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

Brandsen, A.

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).