e.g. the Location[2] *Swifterbant*. In this case, only documents where the Location entity *Swifterbant* has been detected will be returned. Although this helps the user specify their query, it also means that entities that have not been correctly identified will not be returned; in other words, errors in the NER output might propagate to retrieval errors. Therefore, to give the user freedom in the query form that best suits their information need, we combine entity search with full-text search.

We have previously published a prototype of our search engine online. The search engine uses ElasticSearch (Gormley & Tong, 2015) to index the full text, and in the prototype, entities were automatically labelled with a baseline NER model based on Conditional Random Fields (CRF). The resulting entity-based full-text search was experienced as positive by a focus group of archaeologists (Brandsen *et al.*, 2019).

However, the baseline NER model offers room for improvement. As prior work on archaeological NER indicated, CRF with common token-, context- and thesaurus-based features leads to relatively low F1 scores, around 0.50 to 0.70 (Brandsen *et al.*, 2019, 2020). In the last couple of years, transfer learning, and specifically BERT models (Devlin *et al.*, 2019), have been used successfully to get state-of-the-art (SotA) results for NER. On general domain benchmarks the SotA methods yield impressive F1 scores of up to 0.943 (Yamada *et al.*, 2020).

[2]Entity types will be capitalised from here on for clarity.

However, in other domains and languages the performance of NER systems is generally lower (Lee *et al.*, 2019).

BERT has not been applied to the archaeology domain yet in any language, and we believe this domain could benefit from context-dependent embeddings due to the above-mentioned polysemy. Two generic Dutch BERT models have been released (De Vries *et al.*, 2019; Delobelle *et al.*, 2020) which can help our research. Prior work on language- and domain-specific BERT models reports mixed results on the effect of pre-training on language- and domain-specific data (see Section 7.2.4). In this paper, we investigate whether BERT can improve NER in the Dutch archaeology domain, and to what extent further pre-training on domain-specific texts improves the quality of the model. We compare Google's multilingual model (Devlin *et al.*, 2019), the Dutch BERTje model (De Vries *et al.*, 2019), and our own ArcheoBERTje model that we further pre-trained on Dutch excavation reports. We do not compare the Dutch RobBERT model as it has a different training procedure and longer training times. We analyse the differences between the three models and we experiment with ensembles to combine multiple models and a domain-specific thesaurus. As there is unfortunately no test collection with relevance assessments available for the Dutch archaeology domain, we do not evaluate the performance of the information retrieval, only the performance of the NER.

We address the following research question:

1. To what extent does further pre-training a BERT model with domain-specific training data improve the model's quality in our highly specific domain?
2. Can a domain-specific BERT model be improved by adding domain knowledge from a thesaurus in a CRF ensemble model?
3. What errors are made by the models and what are the differences in predicted entities between the three models?

The contributions of our paper are three-fold: First, we propose entity-driven full-text search in which the professional user enters a structured query, and documents are filtered for the occurrence of the query entities detected by our new domain-specific BERT model. Second, we show that for a highly specific domain such as archaeology, further pre-training on domain-specific data increases the model's quality on NER by a much larger margin than shown for other domains in the literature. Third, we show that the domain-specific BERT model outperforms ensemble methods combining different BERT models, and also outperforms a CRF-based ensemble of BERT with explicit domain knowledge from the archaeological thesaurus.

We make our modified training data set, the pre-trained ArcheoBERTje model, and the fine-tuned ArcheoBERTje model for NER publicly available (Brandsen, 2021b).[3]

7.2 Related Work

In this section, we first summarise different approaches to NER (knowledge-driven and data-driven), followed by a discussion of related work on NER for document retrieval, on IR and NER in the archaeological domain, and we summarise the prior work on domain-specific BERT models.

7.2.1 Knowledge-driven and Data-driven NER

Early NER systems were knowledge-based, and relied on thesauri and handcrafted rules to detect entities (Rau, 1991). These methods are limited by the coverage of the thesaurus. Therefore, data-driven methods have become more popular, typically approaching NER as a supervised machine learning problem.

A highly effective machine learning method is Conditional Random Fields (CRF) (Lafferty *et al.*, 2001), which has become a common baseline for NER. Since 2011, word embeddings have become increasingly important as representations in NER. Especially Word2vec (Mikolov *et al.*, 2013) has been used extensively for NER (Sienčnik, 2015; Seok *et al.*, 2016). These embeddings-based methods typically feed the embeddings to CRF and/or Bi-LSTM algorithms to make NER predictions.

A big shift in NLP was introduced by Devlin *et al.* (2019), who presented their BERT (Bidirectional Encoder Representations from Transformers) architecture in 2019. BERT and other contextual embedding architectures are currently achieving SotA results with transfer learning for a large range of NLP tasks, including NER. Two major differences with previous embedding models are (1) that BERT embeddings are contextual, meaning that the same token can have a different embedding based on context, and (2) that it handles out-of-vocabulary words effectively, by dividing tokens into sub-tokens it does have in vocabulary, using the WordPiece (Devlin *et al.*, 2019) or SentencePiece (Kudo & Richardson, 2018) tokeniser.

Recent results indicate that ensemble methods that combining generic and domain-specific BERT models (Copara *et al.*, 2020), combining BERT with dic-

[3] https://doi.org/10.5281/zenodo.4739063, also available via the HuggingFace library for ease of use: https://huggingface.co/alexbrandsen

tionary features (Li *et al.*, 2020), or adding a CRF on top of BERT (Souza *et al.*, 2019) can improve NER quality. In this paper, we investigate whether addition of information from a thesaurus can improve NER in a highly specific domain.

7.2.2 NER for Document Retrieval

In the context of document retrieval, NER can play a role in better ranking or filtering documents based on entities in the query. Guo *et al.* (2009) were the first to address the task of recognising named entities in queries. They found that, despite queries in web search being short, 70% of the queries contained a named entity. They classify the entities according to a predefined taxonomy using a weakly supervised topic modelling approach on the query data. Cowan *et al.* (2015) also address NER in queries, but for the travel domain. They use CRF on the queries for extracting the relevant entities.

More recently, the relevance of NER on queries has been emphasised for the e-commerce domain. Wen *et al.* (2019) and Cheng *et al.* (2020) both implement end-to-end query analysis methods for e-commerce search; the extracted queries are then used to filter the retrieved products.

As opposed to the prior work, we do not focus on query analysis but on document analysis; our expert users prefer the use of structured queries, which makes query analysis unnecessary (see Section 7.4.4). Our documents, on the other hand, are long and unstructured (as opposed to the products in e-commerce search), making NER on the document side necessary for matching structured queries to the relevant documents.

7.2.3 IR and NER in Archaeology

As argued by Richards et al. (Richards *et al.*, 2015), archaeology has great potential for thesaurus-based IR and NER, as it has a relatively well-controlled vocabulary and there are thesauri of archaeological concepts available in multiple languages. However, unlike some other fields, archaeology terminology partly consists of common words, like 'pit', 'well' and 'post'. In addition, words can be archaeological entities or not, depending on the context in which they are used (past or present). For example, the word 'road' is not archaeologically relevant in the snippet "pit next to the main road", but is part of an archaeological entity in the snippet "a Roman road from 34 CE".

Archaeology has started experimenting with IR relatively recently. The focus of the prior work is on Information or Knowledge Extraction, mainly for automatically generating document metadata. An early study by Amrani *et al.*

(2008) aimed specifically at extracting information for archaeology professionals in a knowledge-based approach. A more data-driven approach using machine learning to detect Time Period entities was investigated in the OpenBoek project (Paijmans & Brandsen, 2010, 2009), but since then most studies have been knowledge-driven (Jeffrey *et al.*, 2009; Byrne & Klein, 2010; Vlachidis *et al.*, 2013, 2017).

More recently, Talboom experimented with embeddings in a Bi-LSTM model to recognise zooarchaeological entities (species and specific bones) (Talboom, 2017). A notable exception to the Information Extraction research we often see in archaeology is the work by Gibbs & Colley (2012) who created a full-text search engine on a small Australian corpus (roughly 1,000 documents) combined with facets based on manually entered metadata.

So far, NLP in the archaeology domain has not benefitted from BERT-based models. We believe it is a good candidate domain for BERT as the polysemy mentioned in the introduction and the present/past distinction mentioned above should be easier to detect with the context-dependent embeddings that BERT produces.

7.2.4 Language- and Domain-specific BERT Models

The original BERT paper (Devlin *et al.*, 2019) did not only present an English BERT model, but also a multilingual model (multiBERT) trained on data in 104 languages. This model is often used when no single-language model is available (Hakala & Pyysalo, 2019; Moon *et al.*, 2019; Kim & Lee, 2020). Research by Wu & Dredze (2020) shows that multiBERT achieved higher accuracy on NER and other NLP tasks than monolingual models trained with comparable amounts of data. Moon *et al.* (2019) also showed that fine-tuning multiBERT on a mixed language NER dataset provided better results than fine-tuning on individual languages.

However, recent work has shown that for some languages, multiBERT is outperformed by language-specific BERT models (Nozza *et al.*, 2020). For NER, this has been shown for Finnish (Virtanen *et al.*, 2019), Dutch (De Vries *et al.*, 2019), German (Chan *et al.*, 2021) and Russian (Kuratov & Arkhipov, 2019), among other languages.

For specific domains, it has been shown that further pre-training the English BERT-base model on large amounts of text from that domain increases the quality of the model on multiple tasks, although sometimes by a small margin. BioBERT in the biomedical domain shows an increase in F1 for NER of only 0.62% point (Lee *et al.*, 2019). SciBERT, trained on a large amount of scientific texts from

different domains, shows an increase in F1 for NER of 2 to 5% points, indicating that domain pre-training is useful for NER (Beltagy *et al.*, 2020). They also show that training BERT from scratch with a domain-specific vocabulary does not increase F1 substantially compared to fine-tuning an existing BERT model with an existing generic vocabulary, gaining only 0.6% points.

When we look at research done on non-English in a specialised domain like our study, there is little prior work. A study in the Russian cyber-security domain shows that the Russian model (RuBERT) outperformed multiBERT, and further pre-training RuBERT with domain-specific documents yielded the highest F1 (Tikhomirov *et al.*, 2020). In the Spanish biomedical domain, Akhtyamova (2020) shows a similar result, although their NER BERT model is trained for 30 epochs, possibly leading to over fitting.

To our knowledge, we are the first to address domain-specific NER for Dutch, and we are the first to automatically label a large archaeological document collection with our domain-specific BERT model for the purpose of professional search.

7.3 Data

The unlabelled data set we use for further pre-training the Dutch BERTje model to ArcheoBERTje consists of over 60k documents and 658 Million tokens across 16.6 Million sentences, around 2GB of data. The documents mainly consist of survey/excavation reports, but also include other documents such as research plans, appendices, maps, and data descriptions.

The labelled training data we use for NER we created previously (Brandsen *et al.*, 2020), and consists of fifteen documents that have been annotated by archaeology students. While fifteen reports is a relatively low number, these are longer than average documents, totalling 1,343 pages (average 89 pages per document), containing roughly 440,000 tokens and almost 43,000 annotated entities across six categories: Artefacts, Time Periods, Locations, Contexts, Materials and Species, see Table 7.1. The Inter Annotator Agreement reported is 95% (average pairwise F1 score), so it is of relatively high quality (Brandsen *et al.*, 2020). The data is stored in the BIO annotation schema, and is available for download.[4]

The data set has been split into 5 folds of 3 documents each. All methods are evaluated using this 5 fold split.

[4]Zenodo repository: http://doi.org/10.5281/zenodo.3544544

Entity	Description	Examples
Artefact	An archaeological object found in the ground.	Axe, pot, stake, arrow head, coin
Time Period	A defined (archaeological) period in time.	Middle Ages, Neolithic, 500 BC, 4000 BP
Location	A placename or (part of) an address.	Amsterdam, Steenstraat 1, Lutjebroek
Context	An anthropogenic, definable part of a stratigraphy. Something that can contain Artefacts	Rubbish pit, burial mound, stake hole
Material	The material an Artefact is made of.	Bronze, wood, flint, glass
Species	A species' name (in Latin or Dutch)	Cow, *Corvus Corax*, oak

Table 7.1: Descriptions and examples for each entity type. Examples are translated from Dutch. Adapted from (Brandsen *et al.*, 2020, p. 4574).

7.3.1 Pre-processing

For cross-validation, we divided the fifteen annotated documents across five folds so that each fold has a roughly equal number of tokens. The exact fold split and training data can be found on in the Zenodo repository.

We found that in the data set, sentences often exceed the maximum sequence length of 512 WordPiece tokens. This is not because sentences actually have more than 512 words, but partly because tables and OCRed maps and images create very long 'sentences' that are not cut up by the sentence detection algorithm. The other cause is that words that are uncommon outside of archaeology are cut up into many sub-tokens by the WordPiece tokeniser, as they do not exist in the vocabulary (also see Section 7.6.2).

Since sentences longer than 512 tokens will be trimmed, some of the input tokens will not get a prediction. To counteract this, we wrote a pre-processing script that attempts to break at a punctuation mark ('.', ';' or ',') between the 60th and 90th token and if there are none, it inserts a line break after the 90th token. This shortened the sentences sufficiently to have almost no instances where the sentence was longer than 512 WordPiece tokens. Only 136 tokens in the entire data set fell outside the 512 limit and received no prediction. These tokens only contained two entities, so the effect on the performance metrics will be negligible.

7.4 Methods

7.4.1 Baselines

As the first baseline, we use the method we published previously (Brandsen *et al.*, 2020), where we trained a CRF model using common word shape features (e.g. occurrence of uppercase letters, numbers), part-of-speech tags (e.g. noun, verb) and an archaeological thesaurus in a five word window, and performed hyperparameter optimisation. We used the same features, leading to a micro F1 score of 0.62. This is relatively low when comparing the score to NER in other domains, where F1 scores between 0.8 and 0.9 are common (Akhtyamova, 2020; Lee *et al.*, 2019).

The second baseline is the standard NER pipeline of `spaCy` 2.0, with default parameters (architecture: TransitionBasedParser.v2, random seed, max_steps: 20,000, Adam.v1 optimiser with learn_rate of 0.001). This method uses pre-existing Dutch word embeddings (`nl_core_news_lg`) with a deep convolutional neural network with residual connections, and a transition-based approach as the classifier (Honnibal & Montani, 2017).

7.4.2 Fine-tuning BERT for Dutch Archaeology and NER

Model training for evaluation To train ArcheoBERTje, we started with the Dutch BERTje model (De Vries *et al.*, 2019) and further pre-trained the model with our complete unlabelled archaeological collection, split into a 90/10 train and validation set.[5] We used the same configuration as BERTje, with a batch size of 4. We decided not to train a model from scratch as previous research showed only a minimal increase in quality compared to further pre-training (Beltagy *et al.*, 2020) an existing model, and because our corpus is relatively small and would probably not be enough to train an effective model.

To fine-tune the BERT models for the NER task, we used the labelled data and 5-fold cross-validation as described in Section 7.3.[6] For model comparison and to investigate the stability of each model with different random seeds, we trained all three models 10 times per fold, each time using a different seed (1, 2, 4, 8, 16, 32, 64, 128, 254, 512) and report averages over all runs and folds (50 runs in total per BERT model).

[5] We used HuggingFace's (Wolf *et al.*, 2020) language modelling script version 3.0.2.
[6] We used HuggingFace's token classification script version 3.0.2.

Model for full collection labelling To create the best possible model for inference on the entire corpus, we performed a grid search across hyperparameters as suggested by (Devlin *et al.*, 2019). We optimised the hyperparameters with fold 2 as test set, fold 1 as development set, and the other folds as training set, as this combination had the median F1 score across all models and folds. The grid search yielded the following optimal parameters for our data: 2 training epochs, $5 * 10^{-5}$ learning rate and 0.1 weight decay. We then fine-tuned the inference model on all labelled data with these hyperparameters. This way we maximise the amount of training data available for training the model that we use to label the full collection.

7.4.3 Ensemble Methods

As far as we are aware, we are the first to combine a multilingual model, a language-specific model and a domain-specific model into one ensemble method. We evaluate the following ensemble methods (one run over 5 folds per ensemble):

- Majority voting on the predictions of multiBERT, BERTje and ArcheoBERTje;
- CRF which uses the prediction labels of the three models as features;
- CRF which uses the prediction labels of ArcheoBERTje only;
- CRF which uses the prediction labels of the three models as features, combined with the baseline features;
- CRF which uses the prediction labels of ArcheoBERTje only, combined with the baseline features;
- CRF which uses the embeddings produced by ArcheoBERTje as features.

The above mentioned 'baseline features' are those adopted from prior work (See Section 7.4.1) and include word shape, part-of-speech tags and thesaurus features. We optimised the hyperparameters of each CRF ensemble with gradient descent using the L-BFGS method, optimising c1 and c2 (the coefficients for L1 and L2 regularisation). The optimisation was run separately for each fold. All CRF ensembles use a 5-token window, taking into account the features from the two tokens before and after the current token.

The thesaurus we use in our CRF baseline and ensembles is the ABR (*Archeologisch Basisregister*) (Brandt *et al.*, 1992; Brandsen *et al.*, 2020), a thesaurus containing time periods (e.g. Bronze Age), artefacts (e.g. axe) and materials (e.g. flint). A token is assigned the binary feature 'occurs in period/artefact/material list' if it is part of an n-gram that occurs in the thesaurus. So the token 'Bronze' would only be assigned a positive value for the feature if the token 'age' follows it.

7.4.4 Entity-driven Document Search

Indexing Before we index the documents, we first run the inference NER model on each page to detect the entities. We then store the entities and full text in a JSON file for each document, together with the relevant metadata (authors, DOI, coordinates, document type, etc) retrieved from the DANS repository via an API.

To tackle the synonymy problem for time periods (see Section 7.1), we use a custom script that translates all extracted Time Period entities to year ranges. It uses regular expressions to convert dates (e.g. '100 BCE', 'start of the 9th century') and an extended and customised version of the PeriodO time period gazetteer (Rabinowitz *et al.*, 2016) to translate Time Periods (e.g. 'Bronze Age', 'Medieval period'). These date ranges are added to the JSON and can be used to filter results by allowing users to specify a date range in their query. These JSON files are then sent to an instance of ElasticSearch running on a webserver, which indexes them. At the moment, the retrieval unit is a page, so for any query the terms/entities must occur together on a page. We are aware this is not optimal, as search terms might be split across pages. As such, in future work we will index per document section by using a section detection algorithm.

Query Interface and Analysis Our search engine has a faceted search interface in which metadata filters are combined with entity fields and full-text search (Tunkelang, 2009). We have included facets for document type and subject (metadata fields). In addition, as requested by our target group, we added geographical search via a map functionality, which allows users to draw a rectangle or polygon to search only in a certain region.

At query time, the user can specify if they are looking for a specific entity type and/or specify a date range in which they are interested. The entities and date range are used to filter the result set and can be combined with a standard full text search. This allows for relatively complex queries such as "Artefact: urn AND Context: cremation AND startdate < -2000 AND enddate > -800 AND fulltext: upside down". This example is a real request entered by an archaeologist, who was looking for upside down urns in the Bronze Age in or around cremations. Users do not need to use complex query syntax, but can instead define their query by filling in the relevant fields in the graphical user interface, as shown in Figure 7.1.

Document Ranking Most archaeological information needs are recall-oriented tasks: the users want a complete list and do not mind having irrelevant results

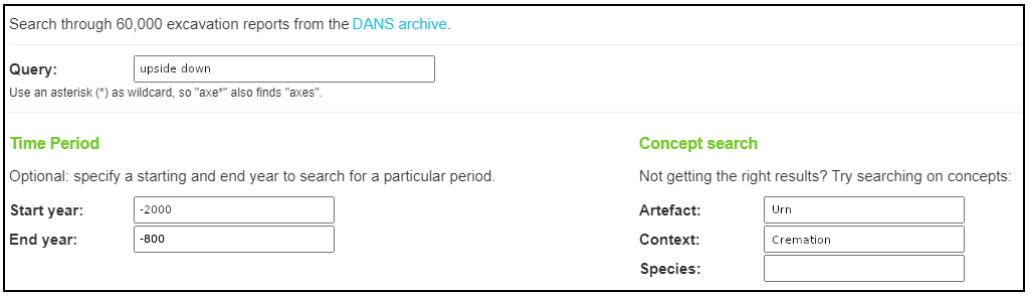


Figure 7.1: Query interface showing query for "Artefact: urn AND Context: cremation AND startdate < -2000 AND enddate > -800 AND fulltext: upside down". Interface and query translated to English for the readers' convenience.

in the (top of) the result set (Brandsen *et al.*, 2019). As the focus of our work is on entity-driven search, we opt for the default ElasticSearch ranking model, consisting of Term Frequency - Inverse Document Frequency (TF-IDF) and the field-length norm (the shorter the field, the higher the relevance) (ElasticSearch, 2018). The only field included for ranking is the page text content, other fields are only used for filtering.

Note that we do not evaluate the ranking, because there is no test collection available yet for Dutch archaeological document retrieval. Therefore, the scope of this paper is limited to the NER and the evaluation thereof.

7.5 Results

7.5.1 Model Stability and Quality

Table 7.2 shows the micro average precision, recall, and F1 score for the three BERT models, compared to the CRF and spaCy baselines. We find that the multilingual BERT model does not outperform the baselines, but the more specialised BERTje and ArcheoBERTje models do, with ArcheoBERTje achieving the highest F1 score.

We also show the average standard deviation over 10 runs with different seeds for 5 folds. The standard deviation between runs is very low, between 0.015 and 0.004. The recent work by Tikhomirov *et al.* reports a standard deviation of 0.015 to 0.008, similar to our results (Tikhomirov *et al.*, 2020). When comparing the predicted labels of each of the models in a pairwise manner, the differences are significant according to McNemar's test (χ^2 between 650 and 4276, $p < 0.00001$).

Model	Precision	Recall	F1 (Std.)	Fails
CRF Baseline	**0.785**	0.526	0.630	n/a
spaCy Baseline	0.717	0.602	0.654	n/a
multiBERT	0.623	0.550	0.583 (0.015)	4
BERTje	0.718	0.682	0.699 (0.005)	0
ArcheoBERTje	0.743	**0.729**	**0.735 (0.004)**	0

Table 7.2: Micro average precision, recall and F1 score at token level (B and I labels), over 10 runs with different seeds, for each of the 5 folds (50 runs total). Standard deviation of F1 over the 10 runs is added in brackets for the BERT models. Standard deviation of precision and recall lies between 0.006 and 0.020. The 'Fails' column indicates the number of times the model failed to learn (F1 = 0).

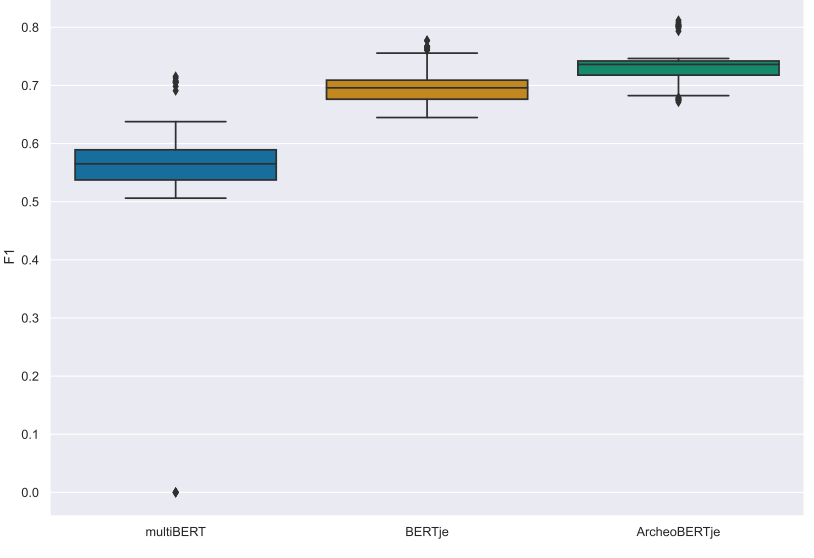


Figure 7.2: Distribution of F1 scores over ten runs with different seeds, for each of the 5 folds (50 runs per model). The zero scores for multiBERT are runs where the model failed to learn.

