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

Brandsen, A.

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