Usability Evaluation for Online Professional Search in the Dutch Archaeology Domain

Alex Brandsen*
Leiden University
a.brandsen@arch.leidenuniv.nl

Suzan Verberne
Leiden University
s.verberne@liacs.leidenuniv.nl

Karsten Lambers
Leiden University
k.lambers@arch.leidenuniv.nl

Milco Wansleeben
Leiden University
m.wansleeben@arch.leidenuniv.nl

Abstract

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

Key Words: *Information Retrieval, Usability Evaluation, User Interface Evaluation, Text Mining*

1 Introduction

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 Besides scholarly publications, another large source of documents are unpublished technical fieldwork reports.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

In the Netherlands, the amount of reports created in the last twenty years is currently estimated at around 60,000, and is growing by approximately 4000 per year (RCE, 2017). Most of these reports are categorised as 'grey literature' (Evans, 2015), and are likely to end up in a proverbial 'graveyard', unread and unknown, unless they are properly archived, indexed, and disclosed.

*Corresponding author

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

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

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

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

The following research questions are addressed in this paper:

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

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

The structure of the rest of this paper is as follows: Section 2 provides an overview of related and prior work; Section 3 is a short introduction to the current version of our system; Section 4 presents the set up of the user study with the results presented in Section 5, followed by a discussion in Section 6. Section 7 describes our conclusions and future work.

2 Background

2.1 Access to archaeological data

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

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

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

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

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

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

2.2 *Feedback on existing systems from our user group*

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

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

2.3 Related work in usability studies

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

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

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

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

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

3 AGNES

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

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

Named entities are terms that refer to important concepts from the real world (Marrero *et al.*, 2013). In the context of this project, the entities are archaeological concepts, mentioned in excavation reports. To give an example, in the following sentence the entities are bold: 'The **burial mound** yielded some **scrapers** from the **Neolithic**', a context, an artefact, and a time period, respectively. The example illustrates that entities can consist of multiple words. Two particular challenges of entity recognition are that a single term can refer to multiple entity types (e.g. 'Swifterbant' can be either a location, a time period, or a type of pottery), and that multiple terms can refer to the same entity (e.g. 'Neolithic' and 'New Stone Age'). For more technical information on the NER process, as well as the methods used, see Brandsen *et al.* (2020).

In AGNES, archaeological entities are recognised and labelled during the indexing of the documents. In version 0.3 of AGNES, all 60,000 reports from the DANS repository were indexed.