Figure 7.2 shows the distribution of F1 scores over the 50 runs per model in a boxplot. Here we again see that the standard deviation is low, and that ArcheoBERTje consistently outperforms the other two models. The F1 scores of 0 for multiBERT are outliers, and we assume these are caused by the ADAM optimiser getting stuck in a local minimum where the loss does not decrease. In this local minimum, predicting the majority class (O) seems to yield the highest accuracy, but of course O labels are not taken into account when calculating an F1 score for NER, so we get a score of zero. This can be solved by changing the learning rate, but this would not change the overall view that BERTje and ArcheoBERTje outperform multiBERT, so we did not investigate further on fixing this for multiBERT.

The low standard deviations for ArcheoBERTje indicate that further pre-training with domain-specific data does not only increase the model quality on average, but also makes the model more stable, reducing the chance of getting a sub-optimal model in a run.

Another way to compare the models is by looking at differences between the errors made. In Table 7.3 we report the top 10 most frequent error combinations for the three models. Here we can see that quite often, BERTje and ArcheoBERTje have similar predictions (whether correct or not), while multiBERT predicted a different label. We see that multiBERT often misses Locations (LOC), Artefacts (ART) and Species (SPE), and sometimes predicts entities that are not there. The first error combination where ArcheoBERTje outperforms BERTje is number 9, having correctly predicted B-ARTs while the other 2 models do not. In Sections 7.5.3 and 7.6.1 we further analyse the output and errors made by the ArcheoBERTje model to provide insight into the model's behaviour.

7.5.2 Ensembles

Table 7.4 shows the results of the ensemble methods.[7] The highest F1 (0.757) is obtained by he optimised production ArcheoBERTje model.

The highest precision is obtained by the CRF ensemble with the baseline features combined with the predicted labels from all three models. The highest recall is achieved by ArcheoBERTje solo.[8] Using a CRF with BERT embeddings

[7] As the standard deviation between multiple runs is low, combining multiple runs of the same model in an ensemble model is very unlikely to increase the F1 score, at the expense of a vastly increased computing time and cost. Hence we do not apply this approach.

[8] For general domain Portuguese NER, Souza *et al.* show the same pattern: Portuguese BERT has the highest recall, while combining BERT with CRF yields the highest precision and F1 (Souza *et al.*, 2019).

Freq.	True	multiBERT	BERTje	ArcheoBERTje
1137	B-LOC	O	B-LOC	B-LOC
1122	B-ART	O	B-ART	B-ART
1015	O	B-ART	O	O
575	B-SPE	O	B-SPE	B-SPE
561	O	B-LOC	O	O
466	B-PER	O	B-PER	B-PER
429	O	O	B-ART	B-ART
425	I-PER	O	I-PER	I-PER
402	B-ART	O	O	B-ART
373	O	I-PER	O	O

Table 7.3: The 10 most frequent error combinations between the 3 models for which at least one model has the correct prediction. Errors are marked in red.

Ensemble	Precision	Recall	F1
ArcheoBERTje (50 runs avg)	0.743	0.729	0.735
ArcheoBERTje (optimised production model)	0.784	**0.731**	**0.757**
Majority Voting	0.784	0.695	0.737
CRF with 3 BERT model prediction labels as features	0.786	0.683	0.731
CRF with only production ArcheoBERTje predictions as features	0.786	0.717	0.750
CRF with 3 BERT model prediction labels + baseline features	**0.795**	0.644	0.712
CRF with production ArcheoBERTje prediction labels + baseline features	0.793	0.649	0.714
CRF with only production ArcheoBERTje embeddings as features	0.767	0.604	0.676

Table 7.4: Micro F1 score, precision and recall for the six ensemble methods, for one run over five folds. ArcheoBERTje results averaged over 50 runs and the optimised production model are added for comparison. The ArcheoBERTje predictions used as features for CRF are from the production model. The baseline features are the word- and context-based features used for CRF in prior work.

Entity	Total	Unique	Top 5
Artefacts	2,520,492	53,675	pottery, charcoal, flint, bone, brick
Contexts	1,602,124	21,319	pit, ditch, posthole, well, house
Materials	457,031	6,146	wooden, flint, wood, metal, bronze
Locations	3,488,698	147,077	nederland, ' , groningen, noord - brabant, gelderland
Species	928,437	34,540	cow, hazel, sheep, goat, pig
Time Periods	4,698,323	98,445	roman period, iron age, 150 - 210, late medieval, modern
Total	13,695,105	361,202	

Table 7.5: Overview of entities detected in the entirecorpus, showing total and unique counts, plus the top 5 for each entity (translated from Dutch where relevant).

as features instead of the default BERT classifier (softmax), does not increase performance. Given the recall-oriented nature of professional search tasks like ours, we prioritise recall over precision for the NER labelling, and use ArcheoBERTje for labelling the full collection.

7.5.3 Analysis of the Retrieval Collection

After labelling the full retrieval collection with ArcheoBERTje, we analyse the extracted entities. Table 7.5 shows for each entity type the total frequency and the amount of unique entities. We also show the top 5 entities extracted for each type (translated from Dutch to English).

As we already mentioned in the introduction, archaeologists are interested in the What, Where and When of excavations. And so we see that Artefacts, Locations and Time Periods are the most common entities.

- For Artefacts, we see that pottery and flint are common, which we expected, but apparently also charcoal, which we did not expect, but could be explained by the use of carbon dating, which often uses charcoal as a sample.
- In the Locations category, we see that the second most common entity is an apostrophe ('). While this is clearly not a location, luckily it will not affect retrieval as it is not something users would search for, and ElasticSearch does not include apostrophes in its index, so it would not match any documents. We speculate that ArcheoBERTje mislabels apostrophes

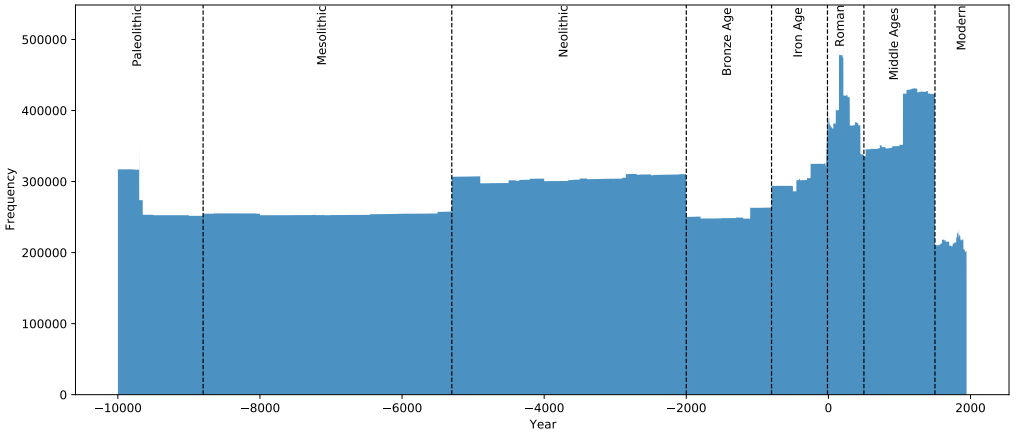


Figure 7.3: Graph showing for each year in each detected time period, how often it occurs in our data set, labelled by ArcheoBERTje. For clarity, years before 10,000 BCE are not included. Major time periods are denoted with dashed lines.

as locations because of the occurrence of apostrophes in some Dutch place names (e.g. *'s Hertogenbosch*).

- For Time Periods, the only unexpected entry in the top 5 is "150 - 210". When we investigated this further, we found this is actually a soil grain size used in coring reports, which have been incorrectly labelled as a time period by ArcheoBERTje. 150-210 µmm is the grain size for medium course sand, apparently the most common grain size in the Netherlands. When we look further down the Time Period top 100, we also see other common grain sizes: 210-300, 105-150 and 105-210. This is an issue when searching for archaeology between 105 and 300 CE, as these irrelevant coring reports will also be returned. We believe that these errors are made because these numbers come from tables, and as such do not have any sentence context, making them difficult to predict correctly. The most likely way to fix this is by making a post-processing correction on the extracted entities. This is something we will improve in the next version of our NER method.

The grain sizes are also clearly visible in Figure 7.3, in which we have plotted the frequency of years found in entities in the corpus. The figure shows a number of plateaus, indicating the use of time periods instead of single dates, i.e. the last plateau is the Late Middle Ages ending in 1500 CE. These plateaus are not completely flat as single dates and subperiods can cause spikes and smaller

sub-plateaus.

The thin spike just after the year 0 can probably be attributed to misclassified entities, i.e. the '10' in '10-02-2006' being labelled by ArcheoBERTje as a Time Period and translated to 10 CE. Other than this we see a big plateau in the middle (5300–2000 BCE), which represents the Neolithic. This indicates that a large amount of data is available describing this period in the Stone Age.

7.6 Discussion

7.6.1 Error Analysis

Figure 7.4 shows the confusion matrix between labels predicted by ArcheoBERTje and the true labels. The diagonal line and the first row and column are typical for NER. The diagonal shows the true positives, the top row is where the model predicted an entity where there isn't one, and the first column is where the model predicted O where there should be an entity. We also see the I / B label confusion quite clearly, mainly for Time Periods and Locations, where the model predicts an I instead of a B, or the other way around.

A more interesting error is the confusion between Materials and Artefacts. This is caused by words like "flint", which can be both an Artefact ("a piece of flint") or a Material ("a flint axe"). In Dutch, "pottery" has the same issue. Even archaeologists struggle with distinguishing between the two (Brandsen *et al.*, 2020), so it is unsurprising that ArcheoBERTje finds this difficult as well. As there is a lot of ambiguity in this entity category, perhaps merging the two categories into one entity type would increase the overall performance. We have seen in previous research that archaeologists will also confuse the two categories when creating queries, so having them both in one search field might not even cause any problems at search time.

Table 7.6 shows the evaluation per entity type. In general, the I labels are more difficult to predict, and Materials are more difficult than the other entities. In fact, Materials are currently not included in the search engine, as archaeologists find it difficult to differentiate between Materials and Artefacts in their queries, so this will not affect retrieval quality. When we remove Materials from the overall micro F1 score calculation, we get an increase of only around 0.01, as there are only a small number in our training data, around 3000.

When we look at some of the errors made by ArcheoBERTje in more depth, we find some interesting patterns. For example, for missing B-ART labels, many errors are adjectives that were assigned the O label, e.g. for "big axe" or "complete pot", the adjectives are labelled O, and axe / pot are labelled B-ART. This

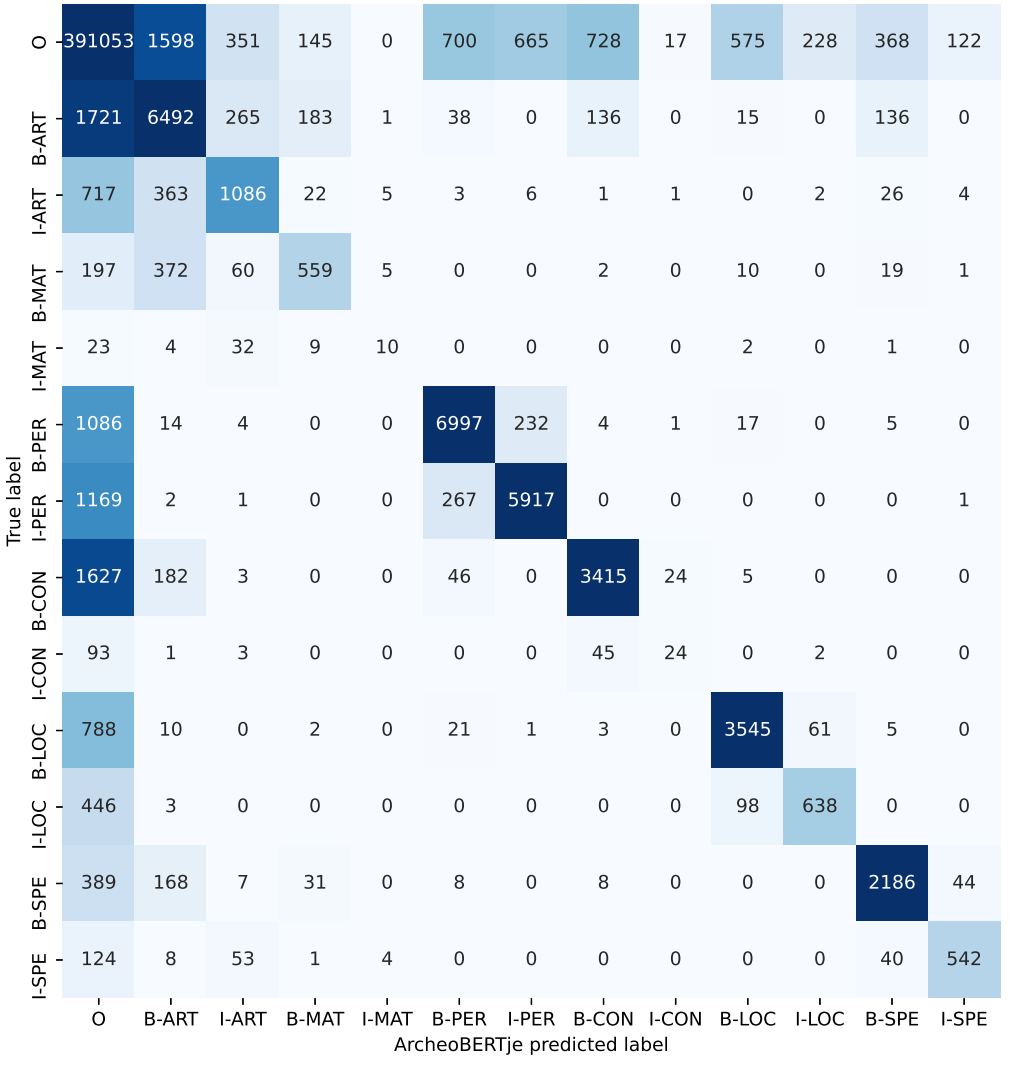


Figure 7.4: Confusion matrix between true labels and ArcheoBERTje predictions.

	Precision	Recall	F1
B-ART (Artefacts)	0.704	0.722	0.713
I-ART	0.582	0.486	0.530
B-CON (Contexts)	0.787	0.644	0.708
I-CON	0.358	0.143	0.204
B-MAT (Materials)	0.587	0.456	0.514
I-MAT	0.400	0.123	0.189
B-LOC (Locations)	0.831	0.799	0.815
I-LOC	0.685	0.538	0.603
B-SPE (Species)	0.785	0.769	0.777
I-SPE	0.759	0.702	0.729
B-PER (Time Periods)	0.866	0.837	0.851
I-PER	0.867	0.804	0.835
Macro Average	0.684	0.585	0.622
Micro Average	0.784	0.731	0.757

Table 7.6: ArcheoBERTje precision, recall and F1 score for each label.

error is not surprising as most archaeologists would probably find it difficult to define these entities as well. In addition, users are more likely to only search for the base artefact and not include an adjective, so they would search for "pot" not "complete pot". In a pilot study evaluating our archaeological search engine, we analysed users' search behaviour and found that of the 148 issued queries, none included an adjective.[9]

For Time Periods, we again see that adjectives are missed from the start of an entity, but also prepositions. Some examples include "from", "between" and "start of". Also we find that connecting words between Time Periods are missed, such as "and", "or" and "Âś" (used to denote the standard deviation of a carbon dating). While this does cause some noise, missing adjectives/prepositions or connecting words are not a considerable issue if the main period has been detected. I.e. for "start of 10th century", if we miss "start of" this means the year range is 900 to 1000 CE, instead of 900 to 925 CE. Again, as archaeologists care more about recall than precision, this should not hinder their search.

The predicted Context[10] entities also have some interesting anomalies. In particular, we analysed the top 10 most misclassified tokens and we found that

[9]Extension and publication of this user study is part of our future work.

[10]For clarity, Contexts are defined as an anthropogenic structures or objects that can contain Artefacts, i.e. rubbish pits, burials, houses, and so on.

these are all words that can denote contemporary objects (and thus not a Context) or actual (pre-)historical Contexts. An example is "*put*", which can mean a trench dug by archaeologists, or a water well found in an excavation, and both instances of *put* can contain an artefact, leading to similar contexts around these words. Other examples are "house", "church", "ditch", "mine" and "settlement". It seems that even with the context-dependent embeddings BERT produces, these ambiguous words are still a challenge. Perhaps future language models are more refined and might be able to distinguish between these types of ambiguous terms.

A special case is the word "*poel*" (pond). We see that this token is always labelled as O while it is in fact a Context. When we checked the sentences this word occurs in, we see they are all very typical of Contexts, i.e. "we found pottery in the pond", which is similar to sentence structures of other Contexts that are classified correctly. The only possible explanation we can find is that the word *poel* only occurs in one of the documents, so when this document is in the test set, the word does not occur at all in the train or dev set. This confirms the importance of creating train-test splits on the document level, to avoid overfitting. At the same time, this might be an issue that could be potentially alleviated by increasing the size of the training data.

More generally speaking, we see that the BERT models make impossible B and I predictions, i.e. an I label without a B label for the previous token. Unlike CRF, which learns the probabilities of two labels occurring after one another, BERT sees every token as an individual classification task without taking into account the predicted label of the previous token. This might explain why the CRF model with ArcheoBERTje labels as features (see Table 7.4) outperforms ArcheoBERTje on precision, as it corrects some of these mistakes. Perhaps another approach to correct this is a rule-based postprocessing step that checks the validity of I labels following B labels, and corrects impossible combinations.

During the annotation process, we used a test document of a hundred sentences (1,962 tokens) to calculate the Inter Annotator Agreement (Brandsen *et al.*, 2020). We added ArcheoBERTje predictions to this data, to see if ArcheoBERTje predictions are more often wrong when humans also have disagreement, indicating that the model mimics human confusion. We disregard tokens where everyone (including ArcheoBERTje) predicts an O label, leaving 292 tokens. In 57.5% of these tokens, all annotators and ArcheoBERTje predict the same label. In 31.5% of tokens, there is some disagreement between annotators, but ArcheoBERTje predicts the same label as the majority, and in 4.4% of tokens, ArcheoBERTje predicts a label different from the majority. In 6.5% of tokens, ArcheoBERTje predicts one label, while annotators all predict the same different label. This is only a small sample, but the above suggests that BERT models

are decently equipped to learn from the majority where there is inter-annotator disagreement.

7.6.2 Tokenisation Issues

The vocabulary of a BERT model is determined by the collection used for pre-training. The WordPiece tokeniser optimises the set of (sub-word) tokens to maximise the coverage of the collection's vocabulary. The same tokenisation is applied to the input sentences at inference. An example is shown below, where we compare tokenisation with the multiBERT and BERTje vocabularies. We see that target entities ("*Swifterbant*", "*aardewerkscherven*" and "*Midden Neolithicum*") are split up into three or more sub-tokens by the multiBERT and BERTje tokenisers.

Original sentence:
"*In put twee werden 3 Swifterbant aardewerkscherven aangetroffen uit het Midden Neolithicum.*" ("In trench two, 3 Swifterbant pottery shards from the Middle Neolithic were found.")

multiBERT tokenisation (23 tokens):
In put twee werden 3 Swift ##er ##bant aarde ##werks ##cher ##ven aan ##get ##roffen uit het Midden Neo ##lit ##hic ##um .

BERTje tokenisation (20 tokens), also used for ArcheoBERTje:
In put twee werden 3 Swift ##er ##ban ##t aardewerk ##scher ##ven aangetroffen uit het Midden Neo ##lith ##icum .

As an additional analysis, we trained a SentencePiece tokeniser on our archaeological collection, with the same vocabulary size as the BERTje model (30k).

Archaeology tokenisation (14 tokens):
In put twee werden 3 Swifterbant aardewerk ##scherven aangetroffen uit het Midden Neolithicum .

The examples show that a more specific pre-training corpus would lead to more complete domain words. However, our collection is small for such from-scratch pre-training and the experiments in the sciBERT paper have shown that even a much larger pre-training collection only gives a +0.6% point F1 increase compared to further pre-training the generic model (Beltagy *et al.*, 2020).

Understandably, the problem of input sequences longer than 512 tokens was occurring more often with the multilingual model, as the vocabulary (with fixed

size) is not solely Dutch. This means that many less common Dutch words are not in the vocabulary, and are cut into many sub-tokens by the WordPiece tokeniser. This effect is aggravated by the Dutch language having a lot of compound words and a much longer average word length (4.8 in English (Norvig, 2013) vs. 8 in Dutch (Corstius, 1981)).

For our experiments comparing the different BERT models, it was sufficient to split up long sentences in the training and test data as a data preprocessing step. However, for the inference described in Section 7.5.3, we did not preprocess the text, and as such, entities found in long sentences after 512 SentencePiece tokens will have been assigned the incorrect "O" label, skewing the results. In future research, we will implement an automatic sentence splitting module, similar to the one implemented in FLAIR (Akbik *et al.*, 2019).

7.7 Conclusion

In this paper, we have evaluated BERT models for Named Entity Recognition in the Dutch archaeological domain, with the purpose of improving our archaeological search engine. We implemented the search engine for a large archaeological text collection, with a structured query interface that allows the specification of entity types. The document collection is automatically annotated with archaeological named entities such as Location, Time Period, and Artefact.

In response to our research questions, first, we found that fine-tuning a BERT model with domain-specific training data improves the model's quality by a large margin for the archaeological domain, larger than in related work addressing domain-specific BERT models. We achieve an average F1 of 0.735 after hyperparameter optimisation, and very small standard deviations over runs with different random seeds.

Second, the domain-specific BERT model was superior in F1 and recall than an ensemble combining multiple BERT models, and could not be further improved by adding domain knowledge from a thesaurus in a CRF ensemble model. This indicates that after pre-training and fine-tuning on a domain-specific collection, the BERT model already covers the relevant information from the domain thesaurus. We did find a higher *precision* when we combined all three BERT models in a CRF model and added domain knowledge. However, as almost all information needs in archaeology are recall-oriented, and combining models is computationally expensive and environmentally taxing Strubell *et al.* (2020), we opt for the ArcheoBERTje model for labelling the full retrieval collection.

Third, our error analysis shows that there is confusion between the Artefact

and Material entities, similar to what humans experienced in the annotation process. For Artefacts and Time Periods, a common error is missing the adjective or preposition in an entity. The detection of Time Periods is a bit noisy, with other non-year numbers erroneously labelled as time ranges. Context entities such as "house" and "ditch" are difficult for the models to distinguish from non-entity words. Creating train-test splits on the document level is important to avoid overfitting, as the consistently misclassified Context "*poel*" shows, which only occurs in one document. An analysis of tokenisation by each of the models indicates that the multiBERT model is hampered by the rough tokenisation, splitting many relevant terms in sub-words.

In the near future, we will evaluate the entity-driven search engine with users, both in a controlled experiment and in natural search contexts. We will also investigate entity-based query suggestion. Once entities are mapped to a thesaurus or embedded in a semantic space, this allows for query improvement by suggesting parent or sibling entities in the thesaurus or nearest-neighbours in the embedding space.

8

Case Study

"*Is is seldom possible to say of the medievals that they *always* did one thing and *never* another; they were marvelously inconsistent.*"
Thomas Cahill, Mysteries of the Middle Ages: The Rise of Feminism, Science and Art from the Cults of Catholic Europe

Previously published as: Brandsen, A. and Lippok, F, 2021. A Burning Question – Using an Intelligent Grey Literature Search Engine to Change our Views on Early Medieval Burial Practices in the Netherlands. *Journal of Archaeological Science*, 133, pp.15456. DOI: 10.1016/j.jas.2021.105456.

