Received 9 March 2023
Revised 24 September 2023
Accepted 30 September 2023

Online subject searching of humanities PhD students at a Swedish university

Koraljka Golub
iInstitute, Linnaeus University, Växjö, Sweden
Xu Tan
School of Information Management, Wuhan University, Wuhan, China
Ying-Hsang Liu
Department of Archival, Library, Information and Museum Studies, Uppsala University, Uppsala, Sweden, and
Jukka Tyrkkö
iInstitute, Linnaeus University, Växjö, Sweden

Abstract

Purpose – This exploratory study aims to help contribute to the understanding of online information search behaviour of PhD students from different humanities fields, with a focus on subject searching.

Design/methodology/approach – The methodology is based on a semi-structured interview within which the participants are asked to conduct both a controlled search task and a free search task. The sample comprises eight PhD students in several humanities disciplines at Linnaeus University, a medium-sized Swedish university from 2020.

Findings – Most humanities PhD students in the study have received training in information searching, but it has been too basic. Most rely on web search engines like Google and Google Scholar for publications' search, and university's discovery system for known-item searching. As these systems do not rely on controlled vocabularies, the participants often struggle with too many retrieved documents that are not relevant. Most only rarely or never use disciplinary bibliographic databases. The controlled search task has shown some benefits of using controlled vocabularies in the disciplinary databases, but incomplete synonym or concept coverage as well as user unfriendly search interface present hindrances.

Originality/value – The paper illuminates an often-forgotten but pervasive challenge of subject searching, especially for humanities researchers. It demonstrates difficulties and shows how most PhD students have

The authors would like to thank the eight study participants as well as one pilot-test person, who remain anonymous. Special thanks are due to Anna Ishchenko for their help with transcripts.

Since acceptance of this article, the following author(s) have updated their affiliation(s): Ying-Hsang Liu is at the Professorship of Predictive Analytics, Chemnitz University of Technology, Germany. Xu Tan is at the *i*Institute, Linnaeus University, Växjö, Sweden.

Funding: Ying-Hsang Liu is in part funded by ERC project, "CApturing Paradata for documenTing data creation and Use for the REsearch of the future" (grant agreement ID: 818210) and Visiting Scholar Program of Chemnitz University of Technology for established research partnerships.

Corrigendum: It has come to the attention of the publisher of the *Journal of Documentation* that the following article by Golub, K., Tan, X., Liu, Y.-H. and Tyrkkö, J. (2023), "Online subject searching of humanities PhD students at a Swedish university", *Journal of Documentation*, Vol. 79 No. 7, pp. 308-329. https://doi.org/10.1108/JD-03-2023-0044, incorrectly listed Jukka Tyrkkö's affiliation, *i*Institute, Linnaeus University, Växjö, Sweden as Uppsala University, Gävle, Sweden. This oversight has now been corrected.

The authors sincerely apologise to the readers for any inconvenience caused.

Journal of Documentation
Vol. 79 No. 7, 2023
pp. 308-329
Emerald Publishing Limited
0022-0418
DOI 10.1108/JD-03-2023-0044

missed finding an important resource in their research. It calls for the need to reconsider training in information searching and the need to make use of controlled vocabularies implemented in various search systems with usable search and browse user interfaces.

Keywords Controlled vocabularies, Information searching, Knowledge organization systems, Search interfaces, Subject searching, Humanities PhD students

Paper type Article

Introduction

While searching for information online is omnipresent among today's PhD students as they conduct literature searches, many of the search services like discovery systems and commercial search engines are based on fully automated free-text information retrieval systems, rather than on controlled vocabularies. This often results in an overwhelming number of retrieved documents, many of which are irrelevant; especially when only general search terms are used. And vice versa, very specific search terms may retrieve only a few documents and miss out many potentially relevant documents. As things stand today, it is becoming more and more difficult even for trained searchers to conduct searches with satisfactory results in large databases such as discovery systems (see, e.g. Dempsey, 2012; East, 2007; Golub, 2018; Liu *et al.*, 2023; Markey, 2007; Tibbo, 1994).

Although the information retrieval practices of PhD students are not under-researched in relevant subfields of information studies such as information behaviour, most studies have focused on general information channels (see, e.g. Catalano, 2013; Spezi, 2016). At the same time, information retrieval research is rarely targeting subject searching in the humanities, as seen from the lack of test collections for testing the system performance in information retrieval experiments in the latter (see, e.g. Ferro and Peters, 2019; Voorhees and Harman, 2005).