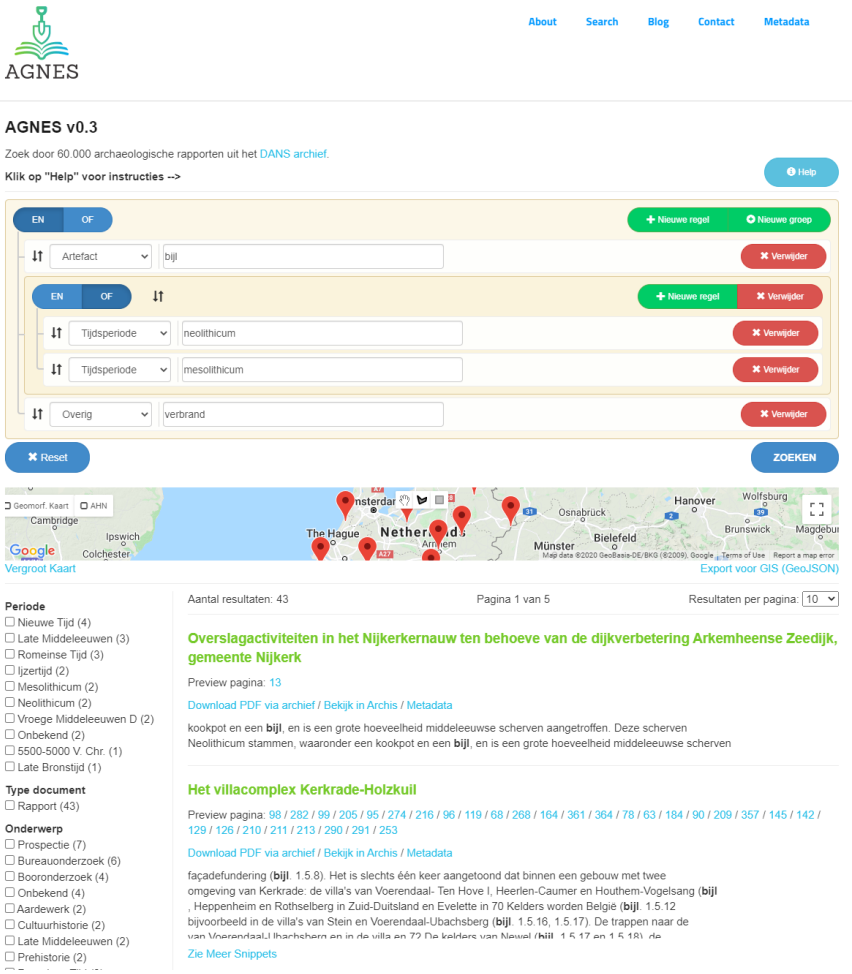

Fig. 1: *Screenshot of AGNES version 0.3. Pictured here is a query for 'artefact:axe AND (period:neolithic OR period:mesolithic) AND fulltext:burnt', with the results on a map and in a list underneath (with snippets). On the left we can see the facets, used to filter results on period, type of document, and subject.*

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

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

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

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

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

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

See Fig. 1 for a screenshot of the AGNES UI. This version is the one that has been evaluated in the current study, and is available at `http://agnessearch.nl/v03`.[2]

4 User Study Setup

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

4.1 Workshops in the AGNES project

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

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

4.2 Compilation of the focus group

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

This group has been selected in such a way that it includes every category of the target audience as defined in Brandsen *et al.* (2019). At the current stage of research, the group consists of six academics, two commercial professionals, and two archaeologists working on different levels in government. See Table 1 for a more detailed break down of the participants.

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

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

Table 1: *Overview of participants in usability evaluation per category*

Regarding the size of the focus group, Nielsen and Landauer (1993) show that the number of additional usability problems encountered when adding more users quickly decreases beyond five users. Thus, the current size of the focus group should be more adequate in this regard.

4.3 Design and procedure

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

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

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

To better understand the user behaviour, we asked the participants to use the Think Aloud Protocol, as introduced in Section 2.3. Specifically, we asked the participants to say what they think, see, expect, do, feel, and motivate their actions. At the end of the session we also asked the user a number of questions which can be found below.

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

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

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

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

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

5 Analysis and Results

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

A total of 148 queries were observed and recorded during the evaluation sessions, for a total of sixty-four information needs, making for an average of 2.3 queries per task. A query is defined here as a combination of search terms entered into the system, an information need as a defined question the researcher wants to answer. The minimum number of query elements is one, as expected, and the maximum is ten, with an average of 2.4. Here, an element of a query is one AND/OR statement, so for example the pseudo query `[artefact : scraper] AND [period : neolithic]` contains two elements.

5.1 Information Needs

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

In our Think Aloud sessions, we asked the users to also state the question they wanted to answer, and noted this down. We noticed that almost all the questions asked by the users are list questions, e.g. the three tasks mentioned in Table 2. This intuitively makes sense for archaeologists, as research most often entails making a list of all known occurrences of a particular topic and then performing some sort of analysis on this list. In our user requirements study, the users also indicated a preference for high recall over high precision, they much prefer getting all the relevant results with some noise, than to miss some results and have only relevant results Brandsen *et al.* (2019).

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

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

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

Find all amber from the Middle Neolithic

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

Find all beakers from graves in the late Neolithic

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

Find all coprolites from the Swifterbant period

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

Table 2: *Three examples of user generated tasks and their associated queries and query reformulations (translated from Dutch).*

5.2 Query Strategies and Effectiveness

We analysed the query reformulation strategies in this data, the process of altering a query to be narrower (specification, making the query longer), broader (generalisation, making the query shorter) or replacing one or more terms by other terms without making the query broader or narrower (parallel movement / reformulation) Rieh (2006). Interestingly, there is no trend to be found across all users between specification and generalisation, with both types of query reformulations occurring almost equally (twenty-five and twenty-four times, respectively). We do note that some users have a tendency to start broad and narrow down, while others do the opposite, but this seems to be a personal preference and not a preference for particular user categories. The full data is available via Zenodo[6], and there are three examples in Table 2.