This paper presents a case study on Early Medieval burial practices using AGNES, an intelligent search engine for Dutch archaeological grey literature. Traditionally, it is assumed that cremations phased out at the start of the Early Middle Ages, when the inhumation practice became more numerous. However, recent research (Lippok, 2020) shows that cremations might be more prevalent than previously assumed. Due to research efforts being concentrated on furnished inhumations, cremations and other types of burials have not received their share of research interest. It is suspected that unknown Early Medieval cremations may be found in grey literature research reports. The rapidly growing document collection requires more efficient methods to search through this big data, as manual searching is too time intensive.

AGNES uses machine learning to allow searching on archaeological concepts (such as time periods and artefacts) in full texts, solving common problems with synonymy and polysemy. This paper describes a controlled search for Early Medieval cremations in the Netherlands, and a comparison of the new information uncovered from the reports to prior knowledge on the topic. The queries resulted in 2541 hits. Twenty-three Early Medieval cremations that were previously unknown to experts were uncovered, and 31 (of 77) known sites were identified. Forty-one possibly interesting documents were noted, where it was not clear from the report if Early Medieval cremations were present. 2446 documents were not relevant to the study for various reasons.

The 23 new sites are an increase of 30% over the existing knowledge of experts. In the last 20 years only nine new sites featuring Early Medieval cremations were discovered, so being able to add another 23 is a major development. Adding previously unknown Early Medieval cremations to the data set challenges the existing bias for inhumation graves, and supports striving towards viewing the Early Medieval burial repertoire as more heterogeneous and representative. This indicates that AGNES is useful for archaeological research, and the uncovered information can lead to a more cohesive view of the past.

8.1 Introduction

Archaeologists produce large amounts of texts, from monographs and articles to research reports written by commercial units. In the Netherlands, it is estimated that around 60,000 of these reports were produced up until 2017, with 4,000 being added every year (Rijksdienst voor het Cultureel Erfgoed, 2019a). Often, these reports are created in-house and not circulated widely, made available only via online repositories such as the DANS archive or the Dutch national archaeology

database, Archis, maintained by the RCE (*Rijksdienst voor het Cultureel Erfgoed*, the Dutch heritage agency).

Such reports are also known as grey literature: documents produced outside of the traditional commercial and academic publishers, often with a small audience and not peer-reviewed. These reports are currently underused especially for synthesising research, even though many authors note that the information in these documents can be of great value (Evans, 2015; Richards *et al.*, 2015; Brandsen *et al.*, 2019).

For Dutch grey literature, it was until now only possible to search through the metadata of the documents, not the texts themselves, using the above mentioned DANS and Archis search systems. This is not ideal, as archaeologists will often want to search more fine grained, and might be interested in a single Iron Age artefact on an otherwise Medieval site, which is not mentioned in the temporal metadata as it is too specific. Such examples are often called 'by-catch', one or a few finds that are different from the rest of the excavation. This by-catch is currently impossible to find effectively, and archaeologists report they currently download large numbers of reports and manually search through each PDF file, a time consuming and inaccurate task (Brandsen *et al.*, 2019).

One way to make this document collection more useful is by applying full text search. Similar to e.g. Google, this allows archaeologists to search through all of the text of all of the documents. This would already be an improvement to the current situation, but from previous research (Vlachidis *et al.*, 2017; Habermehl, 2019) and our own prior work (Brandsen *et al.*, 2020) it is evident that for archaeological texts, synonymy and polysemy are common and can cause problems.

Synonyms are different words that have the same (or similar) semantic meaning. An example is the Middle Ages, which can also be written as the Medieval Period, the Dark Ages, 500-1500 CE, and includes subperiods such as the Merovingian Period, and dates such as 600 CE or 1050 $\pm$ 23 BP. Ideally, an archaeologist will want to find all of these mentions when looking for reports about the Middle Ages.

Polysemy is the other way around: one word having multiple meanings. An example is "Mayen", which can indicate a type of pottery, a pottery production site spanning ten centuries, a volcanic island in the Norwegian sea called "Jan Mayen" or a town in Germany (which the first two are named after). But when you type "Mayen" into a standard text search engine, the system does not know which meaning you are interested in, and will provide results about all the meanings, or maybe just the most popular meaning.

A combination of full text search with entity search is used to solve the potential problem of polysemy. We apply Named Entity Recognition (NER), a method

which automatically finds and labels relevant concepts for archaeologists. This can be done using a rule-based approach, or – in the case of our study – by using Machine Learning. The NER process uses the context of words to attempt to distinguish between meanings. This process is not 100% accurate, as ambiguous entities are a challenge: e.g. "*aardewerk*" which can mean the artefact 'pottery', or the material 'ceramic' in Dutch. But where the NER is correct, users searching for e.g. "flint" will not find someone with the last name 'Flint', only flint artefacts. Once these entities have been found, they are indexed together with the full text in a search system. Detected time periods are automatically translated to year ranges and indexed as additional information. This makes it possible to search on year ranges instead of written timespans, tackling the problem of synonymy for this entity type. The system is called AGNES (Archaeological Grey-literature Named Entity Search), and it is available online at agnessearch.nl. In this project, 65,000 PDF documents were indexed, obtained from the DANS repository in 2017 (Brandsen *et al.*, 2019).

To assess the usefulness and potential of AGNES, we perform a case study on Early Medieval cremation practices. In a recent article, Lippok (2020) has shown that Early Medieval cremations regularly occur next to the traditionally expected inhumation burials. The data for that article was compiled over three years through extensive literary study of published and unpublished books, PhD and master theses, and building on a career-long effort by Prof. Frans Theuws to map all Early Medieval sites in the Netherlands in the Rural Riches project (2021). This resulted in 77 sites containing cremation burials (Lippok, 2020). Commercial reports were rarely consulted, due to the time-intensive process of surveying grey literature. Only after personal prompts by excavators indicating they had found Early Medieval cremations a few of these were added to the database.

Traditionally, cremation burials are presumed to disappear at the beginning of the Early Middle Ages (Effros, 2003; Fehr, 2008; Van Es, 1968; James, 1988). Prior to 2020, their occurrence was never systematically investigated in the Netherlands due to their perceived incidental nature. When cremations were encountered in the past, they did not receive much attention as it was assumed they represented a burial rite that was on its way out of the burial repertoire (Van Es & Schoen, 2008, 858). Cremations were mostly marginalised, the explanations for their occurrence was restricted to them representing individuals of different ethnic backgrounds, different religious persuasion or different social status from the majority of people that were inhumed (Lippok, 2020).

The two burial forms were handled in exclusionary ways: where inhumation was prevalent, cremation was considered an anomaly. This dichotomous interpre-

tation was countered by showing that inhumations and cremations occur most often together, even within the same grave context, rather than apart. There are further similarities: the use of the same material culture in cremations and inhumations, people of all ages and sex were both inhumed and cremated, and external funerary structures such as posts and mounds have been shown as similar too (Lippok, 2020). Establishing the Early Medieval cremation ritual next to the inhumation ritual meant a shift in perspective, moving from homogeneous to heterogeneous burial repertoires.

Cremations are well suited as a case study for AGNES, as grey literature may contain information left out of traditional mediums such as books and papers. The results are relevant as they undermine an unbalanced focus on furnished inhumation graves: every single cremation grave contributes to the view that other ways of burial occurred and that archaeologists should not exclusively concentrate on what is considered the most prevalent type of burial.

This paper presents the latest version of AGNES, and describes a case study on Early Medieval cremations to show the usefulness of the system in archaeological research. The following research questions are addressed:

- How many Early Medieval cremations can we find in existing archaeological research reports that are currently unknown to a group of specialists in that area (the Rural Riches project), and how many known ones?
- What is the effectiveness of AGNES for retrieving relevant documents on this topic?
- To what extent does this new knowledge change our views on Early Medieval cremations?

Our contributions when compared to previous research are two-fold: (1) we present the first combined full text and entity search in the archaeology domain and (2) we provide the first systematic survey of Early Medieval cremations in Dutch grey literature. The creation of the search engine is work by Brandsen, and the research on cremations has been performed by Lippok.

8.2 Methods

In this section, we will first describe the search system, then explain the searching process, and end with a description of the evaluation.

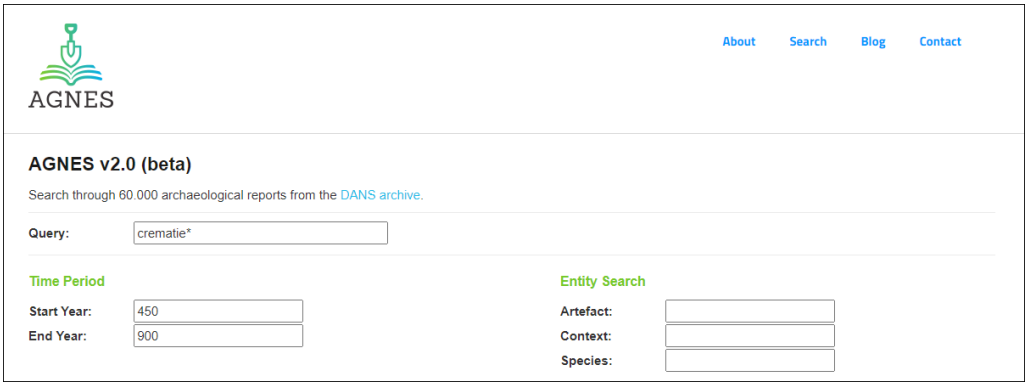


Figure 8.1: Screenshot of the AGNES query interface (translated from Dutch)

8.2.1 AGNES

The latest version of AGNES was created based on a user requirement study (Brandsen *et al.*, 2019) and an interface usability study (Brandsen, 2021b) with a small but representative group of Dutch archaeologists, to ensure the system is fit for purpose. An earlier version of AGNES, the initial prototype, is described in Brandsen *et al.* (2019). Since then, the interface was simplified and the accuracy of the NER has increased. Figure 8.1 shows a screenshot of the current query interface, which allows users to search for any term (in the Query field), or search for particular archaeological entities (artefacts, contexts and species) and/or time periods, which are denoted by a start and end year. There is also a geographical search integrated (not pictured here). The system can be accessed at agnessearch.nl/search/agnesv2.

The entities were detected using machine learning models trained on expert-labelled data. Specifically, we have further pre-trained a BERT (Bidirectional Encoder Representations from Transformers) model, which is a deep neural language model. The BERT architecture is used for many Natural Language Processing (NLP) tasks, and it is currently achieving state of the art results on benchmark tasks in many languages (Devlin *et al.*, 2019; Liu *et al.*, 2019; Xiong *et al.*, 2020). Once the generic language model is pre-trained, it can be fine-tuned for a specific NLP task, in our case NER.

A Dutch BERT model called BERTje has been released (De Vries *et al.*, 2019). We further pre-trained the BERTje model on a collection of 65,000 Dutch archaeological documents obtained from DANS, training on four GPUs for about 24 h. The result is our ArcheoBERTje model. ArcheoBERTje outperforms the

generic Dutch model BERTje, with an F1 score of 73.5% on our test set (Brandsen *et al.*, 2021a).

ArcheoBERTje was used to detect entities in our entire document collection, leading to a total of over thirteen million entities. For time periods, the detected entities are translated to a year range (e.g. "Middle Ages" to 500-1500 CE), using a combination of regular expressions and dictionary lookups. All the entities, year ranges and full text of the documents are subsequently indexed in ElasticSearch, an open source search engine. The entities are stored as uncontrolled entities, i.e., they are not matched to thesaurus entries at the moment.

Currently, pages are used as the index unit, which means that if all the search terms occur on one page of a document, that document is returned as a relevant result. Indexing per page is not ideal, as some query terms might be split over multiple pages. It would be better to index per section, which is planned for future work.

An online user interface (see Fig. 8.1) can be used to query the free text, detected artefact, context and species entities, and the year ranges converted from time period entities. Results are displayed on screen, or can be exported to a Comma Separated Values (CSV) or GeoJSON file for further analysis.

8.2.2 Search Process for our Case Study

To assess the usefulness of AGNES, a case study of Early Medieval cremations in the Netherlands was conducted, as introduced in section 8.2.1. To document the search process and results, the following process was used:

- The information need for this topic is defined based on prior knowledge
- A free search session using AGNES is conducted
- All entered queries are stored
- Results of the queries are stored in CSV format for further analysis, duplicate documents appearing in multiple result sets are removed

Once this was completed, we manually assessed the relevance of the retrieved documents and whether or not the result is already known from an earlier survey. This was done with the CSV export, which also contains links to page previews and the full document, to allow for detailed checking. This CSV approach for assessing the documents was preferred over using the search system's interface as there is a lot of overlap between the result sets of the different queries, thus reducing the number of documents that need to be assessed.

8.2.3 Evaluation: Comparison to Existing Knowledge

To assess the usefulness of the results, the AGNES data was compared to an earlier survey of Early Medieval cremations in the Netherlands (Lippok, 2020). That survey yielded 77 sites, based on comprehensive knowledge of the Early Middle Ages, but excluding grey literature. The reference database was compiled by going through all published, and some unpublished Early Medieval cemetery catalogi, comprehensive overview works of Early Medieval archaeology, master and PhD theses and other relevant material, such as site visits. The reference database is built on the career-long effort by Prof. Frans Theuws to make an overview of Early Medieval sites and has been in the making, specifically for cremations, for three years.

The comparison between the AGNES data and earlier survey data was made through assessing the relevance of the AGNES results, checking for the right time period, and if a cremation was actually found. After deselection of irrelevant records, the relevant AGNES records were cross checked with the list of known sites. A site is considered the same if they either have the same project name, or the same geographical coordinates.

8.3 Results

In this section, we describe the information needs and queries for the case study, show the results retrieved by AGNES, and compare these to existing knowledge.

8.3.1 Information Needs and Queries

An information need can be defined as a user's end goal in a specific search session (Hjørland, 1997). For this case study, the information need is as follows: to find all mentions of Early Medieval cremations in grey literature, with the Early Middle ages being defined as the period 450-900 CE. In the search session, this resulted in the queries listed in Table 8.1. These queries have been thought of and constructed by the expert in this topic in a free search environment, without technical help, for a fair comparison to previous surveys. Synonyms for "cremation" are based on the expert's knowledge, no archaeological thesaurus was consulted.

The start and end year are entered in number fields (see Figure 8.1), and are used to search through detected time period entities translated to absolute year ranges. The free text field was used to search for cremations, the entity search was not useful for these particular queries. The asterisks in the query column

Start Year	End Year	Free Text Query	English Translation	Number of documents retrieved
450	900	crematie	cremation	614
450	900	crematie*	cremation*	2335
450	900	verbrand menselijk bot	burnt human bone	24
450	900	brandstapel	pyre	73
450	900	brandstapel*	pyre*	84
450	900	urn	urn	508
450	900	knochenlager	bone bed	1
450	900	beendernest	bone bed	2
450	900	brandgrube	a pit containing pyre and cremation remains, covered by soil	8
			TOTAL	3035

Table 8.1: All nine queries used to retrieve results, in the order in which they were issued. An English translation is given for Dutch terms. Asterisks (*) are wildcards.

Type	Quantity
Relevant (Early Medieval cremation occurs in report) - known	31
Relevant (Early Medieval cremation occurs in report) - unknown	23
Possibly relevant (period or occurrence of cremation not explicit)	41
Not relevant (Early Medieval cremation does not occur in report)	2446
Total	2541

Table 8.2: Overview of relevant, irrelevant and possibly relevant results. Relevant results are divided into previously known and unknown sites.

denote wildcards, meaning they match zero or more characters appended to the search term. So for example *crematie** will also match *crematieresten* (cremation remains), leading to more possible results. There are two German terms in the query column, *knochenlager* and *brandgrube*. This is because these German phrases are used interchangeably with their Dutch translations in reports.

8.3.2 Retrieved Documents

When the results of these queries are combined and duplicate documents are removed, this leads to a total number of 2541 retrieved documents. The documents are ranked based on the free text query, the year range search is a boolean filter and as such does not influence the ranking For a full list of the results, please see the Zenodo repository[1] containing all data associated with this study (Brandsen & Lippok, 2021). It took one person about 40 h to go through the list and mark the relevance for each document. Out of all the results, 54 documents are relevant to the information need, 41 documents are potentially relevant but unclear from the text, and 2446 documents are not relevant (see Table 8.2).

The large number of irrelevant results leads to a low precision of only 2.1%, with precision being defined as the fraction of relevant documents among the retrieved documents (Powers, 2011). While the precision is low due to the large percentage of irrelevant results, this is not uncommon in systematic review studies. An example is the research by Bramer *et al.* (2013) in the biomedical domain, with a precision of 1.9%, very similar to this study.

While having a higher precision would be useful, as it shortens the time needed to check the results, these kinds of tasks are recall-oriented: having as many relevant results as possible is more important than having a small number of irrelevant results. This has been documented for archaeologists specifically (Brandsen *et al.*, 2019) and professional search more generally (Russell-Rose *et al.*, 2018; Verberne *et al.*, 2019). Unfortunately, the total number of relevant documents in the data set is not known, and therefore the recall can not be calculated. However, it is worth noting that out of the currently known 77 sites, AGNES has found 31, plus an additional 23 unknown sites.

When the irrelevant results are inspected (Table 8.3), it shows the vast majority are due to wrongly identified time periods (number 1 and 4), which can be attributed to NER errors. The other problems are mainly caused by specific types of sections in archaeological reports: lists of abbreviations and time periods, literature lists, etc (numbers 2, 3, 5, 8 and 9). Even though these pages

[1] Available at doi.org/10.5281/zenodo.3758085.

Number	Type of error / type of irrelevant document	Quantity	Percentage
1	Wrong time period	1742	71%
2	Page listing abbreviations	235	10%
3	Page containing research plan (*Plan van Aanpak*[2])	198	8%
4	Unknown time period	122	5%
5	Page containing list of time periods	85	4%
6	Negation ("no cremation")	22	1%
7	Other	21	1%
8	Literature list	18	1%
9	Coring chart	3	0%

Table 8.3: Overview of the different categories of irrelevant results. Percentages are rounded to whole numbers.

will contain the correct search terms, they are always irrelevant as they do not describe an excavated cremation. A possible solution to this problem is described in section 8.4.3. Negations were expected to be a substantial problem, but with only 22 errors this does not seem to be the case for this information need.

8.3.3 Comparison

The results of AGNES were compared with a database containing all Early Medieval cremations known from the earlier survey published by (Lippok, 2020). AGNES found 31 of the 77 known sites, and an additional 23 previously unknown sites containing Early Medieval cremations. 75 of the known sites were originally published in books and PhD theses, and two in excavation reports. However, the latter were published after 2017, which is after the data export from DANS, and as such can not occur in the AGNES results. In that sense it is surprising that AGNES found 31 of these sites published in books and PhD theses, as these are not included in the AGNES dataset. We assume these sites are mentioned in desk-based research reports that used the books as sources.

While 23 new sites might seem like a small number, this is a 30% increase over the existing knowledge. In the last 20 years, only 18 new Early Medieval

[2]A legal requirement in Dutch archaeology, the *Plan van Aanpak* describes the planned research methods.

cemeteries have been discovered in the Netherlands, and only nine sites containing cremation burials. With that as an indication of these site's scarcity, being able to add 23 sites that contain cremation burials is a major development.

The site of Hilvarenbeek is an excellent illustration of the added value of AGNES, it is a site completely missed by the earlier survey (Lippok, 2020), yet it yielded C14 dated cremated remains dating between 550 and 620 CE (Claeys *et al.*, 2012). Whilst this is an exciting find, it is unique in its novelty as it is the only cremation that was newly excavated and reported on in our list of 23. The other cremations are found in desk-research reports that scouted for archaeological sites surrounding the location of their study. Early Medieval cemeteries were most often mentioned there and are not often new finds. This means these sites are not newly excavated, and known to a few people, but practically impossible to find without AGNES.

Fig. 8.2 shows all the known and new sites, as well as all the sites found in the Dutch national cultural heritage database Archis for the search term "*Vroege Middeleeuwen*" (Early Middle Ages), to give an idea of the distribution of sites from this period. New sites containing cremations were located in areas known to yield cremations. Noord Brabant, Limburg, Overijssel, Zuid Holland, Gelderland, Drenthe, Friesland and Groningen have an additional one to six new sites. The sites of Castricum and Den Burg in North-Holland are located in a province that had up till now not yielded any Early Medieval cremations. Generally, Early Medieval sites are scarce in this province due to its geomorphological swampy nature in this period. A notable seven new sites containing cremations were added in the province of Drenthe, where previously only five were known.

Of the 23 new sites containing cremations, 16 are cemetery sites, one was found on a cemetery near a settlement and 6 were single cremations or urns, one of which probably indicates a larger cemetery. No additional context information about the other 5 single cremations was available.

8.4 Discussion

In this section, we discuss how the results from this case study affect our view on Early Medieval burial practices, the overall potential of AGNES, and suggest future work.

8.4.1 Archaeological Significance

Finding an additional 23 Early Medieval sites containing cremation graves draws attention to this type of burial. Previously, cremations have been neglected in

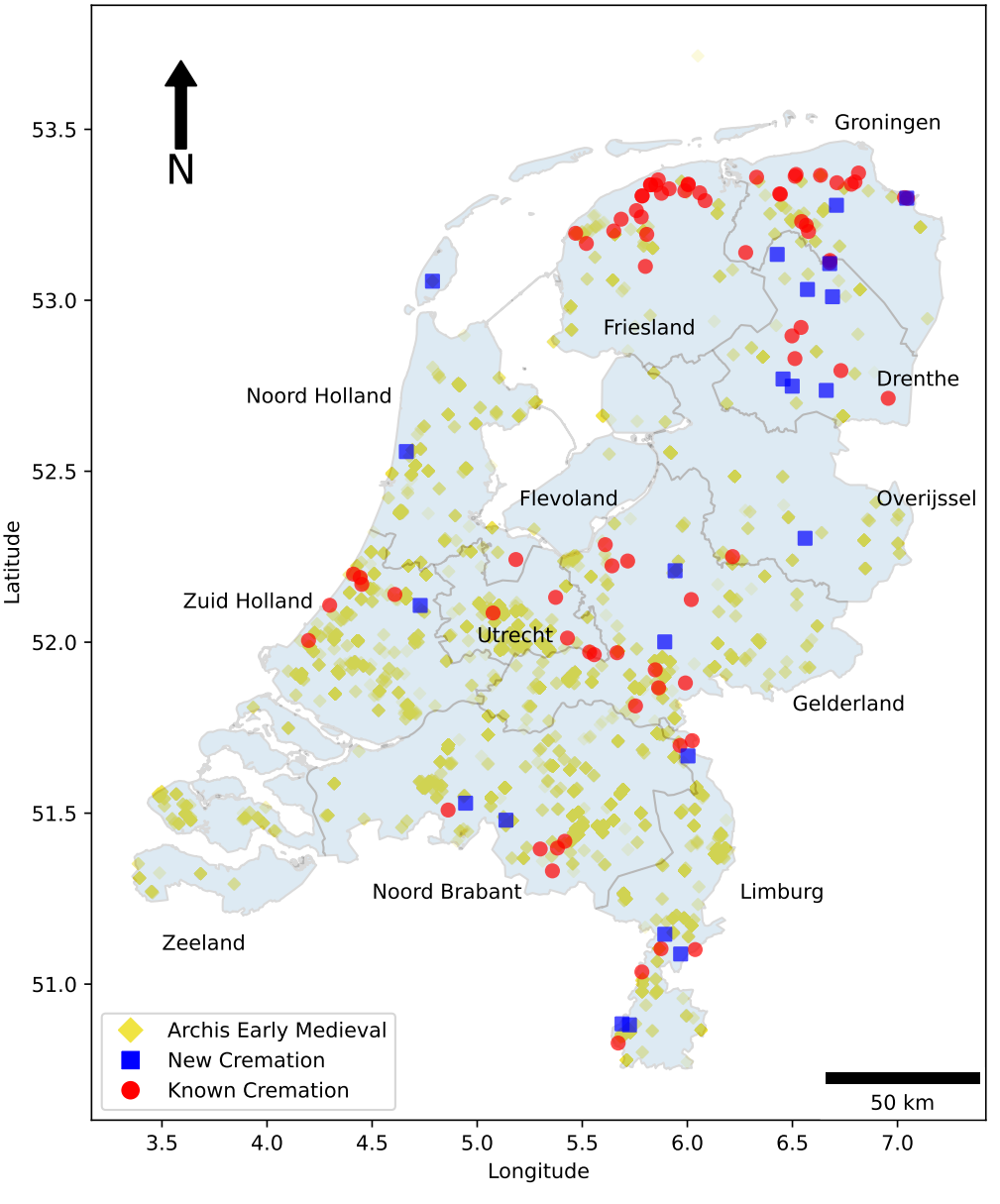


Figure 8.2: Map of The Netherlands showing known sites (red circles) and previously unknown sites found with AGNES (blue squares). Yellow diamonds indicate known Early Medieval sites (with or without cremations) as recorded in the Archis system. Province names marked in black.