This exploratory study aims to help address the aforementioned gap by investigating online information search behaviour of PhD students from different humanities fields, with a focus on subject searching. The key research questions are: What are typical sources that PhD students use when looking for information? How do they approach subject searching? How good are the preferred search systems in supporting humanities PhD students' needs of subject searching? And how do they approach metadata creation when depositing to a local repository?

The methodology is based on a semi-structured interview within which the participants are asked to conduct both a controlled search task and a free search task. The sample comprises eight PhD students in several humanities disciplines at Linnaeus University, a medium-sized Swedish university from 2020.

The paper is structured as follows: the Background section provides the rationale for this study as well as relevant context; Methodology presents the research approach, method of data collection and analysis; and the Results section presents and discusses the results. The results are summarized and reflected upon in terms of further research in the Conclusion.

Background

Information behaviour of humanities researchers

When it comes to the field of information behaviour, compared to science and engineering, research on the humanities started relatively late (Bouazza, 1989, p. 159; Bates, 1986). Chu (1999) describes the process of humanities research as comprising six stages: idea generation, preparation, elaboration, analysis and writing, dissemination, and further writing and dissemination; out of these stages, the preparation stage is where searching, reading and annotating materials occurs. Traditionally, humanities scholars have tended to work alone, which actually puts higher information seeking demands on them (Case and Given, 2016, p. 37). Literature shows that humanities scholars use a particularly wide range of information sources, including books, newspapers, archives, journals, social media and informal communication (Ge, 2010; Kumar, 2013; Shboul *et al.*, 2019).

Lack of information literacy affects the seeking and searching behaviour. Research has shown that humanities researchers have not been trained to understand differences between search tools (limitations/advantages), understand what advanced search refers to, or what Boolean operators or controlled vocabularies are (Encheva *et al.*, 2019). The difficulty of formulating search queries is linked, among others, to the lack of commonly accepted terminology: different sub-disciplines may describe the same phenomena using different terms (Mierzecka, 2015).

Compared to senior humanities researchers, one might assume that PhD students would on average have better information skills and literacy considering their experience of growing up in the world of the Internet, including web search engines like Google, library discovery systems and social media. However, only a few studies seem to have focused on information behaviour of humanities PhD students. Madden (2014) saw that the information-seeking needs of humanities PhD students are particularly varied, that the first year of doctoral studies includes defining and scoping the research topic and knowing which library resources are relevant to the research, but that the research topic is likely to undergo changes and refinement in the first months of doctoral studies, thus making the information-seeking process even more challenging. Likewise, Wu and Chen (2014) studied PhD students from the disciplines of humanities, social sciences, and science and technology. They found that while the students used the Internet, particularly Google Scholar, and library resources, humanities students showed less dependency on Google than their peers in science and technology.

Subject searching for humanities

Research shows that subject access in online library catalogues, repositories and commercial services like bibliographic databases and discovery services has been less than optimal and often fails to meet established objectives for bibliographic systems (see, e.g. Markey, 2007; Golub, 2018; Golub *et al.*, 2020). While the services try to match users' expectations by implementing Google-like single search box interfaces, it seems that efficient mechanisms such as ranking algorithms used by commercial search engines like Google, efficient exploitation of subject indexing or even quality-controlled subject indexing *per se*, are still missing from these services, which are some of the key reasons behind frequent retrieval failures. Furthermore, the design of search interfaces needs to better support the information seeking processes by further consideration of user behavioral models beyond the specified search (see, e.g. Huvila *et al.*, 2022; Liu *et al.*, 2021).

The specific challenges of indexing humanities for successful retrieval have been a cause for concern for several decades (Langridge, 1976; Tibbo, 1994). As part of the general development of digital scholarship, many disciplines and research areas within the humanities have developed new structures both within themselves and in relation to other disciplines within the humanities and beyond (Borgman, 2007, pp. 212–224). For example, in the rapidly growing interdisciplinary field of digital humanities, it has become increasingly important to provide quality subject access to the vast variety of heterogeneous information objects catalogued by digital services. This includes both primary (see, e.g. Choi and Syn (2015), on the use of tags in archival collections) and secondary sources; and if we wish to go beyond bibliographic databases, the management of textual resources for text analysis, lexicographical and linguistic research (Costa *et al.*, 2021; Musgrave and Haugh, 2019; Walsh *et al.*, 2021).