While the users let us know in the feedback that they liked the faceted search (see Section 5.3 below), when we look at the queries we see that they don't use the facets very often. Out of 148 queries, only 23 include the use of facets (15.5%).

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

5.2.1 Query Effectiveness

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

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

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

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

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

5.3 Evaluation and User Satisfaction

5.3.1 Comments per User Group

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

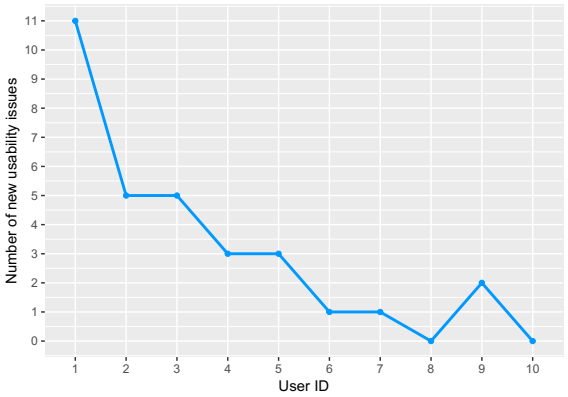

Fig. 2: *Line plot showing the number of new issues raised for each user*

5.3.2 Cumulative New Issues per User

The users mentioned a total of sixty-eight usability issues, averaging 6.8 per user. Where two or more users mentioned the same issue, we grouped it, which leads to a total of thirty-one unique issues. Fig. 2 shows the number of new usability issues found for each user that is added to the evaluation. We can see that after the fifth user evaluation, new users tend to not identify many new issues, confirming prior work on usability studies. The exception is user 9 with two new issues, who is the only commercial excavation archaeologist in our user group. This again underlines the necessity for a diverse user group.

5.3.3 Positive and Negative elements

From the answers to the questionnaire after each session, we got the impression that overall, the users find the system fairly easy to use and clear. The map functionality is mentioned by everyone, and mentioned often, something that was expected by the results of the user requirement solicitation. In Fig. 3 we have plotted a word cloud of all feedback, after translation from Dutch to English. We lowercased all text, removed punctuation, removed stopwords (NLTK list), and then plotted only words which occurred more than once.

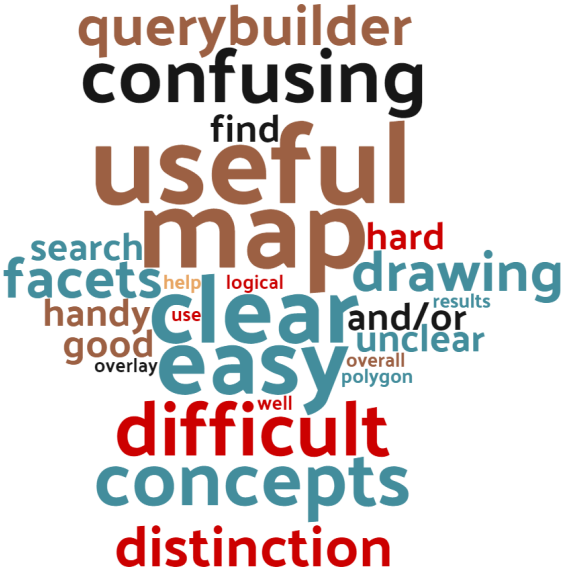

Fig. 3: *Word cloud of all feedback given, both positive and negative (translated from Dutch to English, 'ahn' is the height model of the Netherlands)*

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

In table 3 we show the most frequent words for the positive and negative feedback fields, respectively, where we have removed all stop words, verbs and opinion-bearing words (such as 'clear', 'hard'). Again we only include words mentioned more than once.