Early Medieval mortuary studies, their obscurity is emphasised, or they are altogether left out of discussions on mortuary rites (Lippok, 2020). With renewed interest in cremations, they are not only shown to be more prevalent than before, additional information on the occurrence of cremation burials has come to light. Although the 23 new sites do not reflect this, previous research shows that Early Medieval cremations occur not only in cemeteries, but in settlements too, such as in Oegstgeest and Utrecht-Leidsche Rijn (Lippok, 2019, 2021). Because inhumations also occur in settlements, new questions concerning relations with the dead and their place in the landscape arise. Instead of being restricted to the cemeteries, away from the settlements, a more interactive relationship with the dead might be envisioned. The idea of interaction with the dead fits with the practice of re-opening graves, which consists of targeted retrieval of particular items and/or bones (Van Haperen, 2017). Cremations may have been kept above ground in containers and may have circulated for days or even years before being interred, as is suggested by Williams (2014) for Anglo-Saxon cremations.

In terms of geographical distribution, there are trends in the occurrence of cremations. The cremations in the north of the Netherlands, in the provinces Friesland, Groningen and Drenthe, conform to the prevalent known area of their occurrence. Both these and the cremations around the central Dutch rivers are often said to be restricted to these areas (Van Es & Schoen, 2008). This is accompanied by a heavily critiqued interpretative framework that emphasises ethnicity and religion as the main factor in the choice of burial form. Because of their seemingly incidental nature, cremations were thought to be of 'pagan' immigrants from the North, taking their 'traditional' burial practices with them (e.g. Hombert, 1950; Wamers, 2015). As for the more southern graves, they could fit in with the occurrence of cremation burials in and around the Scheldt area as described by Annaert (2018). Currently, Early Medieval cremations are also explained as belonging to immigrants in that area, with subsequent acculturation to account for cremations in later phases. However, through case studies of Lent and Elst in the Netherlands (Hendriks, 2013; Verwers & van Tent, 2015), Grobbendonk in Belgium (Janssens & Roosens, 1963), and LÃijnen in Germany (Lehnemann, 2008) it became apparent that the cremation and inhumation burials on these sites, and in a broader early medieval context, may be seen as more similar than was assumed before. Material culture in both types of graves are comparable and in the case of Grobbendonk even identical when looking at types of pottery and decoration (Lippok, 2020). A case can be made for burial communities that had both burial methods at their disposal. The ethnic and religious explanations provided by previous research seem less likely, as they oppose the two types of burial in an interpretive sense whereas cremations and inhumations

have been shown as comparable (Lippok, 2020).

Finding 23 sites boosts the numbers of cremations, making it more compelling to argue that they should be included in considerations on the Early Medieval mortuary framework where they currently are neglected (cf. Lippok, 2020). The occurrence of cremations actually helps us rethink Early Medieval burial practices. By incorporating the cremation rite in the Early Medieval mortuary framework, a more accurate portrayal of the archaeological reality is provided, and can therefore also give a more legitimate account of this time period.

8.4.2 Potential of AGNES for Archaeological Research

The results described in the previous section show that AGNES can be useful for synthesising archaeological research. It is not intended as a replacement for the currently available search tools, but more as an additional method to gather data from hitherto underused sources. Of course AGNES is limited in that it can so far only search through reports deposited at DANS, and as such can not yet find information such as metal detectorist finds stored in the Archis system, or information from journals and books. What AGNES excels in though, is finding the by-catch of excavations often not mentioned in books and articles, and finding specialised or uncommon finds and contexts not mentioned in the metadata. This is illustrated by the fact that none of the new sites are returned when searching for "crematie vroege middeleeuwen" (cremation Early Middle Ages) in e.g. the DANS archive. These cremations are a good example of information that can relatively easily be found using AGNES, but would require manually searching through the entire collection of documents using the previously available systems.

While this case study has a very specific topic, AGNES can be used just as easily for other research questions. It is possible to combine any artefact, context, and species entity search with a free text search, a year range search and even geographical search (by drawing areas on a map). There are however a couple of limitations in the current system: it is not possible to search on multiple year ranges at once, and there is no controlled vocabulary for the entity search. Both these issues will be tackled in a follow up project.

Besides being used for research questions in the Netherlands in Dutch documents, the system can also be adapted to other regions and languages (further described in the next section). The types of research questions are fairly similar across countries, most of them deal with What, Where and When questions (Jeffrey *et al.*, 2009), and this would make it relatively easy to adapt to other regions as the entity types remain the same. The only prerequisites are labelled training data for the NER algorithm, and a region/language specific time period

thesaurus to convert entities to year ranges.

Our work is similar to that of the STAR project (Tudhope *et al.*, 2011), which shows the need for these kinds of tools in other regions. Although they focused more on more detailed metadata generation and AGNES focuses more on full text search combined with entity search, both projects used similar methods to find the What, Where and When aspects of grey literature. The STAR project also mapped entities to a thesaurus, leading to a controlled vocabulary of entities. This makes sense for their goal of interoperability, but in a free search scenario uncontrolled entities are more useful for users. However, we do plan to map entities to thesauri where possible in a follow-up project.

The results from this case study are promising, and help change long held views and biases resulting from an underrepresentation of cremations in published literature. We are optimistic that AGNES can help with other archaeological information needs as well, and will hopefully lead to a better understanding of the past.

8.4.3 Future Work

However, in conducting this case study, some areas that could use improvement were found. The main issue is that while the relevant results are very useful, a relatively large amount of irrelevant results are also returned, making the checking process fairly time consuming. However, this is still much less time consuming than searching through the document collection manually, and similar amounts of irrelevant results are found in other systematic reviews. Around 70% of irrelevant results are due to the wrong time period being identified, an error propagation from the NER process. To solve this, we need to invest more effort into making the NER process. To solve this, more effort needs to be invested into making the NER process more accurate, specifically for the time period entities. Experiments with newer architectures such as RoBERTa (Liu *et al.*, 2019) and LUKE (Yamada *et al.*, 2020), as well as increasing the unlabelled training data, might increase the performance of the NER, and thereby decrease wrongly identified time periods. Our time period entity to year range conversion module will also need to be further tested and refined to further decrease false positives.

The second most common type of irrelevant document is due to specific sections that are not useful for searching, such as abbreviation, time period, and literature lists. A way to solve this would be to automatically detect these types of pages after the NER process, and either give terms on these pages a lower weight in the ranking algorithm, or to avoid indexing them altogether. This type of injection of domain or situational knowledge is already being successfully ap-

plied in automated detection in remotely sensed data (Verschoof-Van Der Vaart *et al.*, 2020), and should improve results here too.

As mentioned previously, the current result evaluation process using the CSV export is time consuming, and as such it would be worth experimenting with ways that allow for interactive results checking. A solution would be a system where a query – or group of queries – can be saved in the online environment, which can then be further explored with links to page previews, and a method of marking results as relevant or not. After this process, the user can download a CSV export of just the relevant results. This should streamline the process.

Currently, the NER BERT model is only able to handle Dutch texts, but it is relatively easy to train a model for other languages. This would require annotated texts to train on, which can be produced to a sufficient quantity and quality in about 90 h of annotation (Brandsen *et al.*, 2020). The actual training of the BERT language model would take about 24 h on 4 GPUs, and training the NER model another 4 h. In a follow up project, the system will be expanded to also handle English, German and possibly French texts, as well as diversifying the type of documents: including papers, books and theses, among others.

Besides these technical improvements, the archaeological side of this study also warrants further research. The 41 sites classified as 'possibly relevant' consist of sites where either the Early Medieval date is in question, or the occurrence of cremation is not explicit. The excavations at Park Leeuwenstein in Geldermalsen, for example, yielded cremations and inhumation dating from the Iron Age to the Carolingian period. More information is needed to assess if the cremations are Early Medieval. It is encouraging that verified Early Medieval cremations were found closeby, at Geldermalsen, Meteren de Plantage.This suggests that at least one of the 41 possible relevant sites contained another Early Medieval cremation. Given the small number of sites, it would be prudent to research all 41 sites to take away the doubt over their usefulness. Verification would involve a literature search on those sites and possibly contacting the authors of the report to ask for clarification.

8.5 Conclusions

In total, 23 additional sites containing Early Medieval cremations were found, when compared to a previous survey (Lippok, 2020). This is a 30% increase on the total number of known sites before the study, and more than double the number of Early Medieval cemeteries discovered in the last 20 years.

The amount of information found is promising, but with a precision of 2.1%

there are a large amount of irrelevant results that need to be manually assessed. However, this precision is similar to other systematic review studies, and the total amount of hours spent on assessing (40 h) is much lower than it would be with other systems. More importantly, the time spent querying and assessing was deemed acceptable for the amount of information gained.

The additional 23 sites containing cremation graves further strengthens the importance of a heterogeneous perspective on Early Medieval burial repertoires. In the past, furnished inhumation graves were afforded most scholarly attention. The increasing number of sites containing cremation burials from the Early Middle Ages attests to a more heterogeneous burial repertoire. To understand Early Medieval communities, it is necessary to account for all of their burial practices. Understanding the occurrence of cremation practices will aid answering questions on heterogeneity, burial communities and change therein.

While some work needs to be done to further improve AGNES, the results presented in this case study are relevant and substantial, and the potential of the system seems promising for other information needs. We are confident that AGNES can become a useful tool to add to the archaeologists' searching toolbox, leading to more efficient and more detailed research.

9 Discussion

"*Computers are useless. They can only give you answers.*"
Pablo Picasso

In a way, Picasso was correct: computers do indeed only provide answers. And what use is an answer without a good question? To be able to formulate a useful question, you need creative thinking, innovation, and new ideas. Without this, computers are indeed useless. This is also excellently illustrated in The Hitchhiker's Guide to the Galaxy (Adams, 1979), where a supercomputer takes 7.5 million years to calculate the answer to the "Ultimate Question of Life, the Universe, and Everything", with the answer being 42, a seemingly meaningless number. When asked to produce the Ultimate Question, the computer replies that it can not.

This is also the reason that this research – and other research on artificial intelligence in archaeology – is not going to replace the archaeologist, as computers are (currently) not able to do research from start to finish. This is not something we want either, the combination of processing by a computer and interpretation by a human is what fuels research and provides accountability. Instead, computational tools are meant to further enhance the archaeologist's ability to draw meaningful conclusions from raw data, and to make this process more efficient. Outsourcing menial tasks to e.g. students and volunteers has a long history in archaeology, and science as a whole. The more we can replace this valuable human time with relatively unvaluable computing time, the more we can focus on the interesting parts of archaeology: drawing conclusions and building theories relating to past human behaviour.

In the rest of this chapter, we discuss AGNES in the context of development-led archaeology, synthesising research and Big Data (Sections 9.1 to 9.5), the advantages of less complex methods over computationally heavy models (Section 9.6) and provide some thoughts on evaluation metrics (Section 9.7) and FAIR data (Section 9.4). We then provide some concluding remarks (Section 9.8), and end with ideas for future research.

9.1 Development-led Archaeology and the Role of AGNES

Throughout this research, the point has been made that the number of archaeological documents available in the Netherlands is simply too large for manual inspection. And the reason we have so many documents is mainly due to the Malta Convention (or Valetta Treaty), as discussed in chapter 2. Although reports were created – and to a lesser extent deposited in archives – before, there has been an explosion in the amount of research done after 2007. This development-led

work is mainly done by commercial archaeology units, who due to stiff competition, time constraints, and lack of available funding, might spend the minimum amount of time necessary to produce results that adhere to guidelines, but do not go beyond those guidelines.

While the information uncovered in excavations and other research is often very valuable, the reports describing this information can be seen as a checkbox exercise: a report must be produced, but money might be running out, so the minimum amount of time is spent to produce the report, in an attempt to maximise profit (or in some cases, minimise losses). While this is better than no rescue archaeology at all, the decline in quality due to a clear capitalist rescue archaeological regime is illustrated by the research of Plets *et al.* (2021), who analysed over 4,500 texts from the Dutch speaking parts of Belgium. They show that widespread boilerplate templates and a decrease in complex vocabulary indicates a decrease in quality over time. Also in the Netherlands, the research by Bazelmans *et al.* (2005) shows that only about half of the reports they examined were deemed of sufficient quality. While more sites are excavated – leading to a raw data increase – the relatively low quality of (a portion of) the texts calls into question if the highly competitive development-led research actually leads to an information gain.

Another aspect that possibly contributes to this problem is the perception that the reports are not read and used much, if at all (Habermehl, 2019). This perception makes it feel like making a better report is a waste of time, as nobody is going to read it. And this in turn gives rise to the perception that the reports are low quality and not worth reading, causing a negative feedback loop. While AGNES can not hope to solve the problems surrounding development-led research, this issue of perceived low quality and unwillingness to create better reports is something we can help improve. By increasing the accessibility and findability (as introduced in Section 2.2.5), researchers will more easily be able to find relevant sections for their research (and filter out irrelevant sections), increasing their perceived value of the information available in the corpus. And this increase in usage will hopefully lead the report authors to more carefully consider their writing, as the report is something that can actually have a contribution to research, and is not just a deliverable needed to finish a project.

9.2 Catching the By-Catch

We have already mentioned the 'by-catch' in previous chapters: single or small groups of finds that are dissimilar to the rest of the excavation, things that are

found when looking for – and expecting – other things. Single finds are often seen as less or not important, as such a singular data point says very little about past human behaviour. And as such, these finds are often not given a lot of attention, especially in commercial archaeology due to financial and time constraints. Some examples of by-catch that can be missed completely are the mesolithic sites found in the topsoil that is normally removed by machine (Evans *et al.*, 2014) and Bronze Age metalwork in contexts not normally investigated, and only found by metal detector survey (Bradley *et al.*, 2016). The find concentrations are low, perhaps perceived as not worth studying, and in contexts we do not expect, making them hard to find.

While it is true that such single data points are not very informative, when these data points are combined, patterns emerge that can be very informative. And it is exactly this by-catch that is near impossible to find and study without AGNES. When we look at the research on Early Medieval cremations in Chapter 8, we see that roughly 30% of the cremations we found with AGNES were indeed by-catch in some form: cremations found outside cemeteries, as a singular find within a larger homogeneous context. And we see that all the previously unknown sites are not returned when searching for the term "Early Medieval cremation" in the currently available systems, indicating the strength of AGNES.

And perhaps that is the strength of development-led archaeology: a much more random sample of excavations when compared to targeted academic research, at a much larger scale. As building work occurs just about anywhere, we are finding things in places we did not expect. This more random sampling of past human behaviour allows us to challenge existing ideas and overcome confirmation bias. Searching for particular phenomena in places where we have previously found them is useful for gathering more data, but this data will inevitably be similar to previously gathered data, further entrenching existing ideas. As we have seen in our case study, it is exactly the by-catch that can change our views on the past.

However, for all of this to work, we do have two prerequisites: the information we are looking for needs to be written down in the publications, and we need to be able to find this information. Hopefully, the by-catch is described adequately in reports, and AGNES makes it possible to find and extract the information.

It is worth noting here that while development-led archaeology is more random than targeted research, there is still a bias within the sampling: not all areas are equally often disturbed by building work, and some areas do not see any soil disturbance at all, such as rivers, lakes and protected nature reserves. As noted by Bradley *et al.* (2016) in the UK and Eerden *et al.* (2017) in the Netherlands, certain regions and site types are still underrepresented, and this should be taken

into account when doing synthesising research.

This might also be related to Wheatley's view that correlative predictive modelling is not very useful (Wheatley, 2004). If the data that the model uses to predict archaeology is biased, it will replicate the bias in its predictions. More randomly sampled data might possibly make predictive modelling more accurate.

9.3 Synthesising Research

Without synthesising research, the information in archaeological reports are individual data points with no real use. We need research connecting all the (small and large) dots we have as archaeologists, to create narratives at a larger scale. And for synthesising research to be done, the information must be easily accessible, as it is vital to any understanding of the past.

Most synthesising research is done in the academic sphere, with a notable exception being the work undertaken by the RCE at a governmental level. Here we see that while researchers want to use the reports, they often do not, or do so only to a limited degree, as accessing and finding relevant information is too difficult and time consuming. And if reports are used extensively, this often means that (mainly) early career researchers carry the burden of manually searching through the literature, spending extensive amounts of time and effort to gather data, like in the research by Fokkens *et al.* (2016). As we mentioned at the start of this chapter, these kinds of monotonous and time consuming tasks are exactly the kind of things we should aim to speed up by using computational approaches, leaving more time for actual analysis. This will hopefully lead to more in-depth interpretations, but could also help prevent the common occurrence of projects (especially PhD research) taking longer than expected.

Besides academic research, we would like to mention the *Oogst van Malta* (Valetta Harvest) project, led by the RCE. This project is specifically aimed at extracting new insights from the wealth of information generated by development-led research, and to re-evaluate, homogenise and digitise old data. Up until this point, the research carried out in this project followed almost the same process as most academic research: a pre-selection is made of reports that seem relevant based on metadata, and subsequently this entire pre-selection is read manually and assessed for relevance, after which the analysis can begin. This process is both inaccurate (as the metadata is inaccurate) and time consuming, making these studies very costly and slow moving. Again, computational approaches to speed up and increase accuracy are very much needed to improve this kind of research, and AGNES can help with this.

Besides finding reports about certain topics, the information we extracted from the entire corpus can be used to identify subjects that have ample information for synthesising research. Specifically, the *Nationale Onderzoeksagenda Archeologie* (NOaA), or National Archaeological Research Agenda of the Netherlands, provides a list of research questions currently unanswered (Abrahamse *et al.*, 2017), an example being question number 45 about the changing nature of burial practices, to which we contributed in Chapter 8. The NOaA research question list could be cross-referenced with the information found in the reports, to see which research questions would be most suitable for study.

9.4 MEAN & FAIR Data

In the background chapter we introduced the FAIR principles (Findability, Accessibility, Interoperability, Reusability), which aim to increase re-use of data through making it available, findable and standardised. However, archaeological data tends to be Miscellaneous, Exceptional, Arbitrary, Nonconformist (MEAN), which makes it complicated to archive, digest and re-use (Huvila, 2017), certainly when compared to other disciplines. There are major differences in how data is used and created in archaeology when compared to e.g many science, technology and medical domains. But that does not necessarily mean that we should not try, or that archaeology could not be FAIR in its own terms.

We certainly see that in this project, the data we are working with is very nonconformist and miscellaneous: there are large differences in the structure, format, and quality of the texts, but also in the words used to describe objects and phenomena. This is in contrast to other disciplines such as the biomedical domain, where most literature is published in similar controlled formats (journal articles) and there is much less variation in descriptions, as categories such as drug names, diseases, proteins and chemicals have a much more controlled vocabulary. Due to the 'messiness' of archaeological text, using machine learning to normalise concepts, extract information, and subsequently (re-)publishing this data in a machine readable controlled format substantially increases the FAIRness of the information stored in texts. And this is what we aimed to do in this project: extract relevant entities from text and map them to a controlled vocabulary, and then publishing that data as JSON which can easily be used for other computational approaches.

Interestingly, Huvila (2017) argues that the focus in archaeological information management should not be on "discipline-wide naming of entities and following a shared agenda of explicating interactions between these named entities"

(Huvila, 2017, p. 1), but more on the interactions between creators and users of archaeological data. However, the identification of named entities in this research has increased the FAIRness of the data contained in our corpus, at the very least the Findability. And while we have not researched the interactions between named entities, it is likely that interesting patterns can be found, like in the research by Wilcke *et al.* (2019, also see section 9.9.2).

At the same time, using machine learning does introduce noise through incorrect predictions. This means that while we make data more FAIR, the data also becomes less accurate and more incomplete to some extent. We see this trade-off as unfortunate, but unavoidable, as machine learning (but also manual entry by humans) is never 100% accurate. At the same time, going through the big data we have access to right now by hand is completely unfeasible, and computational methods – even if they are not perfect – are needed to process and analyse the data and make sense of the information we are generating, and use it for synthesising research.

9.5 Taming Big Data

In Chapter 2 we introduced the concept of big data: data having high volume, velocity, variety, and veracity. While in general, archaeological data is relatively small when compared to other disciplines, our corpus definitely falls into the category of big data, as it is over a terabyte in volume, can not be analysed effectively using traditional tools, has a reasonable velocity of over 4,000 reports being added each year, and high variety (as described in the previous section).

This project has been all about making this big data more manageable. By leveraging machine learning and information retrieval techniques, we make it possible to select a portion of the data for further (manual) analysis. We are using computer power to select a subset of the data to focus our efforts on, which would not be feasible to do by hand. We also aim to reduce the variety of the data, which is closely linked to making the data more FAIR: by grouping, disambiguating and interpreting entities in text, it is possible to navigate a heterogeneous mass of data with uncomplicated queries. And although we did not address the velocity of the data in this dissertation, in a follow-up project we will automatically index new documents from a variety of sources (see Section 9.9.1).

Regarding veracity, we certainly encountered problems with completeness and quality. Some examples include OCR and PDF conversion errors creating noise in our texts, and ontologies with varying degrees of accuracy and completeness. Again, in this project we did not explicitly attempt to deal with data quality,

but some of the goals of the follow-up project include creating more complete (multilingual) ontologies and improving the quality of the PDF to text conversion.

More and more archaeological data sets grow so large they become hard to analyse, and efforts such as the ARIADNEplus project to combine data sets across regions and countries will make for even bigger composite data sets (Niccolucci & Richards, 2019). And as these data sets keep getting bigger and more complex, we will need to keep developing new methods to wrangle useful information and patterns out of this big data. We already see many developments in object detection in remotely sensed data such as LiDAR (e.g. Verschoof-Van Der Vaart *et al.*, 2020), but other sources of data currently seem to not get the same level of attention. Of course, other disciplines have been dealing with similar problems, and just like we mix and adapt methods from, e.g. robotics for computer vision in LiDAR, we can similarly look to fields with a high volume of texts (such as biomedical science) for inspiration on how to handle this data.

So while big data can form a problem, we can often leverage computational approaches to make our data small enough to work with and analyse.

9.6 The Problem with Complexity

As *The Zen of Python* states: "Simple is better than complex" (Peters, 2004). This is certainly true for programming code, but also for research in general. If a less complex method produces similar results to more complex methods, it would be preferable to use the former.

We saw that in Chapter 4, the document classification task was performed optimally by the least complex method we tried: the linear SVM model which is commonly used for these types of tasks. While SVMs are a bit more complex than say a logistic regression, they are relatively light-weight to train when compared to newer transformer-based models such as BERT. As we were training many models in that study, using a less complex method meant this was possible to do in a reasonable time scale.

In general, less complex methods have many advantages: being easier to use, less computationally expensive, and generally more explainable. It is therefore always wise to assess these methods before trying more complex techniques, even though these complex methods might be more appealing to put in a paper title.