A major study of online database searching by humanities scholars was conducted over a two-year period by Bates and colleagues at Getty in the 1990s, combining search-log analysis and interviews (final report is given in Bates, 1986). The study showed that most searches were subject searches – 91% of natural language statements indicated a subject of some kind, strongly supporting the need for bibliographic databases to support subject searching. Further, a comparative analysis of the Getty humanities scholars' queries against those of natural science scholars revealed substantial differences in the type of subjects the scholars

searched for (ibid.). Whilst scientific queries typically comprise common terms and only rarely other types of items, the Getty study showed that only 57% of the humanities queries contained any common terms at all (common terms are uncapitalized terms which are none of the following categories: works or publications as subject, individuals as subject, geographical term, chronological term, disciplinary term, other proper term (capitalized) (Bates, 1986, p. 5250); and instead terms that were often present were terms denoting named individuals, geographical terms, chronological terms and disciplinary terms denoting academic disciplines, such as history.

Similar findings have also been reported by Wiberley (1983, 1988), who observed that humanities subject terms are often highly precise proper names, and by Tibbo (1994), who summarized earlier studies observing that single proper terms like authors' names and the titles of works are usual in certain disciplines such as literary studies, while common terms are more characteristic of field such as religion and philosophy. Likewise, Yi *et al.* (2006) found that most search terms of two history databases referred to specific instances of historical events, people and regions, in contrast to the search terms used in a psychology database which were mostly common terms matching those of the dedicated classification scheme.

All these findings suggest that there is a need for a faceted approach to controlled vocabularies, such as the Arts and Architecture Thesaurus for visual arts. Faceted vocabularies are more suitable because they support high specificity and can account for the different facets that are important to humanities scholars, such as geographical, chronological and disciplinary terms (see Bates, 1986; Tibbo, 1994). Furthermore, facet selection and the query expansion based on such controlled vocabularies also needs to be implemented into the search interfaces, which is a feature currently limited to experimental interfaces (see, e.g. Alani *et al.*, 2000; Tudhope *et al.*, 2006; Liu *et al.*, 2022) rather than something applied into practice across bibliographic systems in the humanities (for an example in the medical domain, see EBSCO's Advanced Searching with CINAH database).

It is also worth noting that humanities scholarship is linguistically very diverse: in addition to the considerably high volume of distinct concepts, terminological synonymy and overlaps are also abundant, even within relatively well-defined subdisciplines. The terminological complexity complicates queries by placing the burden on the scholar, who would ideally need to include all possible (near) synonyms in a query if a comprehensive set of results is desired. Homonymy is likewise also present, with the result that queries often end up producing false positives.

Furthermore, terminological changes may also be a problem given that humanities scholars frequently work with materials and sources produced over long periods of time. Consequently, the index terms used may have different meanings when applied to the same term from different periods, or different terms in different periods may be used to refer to the same concept (e.g. changes in countries' common names, such as Ceylon vs Sri Lanka or Persia vs Iran). When it comes to searching by titles, they may be metaphorical, contain allusions or intertextual references, or otherwise be less than descriptive, resulting in low recall (the inability of a retrieval system to present all relevant documents) as well as false positives (the presentation of non-relevant documents). Tibbo further notes that humanities scholars tend to use a "dense" rather than "readable" writing style, making it particularly challenging to create representative metadata (p. 609), the quality pending on information experts also being subject experts.

Similarly complex problems are also commonly encountered when it comes to primary sources. For example, in literary studies sexuality has not been investigated employing computational methods on large text corpora because the phenomenon studied is not manifestly present in the texts and calls for human subject indexing (Bergenmar and Golub, 2020). The lack of conceptualization of terms extracted from text is common in humanities; concepts such as "dramatists", or "persecution" may not be articulated for a paper about particular playwrights or conflicts.