On the negative side we can see that choosing which concept to search for, intuitiveness, the 'help' button, the facets, and the AND/OR toggle buttons are elements that are commonly experienced as negative at the moment. These issues and features will be dealt with in the next

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

Table 3: *Feedback split into positive and negative, with for each word how often it occurs in that context. Words only mentioned once are not included.*

version of the system. We also see that the map, query builder, facets, and overall usability are often experienced as positive.

One of the other observations made during the evaluation is that none of the users use, or even see, the 'Help' button, which we did expect them to. This led to some preventable confusion about the system, as some questions the users had were actually explained in the help section. As a solution, we will include in-context help in the next version; pop ups that appear when hovering on certain elements to further explain the system.

5.3.4 Time Spent per Query

As mentioned before, we observed sixty-four research questions, with a total of 148 queries, so 2.3 query reformulations on average. For each initial query and query reformulation, we recorded the time taken to (re-)formulate the query and the number of elements in the query, among other information. We use the time per element instead of time per query to account for the difference in length of query between users, this way we can easily compare them. In Fig. 4 we plotted the time per element against the succession of queries attempted by a user. Here we see a clear downward trend (average between users shown in black). As the users had to do at least three of their own tasks, but could continue with more if they wanted, means that we have less data between query 6 and 9. However the trend is already clear between query 1 and 6.

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

6 Discussion

Gibbs *et al.* (2012) talk about how the typical humanities user is often neglected in the design of tools, and how tools' visibility can be increased by good attention to the usability. More recent work by Bulatovic *et al.* (2016) agrees, and states that digital humanities tools often suffer from poor user experiences, mainly caused by the lack of resources spent on usability research.

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

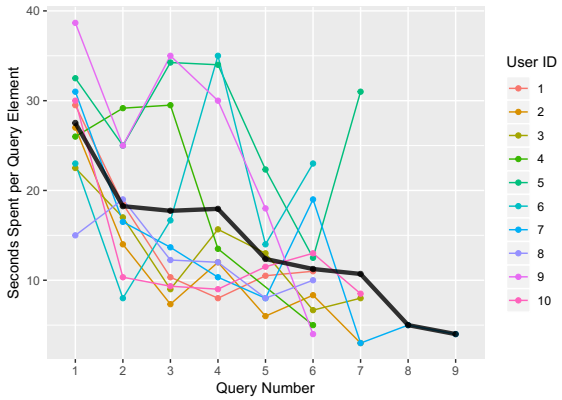

Fig. 4: *Line plot showing for each user, how much time they spent formulating one element of a query, for each new query they attempted. The black line is the average over all the users.*

Bulatovic *et al.* (2016) also mention that early iterative cycles of testing should be implemented in these kinds of projects, to avoid common usability problems. This is what we are doing in this project, and we hope this will see more uptake in the digital humanities, as it seems usability evaluation is something done at the end of most projects as an afterthought, if done at all. In 2012, Gibbs *et al.* called for a shift to user-centred design techniques, and luckily we do see that most of the more recent studies take this approach (e.g. Hinrichs *et al.*, 2015; van Zundert, 2016; Esmailpour *et al.*, 2019).

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

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

7 Conclusions

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

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

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

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

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

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

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

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

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

7.1 Future Work

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

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

At the moment, we only evaluated the system using ten users. We believe that a quantitative study using statistics generated by the system could be useful in finding usability issues, as well as seeing patterns in usage. To this end we will make the next version of the system public and invite a large group of archaeologists to use the system. This should give us a much larger user

group, although this is a more superficial analysis and loses some of the depth of evaluations done on the current group with the one-on-one approach.

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

References

Amrani, A., Abajian, V. and Kodratoff, Y. (2008). A chain of text-mining to extract information in archaeology. In Information and Communication Technologies: From Theory to Applications, ICTTA 2008., pp. 1–5. Damascus, Syria. doi:10.1109/ICTTA.2008.4529905.

Barbour, R. (2018). Doing focus groups. Sage, Los Angeles London.

Bartalesi, V., Meghini, C., Metilli, D. and Andriani, P. (2016). Usability Evaluation of the Digital Library DanteSources. In International Conference on Theory and Practice of Digital Libraries, pp. 191–203. Springer, Cham. doi:10.1007/978-3-319-39513-5{_}18.