However, in the case of Chapter 7, we tested a less complex method (CRF) against a more complex method (BERT) and came to the conclusion that BERT substantially outperformed CRF. In this case, we found that the added complexity was worth the increased performance, but we did run into some issues: the

prediction of entities on the entire corpus by the BERT model took over nine days running on ten GPUs simultaneously. Luckily, this process only needs to be done once, and afterwards any new documents can be added in small batches that will take much less time.

This kind of prolonged use of GPUs for the training and use of Deep Learning models has recently come under scrutiny from another angle: the environmental impact of the power used by these machines. Strubell *et al.* (2020) investigated the CO2 output of training BERT models, and found that pretraining a single BERT model produced the same amount of CO2 as a transatlantic flight. But of course, in most research, multiple – sometimes hundreds – of models are trained to test different data and perform hyperparameter optimisation. Until electricity is fully renewable and CO2 neutral, using this level of power should be carefully considered: does the increase in performance weigh up against the environmental impact?

So to conclude this section, less complex methods should be compared to more computationally expensive methods in the experimentation phase. Only when the performance is substantially better and needed for the application, should computationally expensive methods be used in production systems.

9.7 Evaluation Metrics

In this dissertation, we have introduced and used a variety of evaluation metrics. In general, we have used the metrics that are considered the standard for a particular task, as these are easily comparable to other studies and are often well researched. For NER, we use precision, recall, and the F1 score. In general, it is worth assessing metrics and the calculation of these metrics, to see if they fit in with the goals of the research being done.

Unfortunately, we do see some studies in the archaeology domain where non-standard, non-optimal, or non-reproducible metrics are used for machine learning evaluation. This is mainly seen in the automated detection of features in remotely sensed data, as also discussed by Verschoof-van der Vaart & Landauer (2021). This makes comparing different methods difficult as different measures produce different results.

On the other hand, we have also deviated from standards. In Chapter 3 we evaluated the Inter Annotator Agreement between a group of human annotators. While the standard for IAA is Cohen's Kappa, we found this metric suboptimal, as it needs the number of negative cases, and this is not available for NER. Instead, we used the pairwise F1 score between all annotators, which led to a

more interpretable result.

Hindsight is always 20/20, and as we looked back at the research, we realised that perhaps the F1 score – although adequate and easily comparable – might not have been the optimal metric for our research on NER and IR. This is due to the fact that archaeologists' information needs are most often recall-oriented list questions, meaning recall is more important than precision. To take this into account, perhaps the F2 score would have been more suitable for this research, as this metric considers recall more important than precision. This point is also raised by Hand & Christen (2018) more generally speaking, who argue that the popularity of the F1 score means that this is often used without considering the relative importance of precision and recall, which is an aspect that should be considered when trying to solve a task.

Something related to this is that we can use the F2 (or the precision oriented F0.5) for evaluation after training a classifier, but the classifier itself by default is most likely to optimise on F1 score. This means that the choice of F measure should ideally be decided before any classification is done. So to conclude, it is worth carefully considering evaluation metrics before using them. Balance the comparability of a standard with the specific characteristics of each task, and choose a metric accordingly. However, even with a suitable metric chosen, this does not necessarily mean that it accurately reflects the usability in a real use case. As such, qualitative evaluation (as we did in Chapter 8) should be used in tandem with a quantitative metric.

9.8 Conclusion

Chapters 3 to 8 each covered different aspects of this research, and as such have their own sets of research questions. In the following section the main question from each chapter is answered and discussed.

9.8.1 Answers to Research Questions

Can we use existing labelled data sets for NER in the archaeological domain, or do we need to create our own data set? If so, to what extent does the accuracy increase?

In chapter 3 we discussed the problems we encountered with an existing data set for NER, and how we created new training data. The new data set showed an increase in F1 score of 0.19, from 0.51 to 0.70. This indicates that the previous training data was not optimal, and also shows the importance of rigorous

annotation guidelines and checking of the data afterwards.

In this case, we monitored the quality of the data by having each annotator label the same section of text, and calculated the Inter Annotator Agreement. As the IAA was high (0.95), this is an indication of high quality data. However, as we show in Chapter 7, the data is not perfect and we did find some instances of incorrectly labelled entities. These were detected due to the BERT model correctly predicting the true label instead of the false annotated label, leading to true false positives. While algorithm and feature choices are important for the performance of a model, good quality data lies at the base of any model's performance, and should be tackled before model and feature optimisation.

To what extent can we automatically generate time period and site type metadata for Dutch excavation reports?

In chapter 4 we experimented with methods to automatically label reports on the time period and site type metadata fields. Despite the low quality of the texts and labels in the training data, we managed to obtain F1 scores of 0.752 and 0.542 for time periods and site types respectively.

These scores were obtained by using relatively light-weight methods: an SVM classifier with TF-IDF as features, with basic text pre-processing. Using more advanced methods such as BERTje led to substantially lower results, unlike our results from Chapter 7. This adds to the idea that often, light-weight baseline methods are hard to beat and have relatively high performance with none of the methodological and computational challenges that more advanced models have.

The methods developed are not currently used for the AGNES system, as the data set from DANS already contains metadata for the vast majority of reports. However, in the follow up project to this research (EXALT), we will index documents that often do not have information about time period and site type, for example reports from the KB, which only have standard metadata such as author and year of publication.

Which questions do archaeologists want to ask of this data set, and which user requirements do they have for a search system?

The user requirement solicitation study we describe in Chapter 5 aimed to map archaeologists' wishes for a literature search system, and evaluate a prototype of AGNES. It became clear that the currently available search systems are not adequate, and a more efficient and effective system was highly desirable.

Regarding more specific user requirements, we documented that there was a

strong need for geographic search across the user group, combined with keyword and time period search. This makes sense intuitively, as most archaeologists have questions relating to what, where, and when. We also found that in general, everyone preferred high recall over high precision, even if this means more work for the user evaluating the results.

Feedback on the AGNES prototype indicated that users are generally positive about the system, but work needed to be done on the usability of the front end.

How do Dutch archaeologists use search system interfaces, and what user interface features are experienced as positive or negative?

Based on the feedback we received in the user requirement study, we assessed the front end usability in Chapter 6. We found that overall, the front end is experienced as positive, but some work needed to be done to improve the user experience. We have since updated the front end based on these comments, leading to AGNES v2, the latest online version as of writing.

Nearly all the information needs we recorded in this study are list type questions where a complete list of documents for a particular query is requested. This is interesting, as this is not typical of scholars in the related field of humanities, who often have a mix of list, factoid and yes/no questions (Verberne *et al.*, 2016). This indicates that while archaeology generally can be seen as a humanities field, it does have particular ways of doing research that warrant investigating when building information systems for archaeologists.

We also found that the different categories of archaeologists (e.g. commercial and academic) flag different issues, based on their similar, but slightly differing ways of searching. This highlights that a diverse focus group is important to optimise the number of found usability issues.

To what extent does adding more domain-specific training data to BERT models improve Named Entity Recognition accuracy?

In Chapter 7 we investigated the use of BERT models for NER. We found that further fine-tuning a Dutch BERT model with domain-specific training data improves the model's performance by a large margin, larger than in related work addressing domain-specific BERT models.

We also experimented with ensemble methods of combining multiple BERT models or combining a BERT model with domain knowledge, but could not further improve the overall performance when compared with ArcheoBERTje. We did find higher precision with one of the ensembles, but as almost all informa-

tion needs of archaeologists are recall oriented, we opted for ArcheoBERTje for labelling the full collection. All the extracted entities are available in a DANS repository: doi.org/10.17026/dans-zcs-7b72.

What is the impact of the developed system on archaeological research?

Finally, we performed a case study on Early Medieval cremations in Chapter 8, to evaluate the usefulness of AGNES v2. When compared to a previous literature review and knowledge of experts in the field, we found 23 additional sites containing Early Medieval cremations. This is a 30% increase on the total number of known sites before the study, and more than double the number of sites discovered in the last 20 years.

This rediscovered information further strengthens the idea that the Early Medieval burial practices do not solely consist of inhumations, as previously thought in the field. The common view that only inhumations occurred actually created a bias where cremations in Early Medieval contexts are sometimes assumed to be from earlier periods, as they could not possibly be Early Medieval. The information found in this study helps to undo that bias, and provide a more accurate and heterogeneous view of the Early Medieval burial repertoire.

9.8.2 Answer to Problem Statement

Then finally, we are nearing the end of this dissertation. We started this research with the following research question:

To what extent can a search engine using Text Mining improve archaeological research and aid information discovery in grey literature data sets?

Over the course of this research, we investigated multiple aspects of AGNES, from initial user requirements to testing the system with a case study. But of course, the aspect that is most important for archaeological practice is to what extent the system can actually help us do better and more efficient research.

While some work needs to be done to further improve AGNES, we have seen that all the participants of the focus group responded positively to the system, and the case study on Early Medieval cremations shows that AGNES can provide substantial contributions to archaeological research, while also being more efficient than the previously available search systems.

Digging in documents is perhaps not as glamorous as digging in the ground, but as it is a integral part of archaeology, it is vital we invest time and effort into

improving this process, just as we do with excavations. We are confident that AGNES can help with this problem, leading to more efficient and more detailed research, and a better understanding of the past.

9.9 Future Research

The work presented in this dissertation, while being valuable in its own right, provides a base for further research. There are many new avenues and improvements we would like to explore to further strengthen the usefulness of AGNES. In the next section (9.9.1) we describe further research we will undertake in a follow-up project, and Section 9.9.2 describes ideas not currently in the pipeline, but which would make for interesting research. Finally, we describe some recommendations for future research on this topic and the lessons we learned during the project (Section 9.9.3).

9.9.1 EXALT

In 2020, the AGNES project team were awarded a grant in the 'Future directions in Dutch archaeological research' programme by NWO (*Nederlandse Organisatie voor Wetenschappelijk Onderzoek*, the Dutch research council), to further develop the research described in this dissertation. This new project is called EXcavating Archaeological LiTerature (EXALT), and will take place over four years. The main aims of EXALT are:

- While AGNES currently only gives access to field reports, we will include more archaeological text types (articles and books) from a wide range of additional sources.
- The system will be multilingual, to include documents in Dutch, English and German.
- We will make the novel step from full-text search to semantic search, allowing for searching through a collection of texts with meaning, as opposed to 'normal' text search where we only find literal matches for the search terms (lexical matching). The current entity search already does this to some extent, but we will further develop this.
- We will develop novel NLP methods to extract structured information from texts, building upon the state-of-the-art techniques but geared towards the archaeological domain. This entails the extraction of archaeological concepts and the relations between them. The identification of these concepts

facilitates semantic search, by allowing the mapping between a user's search query and the specific concepts in a document.

We currently have eight partners from four countries who will provide documents in Dutch, German, and English, totalling at least 100,000 documents, and will be adding more partners during the project. Other partners include commercial, academic, and government level archaeologists to function as a focus group, making sure the system is fit for purpose.

Perhaps the most obvious improvement is to include documents from other sources. To keep this Ph.D. project manageable in four years, we opted to only index reports from DANS. But of course there are many other types and sources of literature archaeologists would like to search through, including books, papers, reports from other repositories, and perhaps even other types of data such as numerical data (e.g. databases/spreadsheets) and images. In EXALT we will be integrating more sources into AGNES. This will pose new technical challenges, but will be very beneficial to archaeologists.

Related to this, an automated inflow of newly added documents from different sources would be very useful, as we currently work with a static dump of the DANS archive taken in 2017. We will automatically add new documents to the search engine, which means it stays updated, and would open up the possibility of saved queries for users: when a new document matches a saved query, the user is notified.

Another factor is language: AGNES is completely geared towards Dutch texts, and can not properly deal with texts in other languages. But of course much literature about the Netherlands and surrounding areas is written in English, and to a lesser extent German and French. Being able to integrate all these languages into one search engine would be beneficial to literature studies. For this to work, we would need to update and add a couple of components: a language detection module, NER models for each language (or possibly a multilingual model), and most importantly a mapping of concepts between languages, allowing cross-lingual search. In the EXALT project, we are primarily focusing on English and German, with the possibility of adding French later.

While we are mapping concepts between languages, we can also map relations between concepts, allowing for query broadening or narrowing. An example is "*beugelfibula*" (a type of fibula brooch). When searching for this term, a future version of AGNES could possibly suggest to search for "*fibula*" (the parent concept, broadening the query) or "*Domburgfibula*" (a child concept, narrowing the query), or add all of these concepts to the query for a broad search. Being able to find a group of related concepts like this, instead of having to manually remember

and enter all the terms, is very useful to archaeologists.

This query expansion can be done by looking up terms in a hierarchical ontology like we just described, but a less rigid and predefined way of doing this would by to use semantic similarity. We have introduced the BERT architecture for the NER task in Chapter 7, but these types of language models are also very effective at measuring similarities between terms and documents (Khattab & Zaharia, 2020). This leverages the distributional hypothesis: terms that occur in the same contexts tend to have similar meanings (Harris, 1954). So by looking at the contexts of terms, BERT models can automatically learn which terms are similar, and we can use these relations to automatically expand queries. We will experiment with these techniques in EXALT.

Another improvement we would like to investigate is the indexing of documents by section, instead of by page or whole document, as we do now. This feature was also requested by the focus group in our user requirement solicitation study (Chapter 5). Being able to search through texts with sections as the indexing unit makes more sense than searching per page, as information might be spread across multiple pages. Also, knowing which section a term occurs in could be beneficial to retrieval, think of a section called "Flint analysis" containing the term "Neolithic". This is a very strong indication that this section is relevant to the query "Neolithic flint", perhaps stronger than the words "Neolithic" and "flint" occurring near each other in the text. Lastly, it would be useful to exclude certain sections from indexing, such as generic time period lists often included in reports, which are irrelevant to search.

Besides creating a publicly available search engine, we will also publish all the extracted information as Linked Open Data (LOD) allowing for novel data science approaches by other researchers. As part of the valorisation, we will perform three case studies to assess the system and its influence on archaeological research. The general public will be involved as well, through a 'map of the past' allowing easy access to archaeological information, and a partnership with an archaeological museum and the AWN (*Vereniging van Vrijwilligers in de Archeologie*, the Dutch society for volunteer archaeology) to promote the system.

9.9.2 Long Term Ideas

Here we describe avenues for research that would be very beneficial, but are not currently part of the EXALT project goals.

We already mentioned query expansion in the previous section, but it is also possible to completely bypass entities and ontologies, and search directly on the embeddings created by BERT or other language models (Karpukhin *et al.*, 2020;

Deshmukh & Sethi, 2020). That way we can match query terms to documents that contain the same and similar terms as defined by their similarity in a vector space. It would be very interesting to see if there is enough contextual information in archaeological texts to use this method effectively, and whether or not it would outperform search on entities linked to ontology entries.

As mentioned in Chapter 2, there are problems with the reports being stored in the PDF format, the noise this creates when converting to plain text, and how the structure of the documents (chapters, headings) is very difficult to extract from these files. This document segmentation mentioned above is something we would like to further research to provide more useful results, but ideally, we would like to see new reports to be created in a file format that maintains the document structure, and possibly allows for some semantic annotation. Most layout and design programs will offer the possibility of exporting a document to more structured file formats (mostly HTML or XML), which would already be beneficial, as also shown by Meckseper & Warwick (2003). But perhaps creating a unified standard for archaeological reports is needed, which could be maintained by the SIKB, like they already do for the "*pakbon*" (packing slip), an XML format for excavation data (Stichting Infrastructuur Kwaliteitsborging Bodembeheer, 2016). However, this is a complicated and long-term goal that would need a coalition of all relevant partners in the cultural heritage domain to formulate and maintain the standard. While this is something a research project could not hope to achieve on its own, we aim to start building this coalition and facilitate a discussion on this topic.

Lastly, something we would like to see is the use of the archaeological entities we extracted from our data set, as deposited in the DANS archive (Brandsen, 2021a)[1]. This data set lists the entities found in each document, together with a list of generic metadata, and could be used for interesting computational approaches. Some ideas include the research by Wilcke *et al.* (2019), who aimed to extract meaningful relations between archaeological concepts from the aforementioned *pakbon* XML data, and the research by Plets *et al.* (2021) looking at changes in quality and sentiment in Flemish archaeology over time.

9.9.3 Recommendations

Throughout this research, we have tried and evaluated many methodologies and processes. In this section, we reflect on what did and did not work and give recommendations for any future research on this topic.

[1] Available at: doi.org/10.17026/dans-zcs-7b72

In the Introduction chapter we introduced the Agile principles: creating software in small cycles by building quick prototypes, testing these with users, and updating where needed. We initially aimed at more development/testing cycles, but due to other work (writing, presenting, etc) and the Covid-19 crisis, we ended up doing three cycles. Even though this is fewer than anticipated, this method definitely proved useful: the original prototype (as described in Chapter 5) was built in just a couple of months, but proved invaluable when soliciting requirements from the user group.

In general, having users in the loop during development was very fruitful. Being able to quickly update the system to match user requirements and fix usability issues meant (almost) no programming time was wasted on features that were not needed, or needed changing. We unconsciously adopted user-centred design, putting the user in the centre of focus, as opposed to project goals. This help from our user group was essential in determining the direction of the software development. We would recommend similar software development projects in archaeology (and other disciplines) to also follow this quick cycle and user in the loop approach, as opposed to the more traditional linear development process.

Related to this, we found that while it was easy to calculate performance metrics on e.g. the NER process, to truly measure the usefulness of a system it is needed to apply it to a real world problem, in our case the case study presented in Chapter 8. Without an evaluation with a user in a non-controlled setting it is nearly impossible to get an idea of how useful a system is. This was especially true for this project as we had no data set with relevance assessments to automatically calculate the performance of the Information Retrieval.

We did have a labelled data set for NER at the start of the project, created in the ARIADNE project. However, after working with this data set and investigating prediction errors in classifiers trained on this data, we realised it was not ideal for our methods, as described in Chapter 3. A recommendation is to always check the data quality before starting experimenting, as this would have saved us considerable time.

We had some experience with organising students to annotate entities in text, to create better quality data. The main lessons from this work were: (1) to test annotation guidelines with one or two people outside of the project, as this brought to light issues overlooked by the project, (2) to do the annotation with all annotators in a room, so everyone learns from each other's questions, and update the guidelines on the fly, and (3) to use the pairwise F1 score for Inter Annotator Agreement on NER, instead of Cohen's Kappa which is often used for IAA in other classification tasks.

In a lot of classification tasks in archaeology, we see that a method is tested

and evaluated, but not always compared to a baseline. As we mentioned before, less complex methods can lead to satisfying results, sometimes outperforming more complex methods, and as such should always be experimented with before trying the state of the art. In the case of NER, a common baseline is CRF, which we found to be very effective (although outperformed by BERT).

Bibliography

Abrahamse, J., Blom, A., Bouwmeester, H., Bos, J., Brinkkemper, O., Brounen, F., Cohen, K., Dambrink, R., Bruijn, R.d., Groot, T.d., Kort, J.d., Vries, F.d., Vries, S.d., Eerden, M., Erkens, G., Feiken, H., Gouw-Bouman, M., Groenewoudt, B., Hijma, M., Huisman, D., Jansen, B., Kosian, M., Koster, K., Kriek, M., Lascaris, M., Lauwerier, R., Maas, G., Marges, V., Pierik, H., Rensink, E., Romeijn, E., Schokker, J., Smit, B., Snoek, M., Speleers, B., Stafleu, J., Theunissen, E., Beek, R.v., Jagt, I.v.d., Doesburg, J.v., Reuler, H.v., Vos, P. & Weerts, H. (2017). Knowledge for Informed Choices. Tools for more effective and efficient selection of valuable archaeology in the Netherlands. In R. Lauwerier, M. Eerden, B. Groenewoudt, M. Lascaris, E. Rensink, B. Smit, B. Speleers & J.v. Doesburg, eds., *Nederlandse Archeologische Rapporten*, volume 55. Rijksdienst voor het Cultureel Erfgoed, Amersfoort. ISBN 9789057992773.

Adams, D. (1979). *The Hitchhiker's Guide to the Galaxy*. Pan Books (UK). ISBN 0-330-25864-8.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S. & Vollgraf, R. (2019). FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59. Association for Computational Linguistics, Minneapolis, Minnesota. DOI: 10.18653/v1/N19-4010.

Akhtyamova, L. (2020). Named Entity Recognition in Spanish Biomedical Literature: Short Review and Bert Model. In *26th Conference of Open Innovations Association (FRUCT)*, pp. 1–7. IEEE Computer Society, Yaroslavl, Russia. ISBN 9789526924427. ISSN 23057254. DOI: 10.23919/FRUCT48808.2020.9087359.

Amrani, A., Abajian, V. & Kodratoff, Y. (2008). A chain of text-mining to extract information in archaeology. In *Information and Communication Technologies: From Theory to Applications, ICTTA 2008.*, pp. 1–5. Damascus, Syria. DOI: 10.1109/ICTTA.2008.4529905.

Annaert, R. (2018). *Het Vroegmiddeleeuwse Grafveld Van Broechem, Volume II Analyse*. Habelt Verlag, Bonn.

Athens, J.S. (1993). Cultural resource management and academic responsibility in archaeology: A further comment. *SAA Bulletin*, 11(2), pp. 6–7.

Auger, C.P.C.P. (1975). *Use of reports literature*. Archon Books. ISBN 020801506X.

Auger, C.P.C.P. (1989). *Information sources in Grey literature*. Bowker-Saur. ISBN 0862918715.

Averett, E.W., Counts, D. & Gordon, J. (2016). *Mobilizing the past for a digital future: the potential of digital archaeology*. Technical report, Creighton University. DOI: 10.17613/M6HJ56.

Barbour, R. (2018). *Doing focus groups*. Sage, Los Angeles London. ISBN 9781473912441.

Bartalesi, V., Meghini, C., Metilli, D. & Andriani, P. (2016). Usability Evaluation of the Digital Library DanteSources. In *International Conference on Theory and Practice of Digital Libraries*, pp. 191–203. Springer, Cham. DOI: 10.1007/978-3-319-39513-5_18.

Bazelmans, J., Brinkkemper, O., Deeben, J., Van Doesburg, J., Lauwerier, R. & Zoetbrood, P. (2005). Mag het ietsje meer zijn? Een onderzoek naar de door bedrijven opgestelde Programma's van Eisen voor archeologisch onderzoek uit de periode 2003-2004. *Rapportage Archeologische Monumentenzorg*, 120.

Beck, K., Beedle, M., Bennekum, A.V., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B.,

Martin, R.C., Mellor, S., Schwaber, K., Sutherland, J. & Thomas, D. (2001). Manifesto for Agile Software Development. http://agilemanifesto.org/.

Behnert, C. & Lewandowski, D. (2017). A framework for designing retrieval effectiveness studies of library information systems using human relevance assessments. *Journal of Documentation*, 73(3), pp. 509–527. ISSN 00220418. DOI: 10.1108/JD-08-2016-0099.

Beltagy, I., Lo, K. & Cohan, A. (2020). SCIBERT: A pretrained language model for scientific text. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics, Hong Kong, China. ISBN 9781950737901. DOI: 10.18653/v1/d19-1371.

Bennett, R., Cowley, D. & De Laet, V. (2014). The data explosion: Tackling the taboo of automatic feature recognition in airborne survey data. *Antiquity*, 88(341), pp. 896–905. ISSN 0003598X. DOI: 10.1017/S0003598X00050766.

Bevan, A. (2015). The data deluge. *Antiquity*, 89(348), pp. 1473–1484. DOI: 10.15184/aqy.2015.102.

Bloomberg, J. (2013). The Big Data Long Tail. http://www.devx.com/blog/the-big-data-long-tail.html.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5(1), pp. 135–146.

Boulton, G., Campbell, P., Collins, B., Elias, P., Hall, W., Graeme, L., O'Neill, O., Rawlins, M., Thornton, J., Vallance, P. & Walport, M. (2012). *Science as an open enterprise. The Royal Society Science Policy Centre report 02/12*. Technical Report June, The Royal Society, London. http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf.