Methodology

Purpose and aims

In order to help improve search services and contribute to the body of research on online information search behaviour of humanities PhD students, the study aims to identify: (1) typical resources that the PhD students use when looking for information; (2) ways in which they approach subject searching, including any use of controlled vocabularies; (3) how good the selected search systems are in supporting humanities PhD students' needs when searching by subject; as well as (4) how do the PhD students create metadata when depositing their work in the university repository.

Sample

Although different researchers from different humanities disciplines present different search and information behavior (see, e.g. Bates, 1986), this study included all humanities disciplines available, in order to capture their searching behaviour, especially with respect to subject searching and use of controlled vocabularies.

Invitation to participate in the study was sent out to all PhD students at the Faculty of Arts Humanities, Linnaeus University in Sweden on 21 April 2020, with a reminder sent several weeks later. Of the 29 PhD students, 8 confirmed their participation, 5 declined and the remainder did not provide any feedback. Due to Covid-19 restrictions, the 8 interviews took place online via Zoom, between 7 May and 19 June that year.

The invitation informed the students that participation was voluntary and no compensation would be offered. It described that the interview session, lasting no longer than 90 min, would entail questions on how they normally search for information for research purposes. Also, the students were informed that they would be asked to perform two search tasks online, one on a topic that they have searched for recently and the other on a topic which would be provided. Further, the invitation stated that the interview would be held via Zoom in English and that it would be recorded. The recordings would not be made available to anybody apart from the researchers. The results would be anonymized in the paper and the only information about their identity given would be the fact that they were PhD students in arts and humanities at Linnaeus University. Each participant signed a consent sheet.

Method

The method is based on a semi-structured interview within which the participants are asked to describe their information behaviour, recall their most recent information search and conduct both a controlled search task and a free search task. The data collection techniques include the interviewing data, critical incident technique (Flanagan, 1954) and observations of search behaviour. This allows us to capture the participants' information behaviour and characterize search behaviour regarding subject searching at the same time.

The interview guide comprised two blocks of questions and two blocks of tasks, of which one task was a search task they had recently conducted for their research and the other a controlled one. The first search task was designed to reflect the real information needs, whereas the controlled one can be characterized as an exploratory search task on the topic of digital methods for the humanities, with an emphasis on the activities of learning and investigation (Marchionini, 2006; see Li and Belkin, 2008 for further classification of search tasks).

The four blocks of questions and tasks were as follows:

(1) *Background.* Demographic information such as age and gender were collected along with information related to their research such as their thesis topic, discipline and the stage of their PhD project. They were also asked if they used any digital research

methods or techniques in their research because that was the topic of the controlled search in the fourth block. Then there was a question related to any received training in information searching as well as in metadata creation for depositing their work in the local university repository, DiVA. Related to the latter, we also inquired how they chose the national subject category, a required element during the deposit, as well as how they choose keywords (if any). Also, they were asked about author keywords when publishing their work and whether there were any journal guidelines on how to create them. Finally, this block ended with a general inquiry into whether they have ever missed finding an important resource in their research, only to have found out about it too late (e.g. during peer-review) and what they thought the reason for not finding such resources was.

(2) *Information behavior*. In this block the participants were asked about their frequently used information places when they look for secondary sources and also when looking for primary sources such as data sets, museum objects, etc. They were then asked to name databases specifically and to provide their impression of them as to how good they are in supporting searching. The final questions were about the role of keywords and controlled vocabularies in those search systems.

(3) *Recent search*. Here the participants were asked about their most recent research question and related online search: the database(s), role of keywords. Then they showed how they went about their most recent topic search related to that research question. Wherever relevant, they were asked to do the same using a controlled vocabulary.

(4) *Controlled search task*. In this final task, the interviewees were to conduct a search on a specific, imaginary topic: "Try to find research publications which will help you write an imaginary overview paper for the best journal in your discipline on (dis) advantages of digital research methods or techniques in your main discipline." Then they were asked to conduct the search in each of the four databases:

- Their usual main database.
- Their main disciplinary database.
- Scopus.
- DiVA (the university repository).

In each of the systems they were asked to choose a simple and advanced interface, to inspect search options and to consider which keywords they would use. They were also asked to choose one of the relevant retrieved documents and determine if listed keywords were relevant.

Data analysis comprised the transcription and content analysis by one of the authors in Nvivo.