Behnert, C. and Lewandowski, D. (2017). A framework for designing retrieval effectiveness studies of library information systems using human relevance assessments. Journal of Documentation, 73(3): 509–527. doi:10.1108/JD-08-2016-0099.

Brandsen, A., Lambers, K., Verberne, S. and Wansleeben, M. (2019). User Requirement Solicitation for an Information Retrieval System Applied to Dutch Grey Literature in the Archaeology Domain. Journal of Computer Applications in Archaeology, 2(1): 21–30. doi:10.5334/jcaa.33.

Brandsen, A., Verberne, S., Wansleeben, M. and Lambers, K. (2020). Creating a Dataset for Named Entity Recognition in the Archaeology Domain. In Proceedings of The 12th Language Resources and Evaluation Conference, p. 4573–4577. European Language Resources Association, Marseille, France.

Bulatovic, N., Gnadt, T., Romanello, M., Stiller, J. and Thoden, K. (2016). Usability in digital humanities – Evaluating user interfaces, infrastructural components and the use of mobile devices during research process. In **Fuhr, N., Kovács, L., Risse, T. and Nejdl, W.**, eds., Research and Advanced Technology for Digital Libraries. TPDL 2016. Lecture Notes in Computer Science, volume 9819 LNCS, pp. 335–346. Springer, Cham. doi:10.1007/978-3-319-43997-6{_}26.

Byrne, K. and Klein, E. (2010). Automatic Extraction of Archaeological Events from Text. In **Frischer, B., Crawford, J., and Koller, D.**, eds., Making History Interactive: Computer Applications and Quantitative Methods in Archaeology 2009, pp. 48–56. BAR International Series 2079, Oxford.

Charmaz, K. (2006). Constructing Grounded Theory: A Practical Guide through Qualitative Analysis. SAGE Publications Ltd, London. doi:10.1080/17482620600881144.

Cohen, L., Manion, L., Morrison, K., Manion, L. and Morrison, K. (2002). Research Methods in Education. Routledge. doi:10.4324/9780203224342.

Council of Europe (1992). European Convention on the Protection of the Archaeological Heritage (Revised).
URL: *http://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/143*

DANS (2019). DANS EASY.
URL: *https://dans.knaw.nl/en/about/services/easy*

Dejong, M. and Schellens, P. J. (1997). Reader-Focused Text Evaluation: An Overview of Goals and Methods. Journal of Business and Technical Communication, 11(4): 402–432. doi:10.1177/1050651997011004003.

Dudek, D., Mastora, A. and Landoni, M. (2007). Is Google the answer? A study into usability of search engines. Library Review, 56(3): 224–233. doi:10.1108/00242530710736000.

Esmailpour, R., Ebrahimy, S., Fakhrahmad, S. M., Mohammadi, M. and Abbaspour, J. (2019). Developing an effective scheme for translation and expansion of Persian user queries. Digital Scholarship in the Humanities. doi:10.1093/llc/fqz041.
URL: *https://doi.org/10.1093/llc/fqz041*

Evans, T. N. L. (2015). A reassessment of archaeological grey literature: Semantics and paradoxes. Internet Archaeology, 40. doi:10.11141/ia.40.6.

Gerjets, P., Kammerer, Y. and Werner, B. (2011). Measuring spontaneous and instructed evaluation processes during Web search: Integrating concurrent thinking-aloud protocols and eye-tracking data. Learning and Instruction, 21(2): 220–231. doi:10.1016/j.learninstruc.2010.02.005.

Gibbs, F., Gibbs, F. and Owens, T. (2012). Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs. Digital Humanities Quarterly, 006(2).

Gormley, C. and Tong, Z. (2015). Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine. O'Reilly Media, Sebastopol.