Boyd, D. & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*. ISSN 1369118X. DOI: 10.1080/1369118X.2012.678878.

Bradley, R., Haselgrove, C., Vander Linden, M. & Webley, L. (2016). *The later prehistory of north-west Europe: The evidence of development-led fieldwork*. Oxford University Press, Oxford. ISBN 9780199659777.

Bramer, W.M., Giustini, D., Kramer, B.M. & Anderson, P. (2013). The comparative recall of Google Scholar versus PubMed in identical searches for biomedical systematic reviews: a review of searches used in systematic reviews. *Systematic reviews*, 2(1), p. 115. ISSN 20464053. DOI: 10.1186/2046-4053-2-115.

Branco, P., Torgo, L. & Ribeiro, R. (2015). A survey of predictive modelling under imbalanced distributions.

Brandsen, A. (2018). alexbrandsen/archaeo-CRF: Version 0.1 of Archaeo CRF. *Zenodo Repository*. DOI: 10.5281/ZENODO.1238861.

Brandsen, A. (2019). alexbrandsen/dutch-archaeo-NER-dataset: First version. *Zenodo Repository*. DOI: 10.5281/ZENODO.3544544.

Brandsen, A. (2020). alexbrandsen/archaeo-document-classification-dataset: Second version. *Zenodo Repository*. DOI: 10.5281/ZENODO.4115747.

Brandsen, A. (2021a). Archaeological entities and timespans extracted from all archaeology documents available in DANS EASY in 2017. DOI: 10.17026/dans-zcs-7b72.

Brandsen, A. (2021b). ArcheoBERTje - A Dutch BERT model for the Archaeology domain. *Zenodo Repository*. DOI: 10.5281/zenodo.4739063.

Brandsen, A. & Koole, M. (2021). Labelling the Past: Data Set Creation and Multi-label Classification of Dutch Archaeological Excavation Reports. *Language Resources and Evaluation*. DOI: 10.1007/s10579-021-09552-6.

Brandsen, A., Lambers, K., Verberne, S. & Wansleeben, M. (2019). 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), pp. 21–30. DOI: 10.5334/jcaa.33.

Brandsen, A. & Lippok, F. (2021). AGNES case study data. *Zenodo Repository*. DOI: 10.5281/zenodo.4737564.

Brandsen, A., Verberne, S., Lambers, K. & Wansleeben, M. (2021a). Can BERT Dig It? - Named Entity Recognition for Information Retrieval in the Archaeology Domain. *arXiv*. http://arxiv.org/abs/2106.07742.

Brandsen, A., Verberne, S., Lambers, K. & Wansleeben, M. (2021b). Usability Evaluation for Online Professional Search in the Dutch Archaeology Domain. *arXiv*. http://arxiv.org/abs/2103.04437.

Brandsen, A., Verberne, S., Wansleeben, M. & Lambers, K. (2020). Creating a Dataset for Named Entity Recognition in the Archaeology Domain. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 4573–4577. European Language Resources Association, Marseille, France. https://www.aclweb.org/anthology/2020.lrec-1.562/.

Brandt, R., Drenth, E., Montforts, M., Proos, R., Roorda, I. & Wiemer, R. (1992). *Archeologisch Basisregister*. Technical report, Rijksdienst voor Cultureel Erfgoed, Amersfoort.

Bulatovic, N., Gnadt, T., Romanello, M., Stiller, J. & Thoden, K. (2016). Usability in digital humanities - Evaluating user interfaces, infrastructural components and the use of mobile devices during research process. In N. Fuhr, L. Kovács, T. Risse & W. Nejdl, eds., *Research and Advanced Technology for Digital Libraries. TPDL 2016. Lecture Notes in Computer Science*, volume 9819 LNCS, pp. 335–346. Springer, Cham. ISBN 9783319439969. ISSN 16113349. DOI: 10.1007/978-3-319-43997-6_26.

Byrne, K. & Klein, E. (2010). Automatic Extraction of Archaeological Events from Text. In B. Frischer, J. Crawford, & D. Koller, eds., *Making History Interactive: Computer Applications and Quantitative Methods in Archaeology 2009*, pp. 48–56. BAR International Series 2079, Oxford.

Capannini, G., Nardini, F.M., Perego, R. & Silvestri, F. (2011). Efficient diversification of web search results. In *Proceedings of the VLDB Endowment*, pp. 451 – 459. Seattle, Washington. ISSN 21508097. DOI: 10.14778/1988776.1988781.

Carpineto, C. & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1). ISSN 03600300. DOI: 10.1145/2071389.2071390.

Chan, B., Schweter, S. & Möller, T. (2021). German's Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6788–6796. International Committee on Computational Linguistics, Barcelona, Spain. DOI: 10.18653/v1/2020.coling-main.598.

Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. SAGE Publications Ltd, London. ISBN 0761973532. DOI: 10.1080/17482620600881144.

Cheng, X., Bowden, M., Bhange, B.R., Goyal, P., Packer, T. & Javed, F. (2020). An End-to-End Solution for Named Entity Recognition in eCommerce Search. *arXiv*. http://arxiv.org/abs/2012.07553.

Cherman, E.A., Monard, M.C. & Metz, J. (2011). Multi-label problem transformation methods: a case study. *CLEI Electronic Journal*, 14(1), p. 4.

Chowdhury, G.G. (2005). Natural language processing. *Annual Review of Information Science and Technology*, 37(1), pp. 51–89. ISSN 00664200. DOI: 10.1002/aris.1440370103.

Claeys, J., Baetsen, S., Drenth, E., Huizer, J., Jaspers, N., Kempkens, J., Lupak, T. & Melkert, M. (2012). Onder de Gelderakkers een uitgestrekte meerperiodensite in Hilvarenbeek. Een inventariserend veldonderzoek door middel van proefsleuven (IVO-P). *ADC Archeoprojecten Rapport*, 2613. DOI: 10.17026/dans-zvw-3mpv.

Cohen, L., Manion, L., Morrison, K., Manion, L. & Morrison, K. (2002). *Research Methods in Education*. Routledge. ISBN 9780203224342. DOI: 10.4324/9780203224342.

Concrete5 (2018). Concrete5 Content Management System. https://www.concrete5.org/.

Copara, J., Naderi, N., Knafou, J., Ruch, P. & Teodoro, D. (2020). Named entity recognition in chemical patents using ensemble of contextual language models. *arXiv*. ISSN 23318422. http://arxiv.org/abs/2007.12569.

Copeland, J.M. (1983). Information retrieval systems for archaeological data. In J. Haigh, ed., *Proceedings of the Conference on Computer Applications and Quantitative Methods in Archaeology (CAA) 1983*, pp. 39–45. School of Archaeological Sciences, University of Bradford, Bradford.

Corstius, H.B. (1981). *Opperlandse taal-& letterkunde*. Querido. ISBN 9789021451343.

Costopoulos, A. (2016). Digital Archeology Is Here (and Has Been for a While). *Frontiers in Digital Humanities*, 3. ISSN 2297-2668. DOI: 10.3389/fdigh.2016.00004.

Council of Europe (1992). European Convention on the Protection of the Archaeological Heritage (Revised). http://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/143.

Cowan, B., Zethelius, S., Luk, B., Baras, T., Ukarde, P. & Zhang, D. (2015). Named entity recognition in travel-related search queries. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 3935–3941. AAAI Press, Austin, Texas. ISBN 9781577357032.

Cowley, D.C. (2012). In with the new, out with the old? Auto-extraction for remote sensing archaeology. In C.R. Bostater, S.P. Mertikas, X. Neyt, C. Nichol, D. Cowley & J.P. Bruyant, eds., *Remote Sensing of the Ocean, Sea Ice, Coastal Waters, and Large Water Regions 2012*. SPIE, Edinburgh, UK. ISBN 9780819492722. ISSN 0277786X. DOI: 10.1117/12.981758.

Croft, W.B., Metzler, D. & Strohman, T. (2010). *Search engines: Information retrieval in practice*. Pearson. ISBN 978-0136072249.

Cunningham, H., Gaizauskas, R.J. & Wilks, Y. (1995). *A general architecture for text engineering (GATE): A new approach to language engineering R & D*. University of Sheffield, Department of Computer Science.

DANS (2019). DANS EASY. https://dans.knaw.nl/en/about/services/easy.

De Mauro, A., Greco, M. & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. In *AIP Conference Proceedings*, volume 1644, pp. 97–104. American Institute of Physics Inc. ISBN 9780735412835. ISSN 15517616. DOI: 10.1063/1.4907823.

De Mauro, A., Greco, M. & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, 65(3), pp. 122–135. ISSN 00242535. DOI: 10.1108/LR-06-2015-0061.

De Romas, R. (2019). *Multi-label Text Classification for Ground Lease Documents*. Thesis, University of Amsterdam.

De Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G. & Nissim, M. (2019). BERTje: A Dutch BERT Model. *arXiv*. http://arxiv.org/abs/1912.09582.

Dejong, M. & Schellens, P.J. (1997). Reader-Focused Text Evaluation: An Overview of Goals and Methods. *Journal of Business and Technical Communication*, 11(4), pp. 402–432. DOI: 10.1177/1050651997011004003.

Deleger, L., Li, Q., Lingren, T., Kaiser, M., Molnar, K., Stoutenborough, L., Kouril, M., Marsolo, K. & Solti, I. (2012). Building gold standard corpora for

medical natural language processing tasks. *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2012, pp. 144–153.

Delobelle, P., Winters, T. & Berendt, B. (2020). RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3255–3265. Association for Computational Linguistics, Online. ISSN 23318422. DOI: 10.18653/v1/2020.findings-emnlp.292.

Deshmukh, A.A. & Sethi, U. (2020). IR-BERT: Leveraging BERT for Semantic Search in Background Linking for News Articles. *arXiv*. https://arxiv.org/abs/2007.12603.

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long an, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota. DOI: 10.18653/v1/N19-1423.

Dudek, D., Mastora, A. & Landoni, M. (2007). Is Google the answer? A study into usability of search engines. *Library Review*, 56(3), pp. 224–233. DOI: 10.1108/00242530710736000.

Eerden, M., Groenewoudt, B., de Groot, T., Theunissen, E. & Feiken, H. (2017). Synthesising data from development- led archaeological research. In R. Lauwerier, M. Eerden, B. Groenewoudt, M. Lascaris, E. Rensink, B. Smit, B. Speleers & J. van Doesburg, eds., *Knowledge for Informed Choices Tools for more efective and efcient selection of valuable archaeology in the Netherlands*. Rijksdienst voor het Cultureel Erfgoed, Amersfoort.

Effros, B. (2003). *Merovingian Mortuary Archaeology and the Making of the Early middle ages*. University of California Press, Berkeley.

ElasticSearch (2018). Theory Behind Relevance Scoring. https://www.elastic.co/guide/en/elasticsearch/guide/current/scoring-theory.html.

Eramian, M., Walia, E., Power, C., Cairns, P. & Lewis, A. (2017). Image-based search and retrieval for biface artefacts using features capturing archaeologically significant characteristics. *Machine Vision and Applications*, 28(1-2), pp. 201–218. ISSN 14321769. DOI: 10.1007/s00138-016-0819-x.

Esmailpour, R., Ebrahimy, S., Fakhrahmad, S.M., Mohammadi, M. & Abbaspour, J. (2019). Developing an effective scheme for translation and expansion of Persian user queries. *Digital Scholarship in the Humanities*. ISSN 2055-7671. DOI: 10.1093/llc/fqz041.

Evans, C., Tabor, J. & Vander Linden, M. (2014). Making time work: Sampling floodplain artefact frequencies and populations. *Antiquity*, 88(339), pp. 241–258. ISSN 0003598X. DOI: 10.1017/S0003598X0005033X.

Evans, T.N.L. (2015). A reassessment of archaeological grey literature: Semantics and paradoxes. *Internet Archaeology*, 40. ISSN 13635387. DOI: 10.11141/ia.40.6.

Falkingham, G. (2005). A Whiter Shade of Grey: A new approach to archaeological grey literature using the XML version of the TEI Guidelines. *Internet Archaeology*, 17. DOI: 10.11141/ia.17.5.

Farace, D.J. & Schoptimefel, J. (2010). *Grey literature in library and information studies*. De Gruyter Saur. ISBN 9783598117930. http://hal.univ-lille3.fr/hal-01288536.

Fehr, H. (2008). Germanische Einwanderung oder kulturelle Neuorientierung? Zu den Anfängen des Reihengräberhorizontes. In S. Brather, ed., *Zwischen Spätantike und Frühmittelalter*, pp. 67–102. Walter de Gruyter, Berlin. DOI: 10.1515/9783110210729.2.67.

Feldman, R. & Dagan, I. (1995). Knowledge Discovery in Textual Databases (KDT). In *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*. AAAI Press, Montreal, Canada. http://www.aaai.org/Papers/KDD/1995/KDD95-012.pdf.

Feldman, R. & Sanger, J. (2007). *The text mining handbook : advanced approaches in analyzing unstructured data*. Cambridge University Press. ISBN 9780521836579.

Fiorucci, M., Khoroshiltseva, M., Pontil, M., Traviglia, A., Del Bue, A. & James, S. (2020). Machine Learning for Cultural Heritage: A Survey. *Pattern Recognition Letters*. ISSN 01678655. DOI: 10.1016/j.patrec.2020.02.017.

Fischer, A., Londen, H.v., Bercken, A.B.v.d., Visser, R. & Renes, J. (2021). NAR 68 Urban farming and ruralisation in the Netherlands (1250 up to the nineteenth century), unravelling farming practice and the use of (open) space by

synthesising archaeological reports using text mining. *Nederlandse Archeologische Rapporten (NAR)*, 68.

Fokkens, H., Steffens, B. & van As, S. (2016). *Farmers, fishers, fowlers, hunters. Knowledge generated by development-led archaeology about the Late Neolithic, the Early Bronze Age and the start of the Middle Bronze Age (2850 - 1500 cal BC) in the Netherlands*. Rijksdienst voor het Cultureel Erfgoed, Amersfoort.

Foley, R. (1981). Off-site archaeology: an alternative approach for the short-sited. In I. Hodder, G. Isaac & N. Hammond, eds., *Pattern of the past: studies in honour of David Clarke*. Cambridge University Press, Cambridge. ISBN 9780521108430.

Gartner Glossary (2021). Definition of Big Data - Gartner Information Technology Glossary. https://www.gartner.com/en/information-technology/glossary/big-data.

Gattiglia, G. (2015). Think big about data: Archaeology and the Big Data challenge. *Archäologische Informationen*, 38(1), pp. 113–124. ISSN 2197-7429. DOI: 10.11588/ai.2015.1.26155.

Gerjets, P., Kammerer, Y. & Werner, B. (2011). Measuring spontaneous and instructed evaluation processes during Web search: Integrating concurrent thinking-aloud protocols and eye-tracking data. *Learning and Instruction*, 21(2), pp. 220–231. DOI: 10.1016/j.learninstruc.2010.02.005.

Gibbs, F. & Owens, T. (2012). Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs. *Digital Humanities Quarterly*, 6(2). ISSN 1938-4122. http://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html.

Gibbs, M. & Colley, S. (2012). Digital Preservation,: Online access and historical archaeology 'grey literature' from New South Wales, Australia. *Australian Archaeology*, 75, pp. 95–103. ISSN 03122417. DOI: 10.1080/03122417.2012.11681957.

Glyph & Cog LLC (1996). pdftotext. https://www.xpdfreader.com/pdftotext-man.html.

Golub, K., Hagelbäck, J. & Ardö, A. (2020). Automatic Classification of Swedish Metadata Using Dewey Decimal Classification: A Comparison of Approaches. *Journal of Data and Information Science*, 5(1), pp. 18–38. ISSN 2096157X. DOI: 10.2478/jdis-2020-0003.

Gormley, C. & Tong, Z. (2015). *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine*. O'Reilly Media, Sebastopol.

Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O. & Quintard, L. (2011). Proposal for an Extension of Traditional Named Entities: From Guidelines to Evaluation, an Overview. In S. Pradhan, K. Tomanek, N. Ide & A. Meyers, eds., *Proceedings of the 5th Linguistic Annotation Workshop*, pp. 92–100. Association for Computational Linguistics, Portland, Oregon, USA. https://aclanthology.org/W11-0411.

Gruber, T.R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human - Computer Studies*, 43(5-6), pp. 907–928. ISSN 10959300. DOI: 10.1006/ijhc.1995.1081.

Guo, J., Xu, G., Cheng, X. & Li, H. (2009). Named entity recognition in query. In *Proceedings - 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009*, pp. 267–274. Association for Computing Machinery, Boston, Massachusetts. ISBN 9781605584836. DOI: 10.1145/1571941.1571989.

Habermehl, D. (2019). *Over zaaien en oogsten, de kwaliteit en bruikbaarheid van archeologische rapporten voor synthetiserend onderzoek*. Technical report, Rijksdienst voor Cultureel Erfgoed, Amersfoort. https://www.cultureelerfgoed.nl/publicaties/publicaties/2019/01/01/over-zaaien-en-oogsten.

Hakala, K. & Pyysalo, S. (2019). Biomedical Named Entity Recognition with Multilingual BERT. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pp. 56–61. Association for Computational Linguistics, Hong Kong, China. DOI: 10.18653/v1/d19-5709.

Hand, D. & Christen, P. (2018). A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing*, 28(3), pp. 539–547. ISSN 1573-1375. DOI: 10.1007/s11222-017-9746-6.

Harris, Z.S. (1954). Distributional structure. *Word*, 10(2-3), pp. 146–162.

Hendriks, J. (2013). *Een merovingisch grafveld in het Lentseveld te Nijmegen-Noord. Evaluatie en selectierapport NLa14*. Technical report, Archeologie Gemeente Nijmegen.

Hermjakob, U., Hovy, E. & Lin, C. (2000). Knowledge-based question answering. In *Proceedings of the Sixth World Multiconference on Systems, Cybernetics, and Informatics (SCI-2002)*. International Institute of Informatics and Systemics, Winter Garden, FL.

Hessing, W., Waugh, K., van Heeringen, R. & Visser, C. (2013). *Evaluatie en optimalisatie waarderingssystematiek Kwaliteitsnorm Nederlandse Archeologie. Fase 1: Evaluatie*. Technical report, Vestigia, Amersfoort.

Highsmith, J., Paulk, M.C., Manzo, J., Mcmahon, P.E., Bowers, P., Sleve, G. & Bingham, K. (2002). Agile software development. *The journal of defense software engineering*, 15(10). http://agilesweden.com/doc/oct02.pdf.

Hinostroza, J.E., Ibieta, A., Labbé, C. & Soto, M.T. (2018). Browsing the internet to solve information problems: A study of students' search actions and behaviours using a 'think aloud' protocol. *Education and Information Technologies*, 23(5), pp. 1933–1953. DOI: 10.1007/s10639-018-9698-2.

Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E. & Coates, C.M. (2015). Trading consequences: A case study of combining text mining and visualization to facilitate document exploration. *Digital Scholarship in the Humanities*, 30, pp. 50–75. DOI: 10.1093/llc/fqv046.

Hjørland, B. (1997). *Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science*. Praeger. ISBN 0313298939.

Hombert, P. (1950). Les sépultures mérovingiennes par incinération en Belgique. *Revue Archéologique*, 36, pp. 96–102.

Honnibal, M. & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear, still as of February 2020*.

Hripcsak, G. & Rothschild, A.S. (2005). Agreement, the F-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3), pp. 296–298. ISSN 10675027. DOI: 10.1197/jamia.M1733.

Hu, X. (2018). Usability Evaluation of E-Dunhuang Cultural Heritage Digital Library. *Data and Information Management*, 2(2), pp. 57–69. DOI: https://doi.org/10.2478/dim-2018-0008.

Huggett, J. (2012). Core or periphery? Digital humanities from an archaeological perspective. *Historical Social Research / Historische Sozialforschung*, 37(3), pp. 86–105. ISSN 01726404. DOI: 10.12759/hsr.37.2012.3.86-105.

Huurdeman, H.C. & Piccoli, C. (2020). "More than just a Picture" - The importance of context in search user interfaces for three-dimensional content. In *CHIIR 2020 - Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pp. 338–342. Association for Computing Machinery, Inc, New York, NY, USA. ISBN 9781450368926. DOI: 10.1145/3343413.3377994.

Huvila, I. (2017). Being FAIR when archaeological information is MEAN: Miscellaneous, Exceptional, Arbitrary, Nonconformist. http://www.istohuvila.se/node/526.

International Committee for Documentation (CIDOC) (2014). *Information and documentation - A reference ontology for the interchange of cultural heritage information (ISO Standard No. 21127:2014)*. Technical report, International Organization for Standardization. https://www.iso.org/standard/57832.html.

Jackson, S., Richissin, C.E., McCabe, E.E. & Lee, J.J. (2020). Data-Informed Tools for Archaeological Reflexivity: Examining the substance of bone through a meta-analysis of academic texts. *Internet Archaeology*, 55. ISSN 1363-5387. DOI: 10.11141/ia.55.12.

James, E. (1988). *The Franks*. Blackwells, Oxford.

Janssens, P. & Roosens, H. (1963). Lijkverbranding en lijkbegraving op het merovingisch grafveld te Grobbendonck. *Archaeologia Belgica*, 71, pp. 265–272.

Jeffrey, S., Richards, J., Ciravegna, F., Waller, S., Chapman, S. & Zhang, Z. (2009). The Archaeotools project: faceted classification and natural language processing in an archaeological context. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 367(1897), pp. 2507–19. DOI: 10.1098/rsta.2009.0038.

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In *Machine Learning: ECML-98*, pp. 137–142. Springer, Berlin, Heidelberg. ISBN 978-3-540-69781-7.

Karoulis, A., Sylaiou, S. & White, M. (2006). Usability evaluation of a virtual museum interface. *Informatica*, 17(3), pp. 363–380. ISSN 08684952. DOI: 10.15388/informatica.2006.143.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. & Yih, W.t. (2020). Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781. Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-main.550.

Khattab, O. & Zaharia, M. (2020). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48. ACM, New York, NY, USA. ISBN 9781450380164. DOI: 10.1145/3397271.3401075.

Kim, Y.M. & Lee, T.H. (2020). Korean clinical entity recognition from diagnosis text using BERT. *BMC Medical Informatics and Decision Making*, 20(S7), p. 242. ISSN 14726947. DOI: 10.1186/s12911-020-01241-8.

Kintigh, K.W. (2015). Extracting Information from Archaeological Texts. *Open Archaeology*, 1(1), pp. 96–101. DOI: 10.1515/opar-2015-0004.

Kirkpatrick, L.C. (2018). *Using Computer Screen Recordings and Think Aloud Protocols to Study Students' Cognitive Strategies While Working Online*. SAGE Publications Ltd, London. DOI: 10.4135/9781526444240.

Kleppe, M., Hendrickx, I., Veldhoen, S., Brandsen, A., Vos, H.D., Goes, K., Huang, L., Huurdeman, H., Kim, A., Mesbah, S., Reuver, M., Wang, S. & Zijdeman, R. (2019). *(Semi-) Automatic Cataloguing of Textual Cultural Heritage Objects*. Technical report, KB (National Library of the Netherlands), Den Haag. http://www.kbresearch.nl/brinkeys/report.pdf.

Koolen, M., van Gorp, J. & van Ossenbruggen, J. (2018). Toward a model for digital tool criticism: Reflection as integrative practice. *Digital Scholarship in the Humanities*, 34(2), pp. 368–385. ISSN 2055-7671. DOI: 10.1093/llc/fqy048.

Kudo, T. & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP 2018 - Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings*, pp. 66–71. Association for

Computational Linguistics, Brussels, Belgium. ISBN 9781948087858. DOI: 10.18653/v1/d18-2012.

Kuratov, Y. & Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language. http://arxiv.org/abs/1905.07213.

Lafferty, J., Mccallum, A., Pereira, F.C.N. & Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In C.E. Brodley & D.A. Pohoreckyj, eds., *Proc. 18th International Conf. on Machine Learning*, pp. 282–289. Morgan Kaufmann Publishers Inc., San Fransisco.