Results

Demographics

All of the participants ($n = 8$) are PhD students, two in their twenties, four in their thirties, one in their forties and one in their fifties. Four people identified as male, three as female and one preferred not to say. Three are conducting their PhD in linguistics, two in comparative literature, one in archaeology, one in history and one in learning Swedish as a second language. One participant is at the start stage of their PhD studies (still taking the courses), four are at the medium stage of their PhD studies (finished data collection) and three are at the end stage (analysing data, doing empirical experiments and halfway of writing thesis.

Most of them ($n = 6$) use digital research methods or techniques for data collection, data analysis or data presentation, with one planning to do so at a later stage of their PhD studies. The one person who does not rely on digital methods or techniques instead conducts traditional literary and film analysis without using any tools for analysis. Of the participants who used digital research methods or techniques, they did so to collect data via a questionnaire, interview or a field study as well as for data analysis and presentation.

Six of the participants received training in information searching during their university studies, while the remaining two relied on their supervisor for information. Only two had received training at all the three higher education levels (bachelor, master's and doctoral). The training was provided by university libraries in the use of different search systems, including discovery systems like OneSearch used by the Linnaeus University Library, DiVA repository used by Linnaeus University, Google Scholar, Web of Science and Swedish Union Catalog LIBRIS.

Three participants considered the training to be introductory and basic, without much benefit:

(1) "We got a very general intro like, well, if you want something where the piece is available, then you have to click "full text" or 'available in the library', so just the main functions, not really the databases or anything like that".

(2) "I think the training was not very useful for me, I think, mostly because it was very general. And the general is more targeted towards science".

(3) ". . . not a proper training, but you get these short courses when you begin studying at university from the library personnel, introductions like that".

Thus, the training has not seemed to have targeted any specific disciplinary databases, and this seems to be because the courses at the doctoral level tend to be given collectively to students from varied disciplines: ". . . we were several different students from different disciplines, so they would not really show us that specific information".

Similarly, none had received any training related to depositing their research output in the DiVA repository, which could have included training on metadata creation, including assigning subject keywords. Five had already deposited their research outputs; four relying on written guidelines available on the library website while one asked a librarian for help on DiVA during a training session on another topic.

During a DiVA deposit, the author must choose an appropriate national subject category. The national categories were considered appropriate by two participants, while three thought that they are too broad and that it is not possible to be more specific and interdisciplinary. Indeed, DiVA national categories are broad disciplinary-level, controlled index terms from the Swedish standardized categorization of research areas (Statistics Sweden, 2016). There are three levels of hierarchical division: 6 top level categories, which are divided into a total of 42 categories at the second hierarchical level, followed by about 250 categories at the deepest level. The first two levels are the same as the international OECD's classification of research topics (Organisation for Economic Co-operation and Development, 2007) while the third one is specific to Sweden. The main purpose of the categories is to facilitate the collecting of data for official national statistics on scientific publishing. However, it is also implemented in the DiVA repository and does not really allow hierarchical subject browsing or searching at the most specific level, the latter being a common bibliographic objective.

All the five participants with experience with DiVA depositing also included free keywords; although they are optional, the PhD students seem to appreciate the value that keywords may bring to supporting information discovery. One used to work with search engine optimization before starting their PhD studies and consequently understands their

value. Another student commented that subjects are important for literary works: "We need to put in the keywords for the author and the period, and all of that sort of stuff, because otherwise no one's going to find the paper . . . we also try to put as much of the method in the keywords as well, because since we couldn't fit it into a category, it seemed that was sort of our chance to actually have someone find the paper if they were looking for it". Still, all agreed that there were no guidelines available on how the authors should choose keywords for DiVA.

The participants were also asked about how they create author keywords when publishing their work in journals and of any related journal guidelines. They reported that there were no guidelines on keywords creation in journals where they publish. Their strategies seem to rely on identifying key concepts ("I just improvised, I think. I tried to think of some, you know, what seemed most central to the topic of the article"; "No, I've never seen any guidelines or anything like that for it. So I take whatever the concepts I feel like are the most important in my text"; predicting the readers ("Perhaps I was also thinking about what kind of readers where I want to read the article"); and, relying on their own subject expertise ("I would put keywords . . . for my own classification of my own work because I should know best").