Habermehl, D. (2019). Over zaaien en oogsten, de kwaliteit en bruikbaarheid van archeologische rapporten voor synthetiserend onderzoek. Technical report, Rijksdienst voor Cultureel Erfgoed, Amersfoort.
URL: *https://www.cultureelerfgoed.nl/publicaties/publicaties/2019/01/01/over-zaaien-en-oogsten*

Hermjakob, U., Hovy, E. and Lin, C. (2000). Knowledge-based question answering. In Proceedings of the Sixth World Multiconference on Systems, Cybernetics, and Informatics (SCI-2002).

Hessing, W., Waugh, K., van Heeringen, R. and Visser, C. (2013). Evaluatie en optimalisatie waarderingssystematiek Kwaliteitsnorm Nederlandse Archeologie. Fase 1: Evaluatie. Technical report, Vestigia, Amersfoort.

Hinostroza, J. E., Ibieta, A., Labbé, C. and Soto, M. T. (2018). Browsing the internet to solve information problems: A study of students' search actions and behaviours using a 'think aloud' protocol. Education and Information Technologies, 23(5): 1933–1953. doi:10.1007/s10639-018-9698-2.

Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E. and Coates, C. M. (2015). Trading consequences: A case study of combining text mining and visualization to facilitate document exploration. Digital Scholarship in the Humanities, 30: 50–75. doi:10.1093/llc/fqv046.

Hu, X. (2018). Usability Evaluation of E-Dunhuang Cultural Heritage Digital Library. Data and Information Management, 2(2): 57–69. doi:https://doi.org/10.2478/dim-2018-0008.

Huurdeman, H. C. and Piccoli, C. (2020). "More than just a Picture" - The importance of context in search user interfaces for three-dimensional content. In CHIIR 2020 - Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, pp. 338–342. Association for Computing Machinery, Inc, New York, NY, USA. doi:10.1145/3343413.3377994.
URL: *https://dl.acm.org/doi/10.1145/3343413.3377994*

Karoulis, A., Sylaiou, S., Informatica, M. W. and 2006, u. (2006). Usability evaluation of a virtual museum interface. Informatica, 17(3): 363–380.

Kirkpatrick, L. C. (2018). Using Computer Screen Recordings and Think Aloud Protocols to Study Students' Cognitive Strategies While Working Online. SAGE Publications Ltd, London. doi:10.4135/9781526444240.

Koolen, M., van Gorp, J. and van Ossenbruggen, J. (2018). Toward a model for digital tool criticism: Reflection as integrative practice. Digital Scholarship in the Humanities, 34(2): 368–385. doi:10.1093/llc/fqy048.

Lewis, C. (1982). Using the 'thinking-aloud' method in cognitive interface design. Technical report, IBM TJ Watson Research Center, New York.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J. and Gómez-Berbís, J. M. (2013). Named Entity Recognition: Fallacies, challenges and opportunities. Computer Standards and Interfaces, 35(5): 482–489. doi:10.1016/j.csi.2012.09.004.

Niccolucci, F. and Richards, J. D. (2013). ARIADNE: Advanced Research Infrastructures for Archaeological Dataset Networking in Europe. International Journal of Humanities and Arts Computing, 7(1-2): 70–88. doi:10.3366/ijhac.2013.0082.
URL: *https://www.euppublishing.com/doi/10.3366/ijhac.2013.0082*

Nielsen, J. and Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '93, pp. 206–213. ACM Press, New York, New York, USA. doi:10.1145/169059.169166.

Paijmans, H. and Brandsen, A. (2010). Searching in archaeological texts: Problems and solutions using an artificial intelligence approach. PalArch's Journal Of Archaeology Of Egypt/Egyptology, 7(2): 1–6.

Pescarin, S., Pagano, A., Wallergård, M., Hupperetz, W. and Ray, C. (2014). Evaluating virtual museums: Archeovirtual case study. In **Earl, G., Sly, T., Chrysanthi, A., Murrieta-Flores, P., Papadopoulos, C., Romanowska, I. and Wheatley, D.**, eds., Archaeology in the Digital Era - Papers from the 40th Annual Conference of Computer Applications and Quantitative Methods in Archaeology, volume 74, pp. 74–82. Amsterdam University Press, Amsterdam.