Lambers, K., Verschoof-van der Vaart, W. & Bourgeois, Q. (2019). Integrating Remote Sensing, Machine Learning, and Citizen Science in Dutch Archaeological Prospection. *Remote Sensing*, 11(7), p. 794. ISSN 2072-4292. DOI: 10.3390/rs11070794.

Lancaster, F.W. & Gallup, E. (1973). *Information Retrieval On-Line*. Melville Publishing Company, Los Angeles, California.

Laza, R., Pavón, R., Reboiro-Jato, M. & Fdez-Riverola, F. (2011). Evaluating the effect of unbalanced data in biomedical document classification. *Journal of integrative bioinformatics*, 8(3), p. 177. ISSN 16134516. DOI: 10.1515/jib-2011-177.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), pp. 1234–1240. ISSN 1367-4803. DOI: 10.1093/bioinformatics/btz682.

Lehnemann, E. (2008). Das Gräberfeld von Lünen-Wethmar, Kr. Unna. *Internationale Archäologie*, 108.

Levy, T. (2014). Front Matter. *Near Eastern Archaeology*, 77(3). DOI: 10.5615/neareastarch.77.3.fm.

Lewis, C. (1982). *Using the 'thinking-aloud' method in cognitive interface design*. Technical report, IBM TJ Watson Research Center, New York.

Li, X., Zhang, H. & Zhou, X.H. (2020). Chinese clinical named entity recognition with variant neural structures based on BERT methods. *Journal of Biomedical Informatics*, 107, p. 103422. ISSN 15320464. DOI: 10.1016/j.jbi.2020.103422.

Lippok, F. (2019). Een vroegmiddeleeuws graf aan de rand van de nederzetting. In E. Norde, ed., *Nederzettingsresten uit de vroege middeleeuwen in het plangebied Leeuwesteyn Noord in Leidsche Rijn, Diemen. RAAP-rapport 3855*, pp. 91–99. RAAP.

Lippok, F. (2020). The pyre and the grave: early medieval cremation burials in the Netherlands, the German Rhineland and Belgium. *World Archaeology*, 52(1), pp. 147–162. ISSN 14701375. DOI: 10.1080/00438243.2020.1769297.

Lippok, F. (2021). The early medieval graves of Oegstgeest. In J. De Bruin, C. Bakels & F. Theuws, eds., *Oegstgeest. A riverrine settlement in the early medieval world system*, pp. 84–107. Habelt Verlag, Bonn.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv*. https://arxiv.org/abs/1907.11692.

Manning, C.D., Raghavan, P. & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. ISBN 0521865719.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J. & Gómez-Berbís, J.M. (2013). Named Entity Recognition: Fallacies, challenges and opportunities. *Computer Standards and Interfaces*, 35(5), pp. 482–489. ISSN 09205489. DOI: 10.1016/j.csi.2012.09.004.

Mayer-Schönberger, V. & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McHugh, M.L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), pp. 276–282. ISSN 13300962. DOI: 10.11613/bm.2012.031.

Meckseper, C. & Warwick, C. (2003). The Publication of Archaeological Excavation Reports Using XML. *Literary and Linguistic Computing*, 18(1), pp. 63–75. ISSN 0268-1145. DOI: 10.1093/llc/18.1.63.

Mélanie-becquet, F., Ferguth, J., Gruel, K. & Poibeau, T. (2015). Archaeology in the Digital Age: From Paper to Databases. In *Proceedings of the conference "Digital Humanities 2015"*. Sydney.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

Ministerie van Onderwijs Cultuur en Wetenschap (2007). Wet op de archeologische monumentenzorg. https://wetten.overheid.nl/jci1.3:c:BWBR0021162&z=2008-01-01&g=2008-01-01.

Ministerie van Onderwijs Cultuur en Wetenschap (2015). Erfgoedwet. https://wetten.overheid.nl/jci1.3:c:BWBR0037521&z=2021-07-01&g=2021-07-01.

Mohammed, R., Rawashdeh, J. & Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pp. 243–248. DOI: 10.1109/ICICS49469.2020.239556.

Mohri, M., Rostamizadeh, A. & Talwalkar, A. (2013). *Foundations of Machine Learning*. MIT Press, Cambridge, MA, 2nd edition. ISBN 9780262039406.

Moon, T., Awasthy, P., Ni, J. & Florian, R. (2019). Towards Lingua Franca Named Entity Recognition with BERT. *arXiv*. http://arxiv.org/abs/1912.01389.

Morgan, C. & Eve, S. (2012). DIY and digital archaeology: What are you doing to participate? *World Archaeology*, 44(4), pp. 521–537. ISSN 00438243. DOI: 10.1080/00438243.2012.741810.

Nadkarni, P.M., Ohno-Machado, L. & Chapman, W.W. (2011). Natural language processing: An introduction. DOI: 10.1136/amiajnl-2011-000464.

Nakayama, H. (2019). chakki-works/doccano: Open source text annotation tool for machine learning practitioner. https://github.com/chakki-works/doccano.

Niccolucci, F. & Richards, J. (2019). *The ARIADNE Impact*. Archaeolingua Foundation, Hungary. DOI: 10.5281/ZENODO.4319058.

Niccolucci, F. & Richards, J.D. (2013). ARIADNE: Advanced Research Infrastructures for Archaeological Dataset Networking in Europe. *International Journal of Humanities and Arts Computing*, 7(1-2), pp. 70–88. ISSN 1753-8548. DOI: 10.3366/ijhac.2013.0082.

Nielsen, J. & Landauer, T.K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '93*, pp. 206–213. ACM Press, New York, New York, USA. ISBN 0897915755. DOI: 10.1145/169059.169166.

Norvig, P. (2013). English Letter Frequency Counts: Mayzner Revisited. http://norvig.com/mayzner.html.

Nozza, D., Bianchi, F. & Hovy, D. (2020). What the [MASK]? Making Sense of Language-Specific BERT Models. *arXiv*. ISSN 23318422. http://arxiv.org/abs/2003.02912.

Paijmans, H. & Brandsen, A. (2009). What is in a Name: Recognizing Monument Names from Free-Text Monument Descriptions. In M. van Erp, J. Stehouwer & M. van Zaanen, eds., *Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning (Benelearn)*, pp. 2–6. Tilburg centre for Creative Computing, Tilburg. http://benelearn09.uvt.nl/Proceedings_Benelearn_09.pdf.

Paijmans, H. & Brandsen, A. (2010). Searching in archaeological texts: Problems and solutions using an artificial intelligence approach. *PalArch's Journal Of Archaeology Of Egypt/Egyptology*, 7(2), pp. 1–6.

Paijmans, J. & Wubben, H. (2008). Preparing archeological reports for intelligent retrieval. In A. Posluschny, K. Lambers & I. Herzog, eds., *Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA)*, pp. 2–6. Berlin.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct), pp. 2825–2830.

Pescarin, S., Pagano, A., Wallergård, M., Hupperetz, W. & Ray, C. (2014). Evaluating Virtual Museums: Archeovirtual Case Study. In P. Verhagen & G. Earl, eds., *Archaeology in the Digital Era*, pp. 74–82. Amsterdam University Press. DOI: doi:10.1515/9789048519590-009.

Peters, T. (2004). PEP 20 – The Zen of Python. https://www.python.org/dev/peps/pep-0020/.

Peterson, C. & Seligman, M. (1984). *Content analysis of verbatim explanations: The CAVE technique for assessing explanatory style*. Technical report, Virginia Polytechnic Institute and State University.

Plets, G., Huijnen, P. & van Oeveren, D. (2021). Excavating Archaeological Texts: Applying Digital Humanities to the Study of Archaeological Thought and Banal Nationalism. *Journal of Field Archaeology*, pp. 1–14. ISSN 0093-4690. DOI: 10.1080/00934690.2021.1899889.

Postma, M., van Miltenburg, E., Segers, R., Schoen, A. & Vossen, P. (2016). Open Dutch WordNet. In *Proceedings of the Eight Global Wordnet Conference*, pp. 300–308. Bucharest, Romania.

Powers, D. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology*, 2(1), pp. 37–63.

Pressman, R.S. (2005). *Software engineering : a practitioner's approach*. McGraw-Hill College, New York, NY, USA, 6th edition. ISBN 007301933X.

Rabinowitz, A., Shaw, R., Buchanan, S., Golden, P. & Kansa, E. (2016). Making Sense of the Ways we make Sense of the Past: The PeriodO Project. *Bulletin of the Institute of Classical Studies*, 59(2), pp. 42–55. ISSN 0076-0730. DOI: 10.1111/j.2041-5370.2016.12037.x.

Ramshaw, L.A. & Marcus, M.P. (1999). Text Chunking Using Transformation-Based Learning. In *Third Workshop on Very Large Corpora*, pp. 157–176. Association for Computational Linguistics. DOI: 10.1007/978-94-017-2390-9_10.

Rau, L.F. (1991). Extracting company names from text. In *Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications*, pp. 29–32. Publ by IEEE, Miami Beach, FL, USA. ISBN 0818621354. DOI: 10.1109/caia.1991.120841.

Renfrew, C. & Bahn, P.G. (2019). *Archaeology: theories, methods and practice (8th edition)*. Thames and Hudson London.

Richards, J., Tudhope, D. & Vlachidis, A. (2015). Text Mining in Archaeology: Extracting Information from Archaeological Reports. In J.A. Barcelo & I. Bogdanovic, eds., *Mathematics and Archaeology*, pp. 240–254. CRC Press, Boca Raton. DOI: 10.1201/b18530-15.

Rico, M., Vila-Suero, D., Botezan, I. & Gómez-Pérez, A. (2019). Evaluating the impact of semantic technologies on bibliographic systems: A user-centred and comparative approach. *Journal of Web Semantics*, 59. DOI: 10.1016/J.WEBSEM.2019.03.001.

Rieh, S. (2006). Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing & Management*, 42(3), pp. 751–768. DOI: doi.org/10.1016/j.ipm.2005.05.005.

Rijksdienst voor het Cultureel Erfgoed (2019a). Archeologisch onderzoek - aantal onderzoeksmeldingen | Erfgoedmonitor. https://erfgoedmonitor.nl/indicatoren/archeologisch-onderzoek-aantal-onderzoeksmeldingen.

Rijksdienst voor het Cultureel Erfgoed (2019b). Archis. https://archis.cultureelerfgoed.nl.

Rosenzweig, E. (2015). *Successful user experience: Strategies and roadmaps*. Elsevier, Waltham, MA, USA. ISBN 9780128010617. DOI: 10.1016/c2013-0-19353-1.

Roth, B.J. (2010). An Academic Perspective on Grey Literature. *Archaeologies*, 6(2), pp. 337–345. ISSN 1555-8622. DOI: 10.1007/s11759-010-9141-9.

Rural Riches project (2021). The Rural Riches database. https://www.merovingianarchaeology.org/blog/about/the-rr-database.

Russell-Rose, T., Chamberlain, J. & Azzopardi, L. (2018). Information retrieval in the workplace: A comparison of professional search practices. *Information Processing and Management*, 54(6), pp. 1042–1057. ISSN 03064573. DOI: 10.1016/j.ipm.2018.07.003.

Russell-Rose, T. & Shokraneh, F. (2020). Designing the Structured Search Experience: Rethinking the Query-Builder Paradigm. *Weave: Journal of Library User Experience*, 3(1). ISSN 2333-3316. DOI: 10.3998/weave.12535642.0003.102.

Salton, G. (1968). *Automatic Information Organization and Retrieval*. McGraw Hill Text. ISBN 0070544859.

Salton, G. (1971). *The SMART retrieval system: Experiments in automatic document processing*. Prentice-Hall, Upper Saddle River, NJ.

Sasaki, Y. (2007). *The truth of the F-measure*. Technical report, School of Computer Science, University of Manchester, Manchester. https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf.

Schmidt, S.C. & Marwick, B. (2020). Tool-Driven Revolutions in Archaeological Science. *Journal of Computer Applications in Archaeology*, 3(1), pp. 18–32. ISSN 2514-8362. DOI: 10.5334/jcaa.29.

Selhofer, H. & Geser, G. (2014). *D2.1: First Report on Users' Needs*. Technical report, ARIADNE. http://legacy.ariadne-infrastructure.eu/wp-content/uploads/2019/07/ARIADNE_D2-1_First_report_on_users_needs.pdf.

Seok, M., Song, H.J., Park, C.Y., Kim, J.D. & Kim, Y.S. (2016). Named entity recognition using word embedding as a feature. *International Journal of Software Engineering and its Applications*, 10, pp. 93 – 104. ISSN 17389984. DOI: 10.14257/ijseia.2016.10.2.08.

Seymour, D.J. (2010). Sanctioned Inequity and Accessibility Issues in the Grey Literature in the United States. *Archaeologies*, 6(2), pp. 233–269. ISSN 1555-8622. DOI: 10.1007/s11759-010-9144-6.

Sienčnik, S.K. (2015). Adapting word2vec to named entity recognition. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015*, 109, pp. 239–243. Linköping University Electronic Press, Vilnius, Lithuania.

Song, Y., Zhou, D. & He, L.W. (2011). Post-ranking query suggestion by diversifying search results. In *SIGIR'11 - Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 815–824. Association for Computing Machinery, Beijing, China. ISBN 9781450309349. DOI: 10.1145/2009916.2010025.

Sorel, D. (2018). jQuery QueryBuilder. https://querybuilder.js.org/.

Soto, A., Olivas, J.A. & Prieto, M.E. (2008). Fuzzy approach of synonymy and polysemy for information retrieval. *Studies in Fuzziness and Soft Computing*, 224, pp. 179–198. ISSN 14349922. DOI: 10.1007/978-3-540-76973-6_12.

Souza, F., Nogueira, R. & Lotufo, R. (2019). Portuguese Named Entity Recognition using BERT-CRF. *arXiv*. ISSN 23318422. http://arxiv.org/abs/1909.10649.

Spink, A. (2002). A user-centered approach to evaluating human interaction with Web search engines: an exploratory study. *Information Processing & Management*, 38(3), pp. 401–426. DOI: 10.1016/S0306-4573(01)00036-X.

Sporleder, C. (2010). Natural Language Processing for Cultural Heritage Domains. *Language and Linguistics Compass*, 4(9), pp. 750–768. ISSN 1749818X. DOI: 10.1111/j.1749-818X.2010.00230.x.

Steiner, C.M., Agosti, M., Sweetnam, M.S., Hillemann, E.C., Orio, N., Ponchia, C., Hampson, C., Munnelly, G., Nussbaumer, A., Albert, D. & Conlan, O. (2014). Evaluating a digital humanities research environment: the CULTURA approach. *International Journal on Digital Libraries*, 15(1), pp. 53–70. DOI: 10.1007/s00799-014-0127-x.

Stichting Infrastructuur Kwaliteitsborging Bodembeheer (2016). BRL 4000. https://www.sikb.nl/archeologie/richtlijnen/brl-4000.

Strubell, E., Ganesh, A. & McCallum, A. (2020). Energy and policy considerations for deep learning in NLP. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 3645–3650. Association for Computational Linguistics, Florence, Italy. ISBN 9781950737482. DOI: 10.18653/v1/p19-1355.

Talboom, L. (2017). *Improving the discoverability of zooarchaeological data with the help of Natural Language Processing.* Thesis, University of York.

Talja, S. (1997). Constituting "information" and "user" as research objects: A theory of knowledge formations as an alternative to the information man-theory. *Information seeking in context*, pp. 67–80.

Talks, A. (2019). *An exploration of NLP and NER for enhanced search in osteoarchaeological and palaeopathological textual resources.* Thesis, University of York.

Tanasi, D. (2020). The digital (within) archaeology. Analysis of a phenomenon. *The Historian*, 82(1), pp. 22–36. ISSN 0018-2370. DOI: 10.1080/00182370.2020.1723968.

Theunissen, L. & Feiken, R. (2014). *Analyse archeologische kenniswinst (2000 - 2014).* Technical report, Rijksdienst voor het Cultureel Erfgoed, Amersfoort.

Thomsett-Scott, B.C. (2006). Web site usability with remote users: Formal usability studies and focus groups. *Journal of Library Administration*, 45(3-4), pp. 517–547. ISSN 01930826. DOI: 10.1300/J111v45n03_14.

Tikhomirov, M., Loukachevitch, N., Sirotina, A. & Dobrov, B. (2020). Using bert and augmentation in named entity recognition for cybersecurity domain. In *Natural Language Processing and Information Systems*, volume 12089 LNCS, pp. 16–24. Springer International Publishing, Cham. ISBN 9783030513092. ISSN 16113349. DOI: 10.1007/978-3-030-51310-8_2.

Tjong Kim Sang, E.F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Tong, Z. (2018). elasticsearch-php. https://github.com/elastic/elasticsearch-php.

Traviglia, A. & Torsello, A. (2017). Landscape Pattern Detection in Archaeological Remote Sensing. *Geosciences*, 7(4), p. 128. ISSN 2076-3263. DOI: 10.3390/geosciences7040128.

Trier, Ø.D., Salberg, A.B. & Pilø, L.H. (2018). Semi-automatic mapping of charcoal kilns from airborne laser scanning data using deep learning. In *CAA2016: Oceans of Data. Proceedings of the 44th Conference on Computer Applications and Quantitative Methods in Archaeology*, pp. 219–231. Archaeopress Oxford.

Truyens, M. & Van Eecke, P. (2014). Legal aspects of text mining. *Computer Law & Security Review*. ISSN 02673649. DOI: 10.1016/j.clsr.2014.01.009.

Tudhope, D., May, K., Binding, C. & Vlachidis, A. (2011). Connecting archaeological data and grey literature via semantic cross search. *Internet archaeology*, 30. DOI: doi.org/10.11141/ia.30.5.

Tunkelang, D. (2009). Faceted Search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1), pp. 1–80. ISSN 1947-945X. DOI: 10.2200/s00190ed1v01y200904icr005.

Van den Bosch, A., Busser, B., Canisius, S. & Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In F. van Eynde, P. Dirix, I. Schuurman & V. Vandeghinste, eds., *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pp. 99–114. Leuven.

Van den Dries, M. (2016). Is everybody happy? User satisfaction after ten years of quality management in European archaeological heritage management. In P. Florjanowicz, ed., *When Valletta meets Faro, the reality of European archaeology in the 21st century, proceedings of the International Conference*, pp. 126–135. Archaeolingua, Lisbon.

Van Es, W. (1968). *Grafritueel en Kerstening*. Inaugural lecture Vrije Universiteit Amsterdam, Amsterdam.

Van Es, W. & Schoen, R. (2008). Het vroegmiddeleeuwse grafveld van Zweeloo. *Palaeohistoria*, 45/50, pp. 795–935. https://ugp.rug.nl/Palaeohistoria/article/view/25161.

Van Gompel, M. & Reynaert, M. (2013). FoLiA: A practical XML format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3, pp. 63–81.

Van Haperen, M. (2017). *Early Medieval Grave Reopenings in the Low Countries*. Thesis, Leiden University. DOI: 10.17026/dans-x6b-bvgj.

Van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T. & Van de Walle, R. (2015). Exploring entity recognition and disambiguation for cultural heritage collections. *Digital Scholarship in the Humanities*, 30(2), pp. 262–279. ISSN 2055-7671. DOI: 10.1093/llc/fqt067.

Van Waes, L. (2000). Thinking aloud as a method for testing the usability of Websites: the influence of task variation on the evaluation of hypertext. *IEEE Transactions on Professional Communication*, 43(3), pp. 279–291. DOI: 10.1109/47.867944.

Van Zundert, J.J. (2016). The case of the bold button: Social shaping of technology and the digital scholarly edition. *Digital Scholarship in the Humanities*, 31(4), pp. 898–910. ISSN 2055-7671. DOI: 10.1093/llc/fqw012.

Verberne, S., Boves, L. & Bosch, A. (2016). Information access in the art history domain. Evaluating a federated search engine for Rembrandt research. *Digital Humanities Quarterly*, 10(4), p. online. ISSN 1938-4122. http://www.digitalhumanities.org/dhq/vol/10/4/000265/000265.html.

Verberne, S., He, J., Kruschwitz, U., Wiggers, G., Larsen, B., Russell-Rose, T. & de Vries, A.P. (2019). First International Workshop on Professional Search. *ACM SIGIR Forum*, 52(2), pp. 153–162. ISSN 0163-5840. DOI: 10.1145/3308774.3308799.

Verschoof-van der Vaart, W. & Brandsen, A. (2020). Boundingbox Localizer Tool (BLT) - Brandenburgische Technische Universität Cottbus-Senftenberg version. *Zenodo Repository*. DOI: 10.5281/ZENODO.3888053.

Verschoof-Van Der Vaart, W.B., Lambers, K., Kowalczyk, W. & Bourgeois, Q.P. (2020). Combining deep learning and location-based ranking for large-scale archaeological prospection of LiDAR data from the Netherlands. *ISPRS International Journal of Geo-Information*, 9(5), p. 293. ISSN 22209964. DOI: 10.3390/ijgi9050293.

Verschoof-van der Vaart, W.B. & Landauer, J. (2021). Using CarcassonNet to automatically detect and trace hollow roads in LiDAR data from the Netherlands. *Journal of Cultural Heritage*, 47, pp. 143–154. ISSN 12962074. DOI: 10.1016/j.culher.2020.10.009.

Verwers, W. & van Tent, W. (2015). *Merovingisch grafveld Elst-'t Woud. Rapportage Archeologische Monumentenzorg 223*. Technical report, Rijksdienst voor het Cultureel Erfgoed, Amersfoort.

Vince, A. (1996). Editorial. *Internet Archaeology*, 1. ISSN 13635387. DOI: 10.11141/ia.1.7.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F. & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *arXiv*. http://arxiv.org/abs/1912.07076.

Vlachidis, A. (2012). Semantic Indexing via Knowledge Organization Systems: Applying the CIDOC-CRM to Archaeological Grey Literature. *Unpublished PhD Thesis, University of South Wales (USW)*.

Vlachidis, A., Binding, C., May, K. & Tudhope, D. (2013). Automatic metadata generation in an archaeological digital library: Semantic annotation of grey literature. *Studies in Computational Intelligence*, 458, pp. 187–202. ISSN 1860949X. DOI: 10.1007/978-3-642-34399-5_10.

Vlachidis, A. & Tudhope, D. (2012). A pilot investigation of information extraction in the semantic annotation of archaeological reports. *International Journal of Metadata, Semantics and Ontologies*, 7(3), p. 222. ISSN 1744-2621. DOI: 10.1504/IJMSO.2012.050183.

Vlachidis, A. & Tudhope, D. (2016). 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), pp. 1138–1152. DOI: 10.1002/asi.23485.

Vlachidis, A., Tudhope, D., Wansleeben, M., Azzopardi, J., Green, K., Xia, L. & Wright, H. (2017). *D16.4: Final Report on Natural Language Processing.* Technical report, ARIADNE. http://legacy.ariadne-infrastructure.eu/wp-content/uploads/2019/01/D16.4_Final_Report_on_Natural_Language_Processing_Final.pdf.

Voorhees, E. (2001). Overview of TREC 2001. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, pp. 1–13.

Wamers, E. (2015). *Das bi-rituelle Kinderdoppelgrab der späten Merowingerzeit unter der Frankfurter Bartholomäuskirche (»Dom«).* Schnell and Steiner, Regensburg.

Wei, J. & Zou, K. (2020). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 6382–6388. ISBN 9781950737901. DOI: 10.18653/v1/d19-1670.

Wen, M., Vasthimal, D.K., Lu, A., Wang, T. & Guo, A. (2019). Building large-scale deep learning system for entity recognition in e-commerce search. In *BDCAT 2019 - Proceedings of the 6th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, pp. 149–154. Association for Computing Machinery, Inc, New York, New York, USA. ISBN 9781450370165. DOI: 10.1145/3365109.3368765.