The lack of appropriate training in information searching (e.g. in disciplinary databases using controlled vocabularies) as well as in metadata creation for depositing their work in the local university repository, DiVA, may have direct effect on ways of searching that follows. This calls for strategic change in conducting training in information searching at the university library. Further, the lack of guidelines in keyword creation either in the repository or academic journals is significant when it comes to the lack of quality control in the final metadata which provide the basis for searching in many information databases that humanities researchers rely on.

Finally, the first block of questions ended with a general inquiry into whether they have ever missed finding an important resource in their research, only to have found out about it too late (e.g. during peer-review) and what they thought the reason for not finding such resources was. This has happened to five participants for the following reasons:

(1) A mismatch between the keywords of the article and the search terms used by the user. For example, a participant missed several publications, all by the same author, at the time of writing their Master's thesis on Hemingway's stylistics. They only found out two years later at a conference mingle where they met an author with three or four relevant publications on that topic. The reason for this was that they used certain terms to name the relevant methods as search words, but the author used different ones, and the search systems did not provide a controlled vocabulary to align those synonyms.

(2) The relatively large number of synonyms in the humanities and literary studies. A participant said that ". . . a lot of the time people are talking about the same thing, but they are not using the same words . . .". Although now that the participant has become aware of it, they try to guess the different terms for the same concept, but this still does not result in including all relevant terms. This also points to the need for controlled vocabularies with lots of synonyms, which are also being frequently updated.

(3) Too many retrieved documents with irrelevant results. Because one needs to prioritize other tasks, the participant often lacks the time to browse to the 10th or 20th results page where they used to find something which is "really relevant to my work". And she is also aware that indexing and ranking is not done well (she says "not very democratically done") which leads to excluding some resources that are very relevant to the search term.

(4) Search habits. A participant reports on missing resources because of forgetting to search multiple databases as well as forgetting to search in different Scandinavian languages. "Like I'm not too good at Swedish film music scholars, even though I write about film music, because I normally use Danish or English sources, even though I should be able to read Swedish."

Information search

Where do users usually look for information to support their research?

For secondary information resources like publications, most resort to web search engines, discovery systems and bibliographic databases. Five participants use the general web search engine Google and four used the academic search engine Google Scholar. Five use OneSearch, the discovery system provided by the Linnaeus University Library. Three participants use the multidisciplinary bibliographic databases Web of Science ($n = 2$) and JSTOR ($n = 1$) WOS ($n = 2$) and JSTOR ($n = 1$); two participants use disciplinary databases (RILM Abstracts of Music Literature with Full Text and ERIC (Educational Resources Information Centre), one each). In addition to these, the participants also use informal sources, such as supervisors ($n = 2$), social media ($n = 1$) and email ($n = 1$). Of primary resources, they use the English Corpora within which the BNC (British National Corpus) ($n = 2$) and Corpus of Contemporary American English (COCA) ($n = 1$), IRIS (database of instruments, materials, stimuli, data, data coding and analysis tools used for research into language) ($n = 1$); as well as Swedish National Data Service (SND) ($n = 1$) and Fornsök (Swedish national registry of ancient sites and other cultural heritage sites) ($n = 1$).

Reasons mentioned for choosing Google are that it is fast, has broad coverage and provides direct links to websites like Wikipedia. One person also describes how they first use Google to get general information resources which then help her formulate a search query in Google Scholar. Similarly, Google Scholar is often used because it is easy to use, fast, has broad coverage and additionally shows the citation rate and metadata of the publication. Disadvantages of Google Scholar are many irrelevant hits, and it takes time to go through pages of results: "In Google Scholar, it takes a while because you have to go through all the pages, and it's not very good at figuring out exactly what I want. And sometimes something very relevant will be hidden on page four. But to get there I'll have to read the abstracts for like a dozen or more publications before I eventually find it". Or "I think it's an inherent problem with Google Scholar . . . because it sorts it according to citations. So, if it's not a very popular topic, it's going to hide the result away, even if it's very relevant and it matches very well, but not enough people have cited it."

The University Library's discovery system OneSearch was typically chosen for known item searching because it links directly to full-text documents or the library building on campus (an estimate of 70% given by a respondent), is especially good for monographs as well as for older information resources. Discovery systems are often criticized for an overwhelming number of results and a black-box approach to searching across a multitude of databases. While one participant criticized it for those reasons (". . . there are too many filters you could apply, and then of course you don't, because you want to get a list rather quickly . . . then you have lots of noise. So lots of hits that are not really suitable for what you're looking for"), two participants considered different filtering options useful to help address that problem (". . . in OneSearch, they're almost always relevant, because I have so many sorting filters to use").