RCE (2017). De Erfgoedmonitor.
URL: *https://erfgoedmonitor.nl/indicatoren/archeologisch-onderzoek-aantal-onderzoeksmeldingen*

Rico, M., Vila-Suero, D., Botezan, I. and Gómez-Pérez, A. (2019). Evaluating the impact of semantic technologies on bibliographic systems: A user-centred and comparative approach. Journal of Web Semantics. doi:10.1016/J.WEBSEM.2019.03.001.

Rieh, S. (2006). Analysis of multiple query reformulations on the web: The interactive information retrieval context. Information Processing & Management, 42(3): 751–768. doi:doi.org/10.1016/j.ipm.2005.05.005.

Rijksdienst voor het Cultureel Erfgoed (2019). Archis.
URL: *https://archis.cultureelerfgoed.nl*

Rosenzweig, E. (2015). Successful user experience: Strategies and roadmaps. Elsevier. doi:10.1016/c2013-0-19353-1.

Russell-Rose, T., Chamberlain, J. and Azzopardi, L. (2018). Information retrieval in the workplace: A comparison of professional search practices. Information Processing and Management, 54(6): 1042–1057. doi:10.1016/j.ipm.2018.07.003.

Russell-Rose, T. and Shokraneh, F. (2020). Designing the Structured Search Experience: Rethinking the Query-Builder Paradigm. Weave: Journal of Library User Experience, 3(1). doi:10.3998/weave.12535642.0003.102.
URL: *http://hdl.handle.net/2027/spo.12535642.0003.102*

Selhofer, H. and Geser, G. (2014). D2.1: First Report on Users' Needs. Technical report, ARIADNE.
URL: *http://legacy.ariadne-infrastructure.eu/wp-content/uploads/2019/07/ARIADNE_D2-1_First_report_on_users_needs.pdf*

Sorel, D. (2018). jQuery QueryBuilder.
URL: *https://querybuilder.js.org/*

Spink, A. (2002). A user-centered approach to evaluating human interaction with Web search engines: an exploratory study. Information Processing & Management, 38(3): 401–426. doi:10.1016/S0306-4573(01)00036-X.

Steiner, C. M., Agosti, M., Sweetnam, M. S., Hillemann, E.-C., Orio, N., Ponchia, C., Hampson, C., Munnelly, G., Nussbaumer, A., Albert, D. and Conlan, O. (2014). Evaluating a digital humanities research environment: the CULTURA approach. International Journal on Digital Libraries, 15(1): 53–70. doi:10.1007/s00799-014-0127-x.

Thomsett-Scott, B. C. (2006). Web site usability with remote users: Formal usability studies and focus groups. Journal of Library Administration, 45(3-4): 517–547. doi:10.1300/J111v45n03{_}14.

van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T. and Van de Walle, R. (2015). Exploring entity recognition and disambiguation for cultural heritage collections. Digital Scholarship in the Humanities, 30(2): 262–279. doi:10.1093/llc/fqt067.
URL: *https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqt067*

van Waes, L. (2000). Thinking aloud as a method for testing the usability of Websites: the influence of task variation on the evaluation of hypertext. IEEE Transactions on Professional Communication, 43(3): 279–291. doi:10.1109/47.867944.

van Zundert, J. J. (2016). The case of the bold button: Social shaping of technology and the digital scholarly edition. Digital Scholarship in the Humanities, 31(4): 898–910. doi:10.1093/llc/fqw012.
URL: *https://doi.org/10.1093/llc/fqw012*

Verberne, S., Boves, L. and van den Bosch, A. (2016). Information access in the art history domain: Evaluating a federated search engine for Rembrandt research. Digital Humanities Quarterly, 10(4): 69–87.

Vlachidis, A., Tudhope, D., Wansleeben, M., Azzopardi, J., Green, K., Xia, L. and Wright, H. (2017). D16.4: Final Report on Natural Language Processing. Technical report, ARIADNE.
URL: *http://legacy.ariadne-infrastructure.eu/wp-content/uploads/2019/01/D16.4_Final_Report_on_Natural_Language_Processi*

Voorhees, E. (2001). Overview of TREC 2001. In Proceedings of the Tenth Text REtrieval Conference (TREC 2001), pp. 1–13.