Wesson & Cottier (2014). Big Sites, Big Questions, Big Data, Big Problems: Scales of Investigation and Changing Perceptions of Archaeological Practice in the Southeastern United States. *Bulletin of the History of Archaeology*, 24(0), p. 16. ISSN 2047-6930. DOI: 10.5334/bha.2416.

Wheatley, D. (2004). Making space for an archaeology of place. *Internet Archaeology*, 15. DOI: 10.11141/ia.15.10.

Wilcke, W.X., de Boer, V., de Kleijn, M.T., van Harmelen, F.A. & Scholten, H.J. (2019). User-centric pattern mining on knowledge graphs: An archaeological case study. *Journal of Web Semantics*, 59, pp. 1–10. ISSN 15708268. DOI: 10.1016/j.websem.2018.12.004.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E.,

Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3. ISSN 2052-4463. DOI: 10.1038/sdata.2016.18.

Williams, H. (2014). A well-urned rest: Cremation and inhumation in early Anglo-Saxon England. In I. Kuijt, C.P. Quinn & G. Cooney, eds., *Transformation by fire: The archaeology of cremation in cultural context*, pp. 93–118. University of Arizona Press, Tucson.

Wiseman, R. & Ronn, P. (2020). *Archaeology on Furlough: Accessing Archaeological Information Online: A Survey of Volunteers' Experiences*. Technical report, Cambridge University. DOI: 10.17863/CAM.54876.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. & Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Association for Computational Linguistics, Stroudsburg, PA, USA. DOI: 10.18653/v1/2020.emnlp-demos.6.

Wu, S. & Dredze, M. (2020). Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pp. 120–130. Association for Computational Linguistics, Stroudsburg, PA, USA. DOI: 10.18653/v1/2020.repl4nlp-1.16.

Xiong, Y., Huang, Y., Chen, Q., Wang, X., Ni, Y. & Tang, B. (2020). A joint model for medical named entity recognition and normalization. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, 2664, pp. 499–504. ISSN 16130073.

Yamada, I., Asai, A., Shindo, H., Takeda, H. & Matsumoto, Y. (2020). LUKE: Deep Contextualized Entity Representations with Entity-aware Self-

attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6442–6454. Association for Computational Linguistics, Stroudsburg, PA, USA. DOI: 10.18653/v1/2020.emnlp-main.523.

Zubrow, E.B. (2006). Digital archaeology: A historical context. In P. Daly & T. Evans, eds., *Digital Archaeology: Bridging Method and Theory*. Routledge, London, 1st edition. ISBN 9780415310505. DOI: 10.4324/9780203005262.

Appendices

A

Category frequencies

Site type categories frequency overview									
Label	Freq	Label	Freq	Label	Freq	Label	Freq	Label	Freq
xxx	33532								
cthd	1463	bewv.wp	10	**idnh**	2015	sv.vorg	0	bgv.meg	9
cthd.x	379	bewv.n	0	idnh.x	1074	sv.bsb	0	**bgr**	1502
cthd.klo	299	bewv.rv	501	idnh.tk	0	**gw**	169	bgr.gvic	1502
cthd.kpl	24	bewv.stel	4	idnh.tn	128	gw.x	40	**infr**	7327
cthd.sgmw	0	bewv.bw	0	idnh.br	36	gw.vw	58	infr.x	1248
cthd.kerk	581	bewv.hp	1566	idnh.zp	0	gw.hout	0	infr.weg	1575
cthd.rcp	372	bewv.th	0	idnh.sb	71	gw.ijw	9	infr.dam	53
cthd.oloc	1	bewv.inka	0	idnh.hkb	127	gw.zw	3	infr.werf	0
cthd.temp	2	bewv.sv	0	idnh.bb	5	gw.kw	50	infr.gem	5
bewv	25264	bewv.bext	5872	idnh.ll	112	gw.griw	0	infr.rede	0
bewv.x	15236	bewv.vkm	63	idnh.hb	4	gw.mw	4	infr.per	3766
bewv.lg	89	bewv.tw	388	idnh.m	150	gw.vsw	8	infr.strek	0
bewv.wb	51	bewv.lw	130	idnh.rom	1	**bgv**	6317	infr.wat	228
bewv.sch	65	**apvv**	3152	idnh.wam	2	bgv.x	731	infr.dui	221
bewv.vx	815	apvv.x	1415	idnh.wim	1	bgv.gvc	522	infr.vijv	0
bewv.vlp	102	apvv.vw	0	idnh.gp	0	bgv.tpgb	0	infr.kan	273
bewv.lk	0	apvv.vk	166	idnh.pb	227	bgv.gvi	536	infr.slu	103
bewv.ct	0	apvv.vs	6	idnh.vb	312	bgv.gvx	983	infr.kslu	0
bewv.cstl	5	apvv.stel	0	idnh.mb	388	bgv.kh	605	infr.lv	0
bewv.mbh	125	apvv.ek	0	idnh.mbnf	0	bgv.rgv	1	infr.hav	982
bewvv.pls	0	apvv.cf	23	idnh.mbf	0	bgv.ghv	2476	infr.kade	1
bewv.kwb	490	apvv.dp	142	idnh.kb	2	bgv.bhv	0	infr.vweg	7
bewv.ht	332	apvv.la	1879	**sv**	971	bgv.vgv	0	infr.brug	256
bewv.aw	7	apvv.ak	0	sv.x	971	bgv.cjbp	536	infr.dok	0
bewv.dump	0	apvv.tuin	20	sv.obsb	0	bgv.uv	1295	infr.vs	0
bewv.vic	332	apvv.pdek	2	sv.ijz	0	bgv.gh	1221	infr.vrde	1
bewv.kaze	0	**wrak**	384	sv.h	0	bgv.gx	731	infr.spre	0
bewv.fort	8	wrak.schip	384	sv.lad	0	bgv.vg	163	infr.watw	11
bewv.sk	2470	wrak.vlgtg	5	sv.hijz	0	bgv.dier	131	infr.dij	721

Table A.1: An overview of the frequencies for all site type categories. Main categories are denoted in bold.

Time periods categories frequency overview									
Label	Freq	Label	Freq	Label	Freq	Label	Freq	Label	Freq
paleo	2077	neov	6459	bronsm	8494	romvb	12414	vmec	11767
paleov	1197	neova	6456	bronsma	8397	romm	12427	vmed	12194
paleom	1460	neovb	6445	bronsmb	8312	romma	12381	lme	18832
paleol	1816	neom	6127	bronsl	7910	rommb	12348	lmea	17053
paleola	1732	neoma	6098	**ijz**	13876	roml	11939	lmeb	18235
paleolb	1816	neomb	5893	ijzv	10356	romla	11921	**nt**	19833
meso	4290	neol	8954	ijzm	11307	romlb	11850	nta	17511
mesov	3133	neola	8200	ijzl	12033	**xme**	20593	ntb	17514
mesom	3152	neolb	8947	**rom**	13299	vme	12642	ntc	18525
mesol	4180	**brons**	10414	romv	12421	vmea	11645		
neo	9916	bronsv	8380	romva	12275	vmeb	11874		

Table A.2: An overview of the frequencies for all time period categories. Main categories are denoted in bold.

B

Filter list

Terms used for document filtering	
List name	**Terms**
genList	notulen, bijlage, meta
rapList	dagrapport, dag_rapport, weekrapport, week_rapport, weekverslag, week_verslag, logboek
pvaList	draaiboek, plan_van_aanpak, pva
omnList	onderzoeksmeldingsnummer, onderzoeksmeldings_nummer, onderzoeks_meldings_nummer
totList	rapList + pvaList + pveList + omnList + genList

Table B.1: An overview of different types of lists and included terms.

C

Category frequencies test set

Time periods categories frequency overview test set									
Label	Freq	Label	Freq	Label	Freq	Label	Freq	Label	Freq
paleo	13	neov	24	bronsm	15	romvb	25	vmec	30
paleov	7	neova	24	bronsma	14	romm	27	vmed	29
paleom	8	neovb	24	bronsmb	15	romma	27	lme	48
paleol	12	neom	23	bronsl	18	rommb	27	lmea	41
paleola	12	neoma	23	**ijz**	37	roml	25	lmeb	48
paleolb	12	neomb	23	ijzv	34	romla	25	**nt**	49
meso	21	neol	26	ijzm	28	romlb	25	nta	42
mesov	17	neola	25	ijzl	30	**xme**	54	ntb	43
mesom	18	neolb	26	**rom**	29	vme	30	ntc	39
mesol	20	**brons**	24	romv	25	vmea	28		
neo	29	bronsv	19	romva	25	vmeb	28		

Table C.1: An overview of the frequencies for all time period categories captured by the reference test set. Main categories are denoted in bold.

Site type categories frequency overview test set									
Label	Freq	Label	Freq	Label	Freq	Label	Freq	Label	Freq
cthd	1	bewv.hp	7	idnh.hkb	2	bgv.x	4	**bgr**	3
cthd.klo	1	bewv.bext	3	idnh.ll	1	bgv.gvc	2	bgr.gvic	3
bewv	65	**apvv**	8	idnh.m	1	bgv.gvi	3	**infr**	19
bewv.x	53	apvv.x	3	idnh.pb	1	bgv.gvx	3	infr.x	1
bewv.vx	1	apvv.cf	1	idnh.vb	2	bgv.kh	1	infr.weg	4
bewv.vlp	1	apvv.la	4	idnh.mb	2	bgv.ghv	6	infr.per	6
bewv.kwb	1	**wrak**	2	**sv**	1	bgv.cjbp	3	infr.kan	2
bewv.ht	10	wrak.schip	2	sv.x	1	bgv.uv	4	infr.brug	2
bewv.vic	1	**idnh**	11	**gw**	1	bgv.gx	4	infr.dij	5
bewv.sk	4	idnh.x	4	gw.vw	1	bgv.vg	1	**xxx**	17
bewv.rv	1	idnh.tn	1	**bgv**	16	bgv.dier	1		

Table C.2: An overview of the F1 scores for the main and sub-categories for site type classification as captured by the reference test set. Sub-categories not present within the reference test set are not included. Again, main categories are denoted in bold.

D

Curriculum Vitae

Alex **Brandsen**

PhD Candidate · Code Wrangler

Leiden University, Faculty of Archaeology, Einsteinweg 2, 2333CC Leiden, The Netherlands

+31 681 833 764 | a.brandsen@arch.leidenuniv.nl | alexbrandsen.nl | alexbrandsen | alex-brandsen | @alex_brandsen

"Programming isn't about what you know; it's about what you can figure out." - Chris Pine

Summary

I studied archaeology for my BA and MSc, but always with a focus on digital techniques and web technologies. After my studies I pursued a career in web development, working on front and back end as well as server maintenance and client liaison for six years. I completed a PhD in Digital Archaeology at Leiden University in 2022, researching the use of text mining in archaeological grey literature. I am currently a postdoc in the EXALT project, further building on my PhD research.

I'm a fast learner, tinkerer, and passionate about open science and reproducibility.

Work Experience

Leiden University
Leiden, The Netherlands

Postdoc
June 2021 - PRESENT

- Further building on my PhD research in the EXALT project
- Creating a multilingual search engine for all available archaeological texts about the Netherlands and surrounding areas
- Developing and teaching courses at BA and MA level

Leiden University
Leiden, The Netherlands

PhD Candidate
May 2017 - May 2021

- Investigating the use of Text Mining techniques to make archaeological excavation reports more accessible
- Using machine learning techniques to perform Named Entity Recognition
- Building an intuitive search UI for archaeologists
- Teaching assistance in various courses (Databases, Text Mining, etc)

Space Creative / The Wrapped Agency
Leeds, United Kingdom

Web Developer
Jan 2010 - Apr. 2017

- Developing websites on the Magento eCommerce and Concrete5 CMS platforms, as well as designing and developing custom applications
- Worked on complex projects from initial specification and database design right up to the final stages of responsive testing
- Gained experience in Linux sysadmin & server maintenance, custom (Google) mapping, geographical searches, XML, complex jQuery applications, Photoshop, Inkscape & Illustrator

Impulse Media
York, United Kingdom

Freelance Web Developer
Sep. 2010 - Dec. 2010

- Mainly working for Impulse Media, developing PHP web applications, converting PSD files to HTML/CSS templates, developing iPhone apps and working with Concrete5 CMS and eCommerce platform Magento

BioArch, University of York
York, United Kingdom

Internship: Assistent Web Developer
Jan. 2010 - Jun. 2010

- Assisted David Harker in developing SHAARKWeb, a web-based UI that inputs and processes information from users of the NEAAR lab
- Used Object-Oriented PHP, Propel, XHTML and CSS to develop parts of the system; mainly the login system, AJAX dropdown boxes populated by external database queries and programmatically importing MS Excel spreadsheets into the database

Antiquity Journal
York, United Kingdom

Internship: Assistent Web Designer
Oct. 2009 - Dec. 2009

- Created a new Project Gallery Archive homepage

The Open Boek Project, at Rijksdienst voor Cultureel Erfgoed
Amersfoort, The Netherlands

Internship: Assistant Web/Software Developer
Nov. 2008 - Mar. 2008

- Assisted in making the smart index- and search engine 'Open Boek' more user-friendly for archaeologists
- Adapted the system to be able to process Dutch texts as well as English texts
- Gained experience with LaTeX, MySQL, PHP, AWK and Bash-shell scripting

Education

University of York — *York, United Kingdom*

MSc in Archaeological Information Systems — *Sep. 2010*

- Grade: 1st Class Distinction
- Major in Archaeological Information Systems
- Main Topics: Web Design, Database Design, Geographical Information Systems & Virtual Reality
- Minor in Zooarchaeology
- Dissertation: Digital Medieval Graffiti; Using online-GIS to display and edit non-geographical data

Leiden University — *Leiden, The Netherlands*

BA in Archaeology — *Sep. 2009*

- Grade: 7.5
- Major in European Prehistory
- Minor in Archaeological Information Systems (Database Design, GIS & Use of Total Station)
- Additional courses: Physical Anthropology
- Thesis Topic: Using Ground Penetrating Radar to assess burial mounds

Tabor College, Location Werenfridus — *Hoorn, The Netherlands*

VWO (Pre-university secondary education) — *Jul. 2005*

- Specialisation: Economics and Society. Main topics are History and Economics
- Additional course: Computer Science (MS Access, SQL, Java & HTML)

Publications

Brandsen, A, & Lippok, F, 2021. A burning question – Using an intelligent grey literature search engine to change our views on early medieval burial practices in the Netherlands. *Journal of Archaeological Science*, 133. DOI: 10.1016/j.jas.2021.105456

Brandsen, A & Koole, M, 2021. Labelling the Past: Data Set Creation and Multi-label Classification of Dutch Archaeological Excavation Reports. *Language Resources and Evaluation*. DOI: 10.1007/s10579-021-09552-6

Brandsen, A, Verberne, S, Lambers, K & Wansleeben, M, 2020. Creating a Dataset for Named Entity Recognition in the Archaeology Domain. *Proceedings of the 12th Language Resources and Evaluation Conference*, pp.4573-4577. URL: www.aclweb.org/anthology/2020.lrec-1.562

Brandsen, A, Lambers, K, Verberne, S & Wansleeben, M, 2019. 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), pp.21–30. DOI: 10.5334/jcaa.33

Paijmans, H & Brandsen, A, 2010. Searching in Archaeological Texts. Problems and Solutions Using an Articial Intelligence Approach, *PalArchs Journal of Archaeology of Egypt/Egyptology*, 7(2).

Paijmans, H & Brandsen, A, 2009. What is in a Name: Recognizing Monument Names from Free-Text Monument Descriptions, *Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning*. Tilburg centre for Creative Computing, Tilburg.

Selected Presentations

2021	**CLIN31,** Can BERT Dig It? Named Entity Recognition for Information Retrieval in the Archaeology Domain	*Online*
2020	**Keeping Archaeology Together,** Using Transfer Learning for NER in Dutch Excavation Reports	*Online*
2020	**CLIN30**, BERT-NL: a set of language models pre-trained on the Dutch SoNaR corpus	*Utrecht*
2019	**Machine Learning in Archaeology**, Using Machine Learning for NER in Dutch Excavation Reports	*Rome*
2019	**CAA**, User Interface Design and Evaluation for Online Professional Search in Dutch Archaeology	*Krakow*
2019	**ICT Open,** Brinkey-generator - Computer Assisted Assignment of Thesaurus Topics for Scientific Texts	*Hilversum*
2018	**EAA**, Utilising Text Mining to Unlock the Hidden Knowledge in Dutch Archaeological Reports	*Barcelona*
2018	**DHBenelux 2018,** Knowledge dissemination and discovery in Dutch excavation data	*Amsterdam*
2017	**DIR2017,** Archaeological Entity Recognition for Information Retrieval in Dutch Archaeological Reports	*Hilversum*

Committees

2020	**Committee member**, ARCHON Digital Archaeology Workgroup	*Amsterdam*
2020	**Outreach Officer**, CAA NL/FL	*Amsterdam*
2019	**Main organiser**, Digital Archaeology Group	*Leiden*
2019	**Committee member**, ARCHON Digital Archaeology Workgroup	*Leiden*
2018	**Committee member**, Dutch-Belgian Information Retrieval Workshop 2018	*Leiden*
2018	**Committee member**, Digital Archaeology Workshops Netherlands-Flanders (DAWN) 2018	*Leiden*

Glossary

ABR *Archeologisch Basisregister*. 33, 52–57, 62, 63, 82, 89, 124

ADS Archaeology Data Service. 30, 79

AGNES Archaeological Grey-literature Named Entity Search. 3–10, 37, 48, 53, 78, 79, 83, 90, 92–99, 110, 111, 140, 142–146, 148–151, 153, 154, 156, 158–161, 167–171, VI, X

ALICE Academic Leiden Interdisciplinary Cluster Environment. 33

API Application Programming Interface. 30, 79, 90, 125

Archis *Archeologisch Informatiesysteem*, a system for registering and accessing data about archaeological research, finds and monuments in the Netherlands, maintained by the RCE. 87, 94, 95, 141, 150, 151, 153, X

ARIADNE Advanced Research Infrastructure for Archaeological Dataset Networking in Europe. 30, 31, 34, 37, 38, 79, 81, 94, 174

BERT Bidirectional Encoder Representations from Transformers. 4, 8, 9, 52, 63, 64, 66, 73, 114, 116–121, 123, 126–128, 130, 135–137, 144, 155, 164, 165, 167, 168, 172, 175, XIII

Bi-LSTM Bidirectional Long Short Term Memory. 30, 118, 120

BIO Beginning, Inside, Outside. 27

CATCH Continuous Access To Cultural Heritage. 78

Corpus A large collection of texts. It is a body of written or spoken material upon which a linguistic analysis is based. 7, 8, 32, 36, 38, 40, 45, 60, 77, 81, 90, 115, 116, 120, 123, 124, 130, 131, 136, 159, 162, 163, 165, XIII

CRF Conditional Random Fields. 8, 30, 42, 43, 45, 76, 80, 82, 90, 114, 116–119, 123, 124, 126, 128, 135, 137, 164, 175

CSV Comma Separated Values. 145, 155

DANS Data Archiving and Networked Services. 2, 3, 19, 32, 38, 53, 54, 59, 77, 82, 83, 87, 90, 94–97, 115, 125, 140, 144, 153, 167, 169, 171, 173

Deep Learning A subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks. These Deep Learning algorithms are more complex and computationally expensive when compared to traditional Machine Learning approaches, but also generally produce better results. 32, 33, 165

DSRP Data Science Research Programme. 33, 77, 79

ElasticSearch An open source, full-text search engine. Used for all the indexing and retrieval tasks in this project. 76, 82, 83, 97, 116, 125, 126, 130, 145

EXALT EXcavating Archaeological LiTerature. 167, 170–172

F1 Score The F1 score (or F measure) combines recall and precision to provide an overall evaluation metric. More specifically, it is the harmonic mean of precision and recall. 6, 8, 28, 29, 34, 36, 37, 41–43, 45, 49, 51, 52, 63, 64, 66, 67, 69, 72, 74, 114, 116, 120, 121, 123, 124, 126, 128, 132, 136, 137, 145, 165, 166, 174

FAIR Findability, Accessibility, Interoperability, Reusability. 16, 20, 21, 158, 162, 163

GATE General Architecture for Text Engineering. 31, 79

GIS Geographical Information Systems. 90

Gold Standard A test set of human annotated documents describing the desirable system outcome. 31, 34, 82, 88–90

Grey Literature Materials and research produced by organisations outside of the traditional commercial or academic publishing and distribution channels. 2, 3, 8, 9, 16–18, 22, 29–32, 36, 48, 76–79, 87, 90, 92, 96, 110, 114, 140–143, 146, 154

HTTP Hypertext Transfer Protocol. 82

IAA Inter Annotator Agreement. 8, 36–39, 41, 121, 135, 165, 167, 174

Information Need A user's end goal in a specific search session, or a description of the information or the answer they are looking for. 8, 23, 24, 29, 48, 54, 90, 92, 93, 97, 100–104, 115, 116, 125, 137, 146, 148, 149, 154, 156

IR Information Retrieval. 22–24, 29–31, 76, 78, 86, 87, 92, 93, 95, 96, 114, 118, 119, 166, 174

JSON JavaScript Object Notation. 7, 82, 83, 125, 162

KB *Koninklijke Bibliotheek*. 19, 77, 90, 94, 167

LIACS Leiden Institute of Advanced Computer Science. 33
LOD Linked Open Data. 172

MEAN Miscellaneous, Exceptional, Arbitrary, Nonconformist. 162
Metadata Data that provides information about other data, often describing certain properties of a data set. 2, 6, 8, 21, 29, 30, 36, 48–50, 52–57, 59, 73, 77, 79, 80, 92, 94, 95, 97, 104, 114, 115, 119, 120, 125, 141, 153, 154, 161, 167, 173

NER Named Entity Recognition. 7, 8, 24–31, 33, 34, 36–38, 45, 50, 63, 73, 76, 78–82, 87–90, 97, 104, 114, 116–121, 123, 125, 126, 128, 130–132, 137, 141, 142, 144, 148, 153–155, 165, 166, 168, 171, 172, 174, 175
NLP Natural Language Processing. 22, 27, 29, 30, 51–53, 93, 118, 120, 144, 170
NOaA *Nationale Onderzoeksagenda Archeologie*. 162
NWO *Nederlandse organisatie voor Wetenschappelijk Onderzoek*. 19

OCR Optical Character Recognition. 15, 22, 38, 163

PDF Portable Document Format. 21, 31, 32, 36, 38, 49, 53, 83, 96, 97, 115, 141, 142, 163, 164, 173
POS Part Of Speech. 27, 81
Precision An evaluation measure that indicates, out of all the labelled entities, what percentage has been assigned the correct label. 6, 28, 29, 52, 64, 81, 82, 88, 92, 103, 115, 126, 128, 130, 134, 135, 148, 155, 156, 165, 166, 168
Python A widely used high-level programming language, used for most of the programming in this project. 81, 164

RCE *Rijksdienst voor het Cultureel Erfgoed*. 19, 33, 53, 77, 82, 85, 87, 89, 90, 94, 141, 161
RDF Resource Description Framework. 30
Recall An evaluation measure that indicates out of all the entities in a text, what percentage have been correctly labelled as an entity. 6, 24, 28, 29, 51, 52, 63, 64, 67, 69, 81, 82, 88, 92, 103, 115, 116, 125, 126, 128–130, 134, 137, 148, 165, 166, 168, 169, XII, XIII

SIKB *Stichting Infrastructuur Kwaliteitsborging Bodembeheer*. 19, 77, 173
SVM Support Vector Machine. 52, 62, 66, 67, 72, 164, 167

Text Mining The process of analysing text to extract information from it. 2, 3, 9, 12, 21, 24, 30, 31, 48, 76–79, 169
TF-IDF Term Frequency - Inverse Document Frequency. 126, 167

UI User Interface. 97, 109–111

XML eXtensible Markup Language. 34, 53–55, 81, 173