A participant commented that they now only sometimes use the RILM (Répertoire International de Littérature Musicale) database because they have discovered that OneSearch covers it. Although this means they must deal with a lot of irrelevant results, they prefer OneSearch because it also finds resources not covered by the specialized database. Another

participant mentioned the internet Archaeology journal as a good resource because it also allows inclusion of datasets.

One participant considers the English Corpora database, an online corpus platform, of great value because there are many different corpora there, although they noticed the problems with the lack of browsing: "you could not really access the texts that are stored there without searching for something". They also mentioned there are many automatically tagged mistakes in the corpora, such as words tagged as nouns that should be tagged as verbs and vice versa. Another respondent considered the interface user friendly and appreciated their online tutorials.

Social media are reported to be used as both a primary and secondary resource as well as for interactions with the community. Specifically mentioned was a Reddit community called "Data is Beautiful" which provides curated visualizations over datasets. The participant commented that they are not very good at data management and appreciated the community for telling him what one can do with the data sets. Another respondent researching migration, literature and cinema uses social media to look for reviews, for materials from the production process, or people's feedback on works, like in Goodreads.

Colleagues like supervisors, other subject experts (e.g. at conferences), fellow PhD students working in the same field and librarians are also an important source of information. One relied on the supervisor especially in the early stage of the PhD, while another asked for help when they were not able to find anything or were looking for advice on methodology. Another participant in the early stages asks a librarian for help to increase one's confidence that they are searching in the right direction.

Since this study specifically targeted the role of subject keywords, here we also asked about general use of keywords for searching: whether they use keywords in the databases, whether they are aware of any controlled terms and if yes, whether they use them. Considering previous research pointing to problems of controlled terms not being available at search interfaces and based on the types of commonly used databases by the study participants, it is not surprising that none of the participants know about controlled vocabularies, their advantages for disambiguation, achieving potentially better search results that are evaluated by precision and recall measures. They report on the need to use very many search terms, the need to save all the search keywords in order to remember them next time and not miss anything, the need to use specific terms so as to avoid too many results (and if too few results, then they broaden the search with broader terms), the problems of discovery systems and their relevance ranking as well as the "black-box' algorithm problem (". . . something happened to the algorithm a few years back where I normally only got relevant results; and then all of a sudden you know you would start having swingy articles about some medical subject before all the articles you wanted . . ."). One person does not rely so much on keywords, but on the author, as keywords can be "a bit trickier".

Recent search

Six participants' latest search was related to their PhD thesis topic, while two participants were looking for information they needed to complete the publications they were working on at the time. Four participants used Google Scholar, three resorted to OneSearch, two to Google. An archaeology student used one journal official website (Internet Archaeology) and two specialized databases: Swedish National Data Service (SND), and Fornsök.

Specifically, when they were looking for information on remix studies in relation to social media to gain general knowledge on what this theme could entail, Google Scholar was selected for information searching rather than OneSearch because they felt that the latter would not have resources on remix studies, and they felt that Google Scholar is a good place to start because it "catches very broadly" (P3). When looking for information about the history

of the noun "police", they resorted to OneSearch and used the search term "history police" that retrieved 430,722 results. They further narrowed down the search to books available in the physical library on campus, resulting in 53 hits. The next step would be to go to the library and inspect the books for their relevance (P5). When the search topic was about how digitalization is affecting our knowledge production, they used four databases in the following order: Google, Internet Archaeology, Swedish National Data Service (SDN) and Fornsök for various resources, including research publications and datasets (P7). When their recent search was about decolonial theory, they tried to see if it would be a good fit for their thesis. Mostly recently they were looking for decolonial journals. They used Google and entered the term "decolonial journal", resulting in 412,000 hits. They usually start with Google and when finding the right journal, they go on to the University Library to find specific articles (P8). Overall, these results reflect the task-based information searching, which recognizes a user's task as important factor in information search activities and the importance of database selection during information searching processes (Kim, 2007; Vakkari, 2003).

We found the problem of vocabulary mismatch between the user's query terms and the keywords represented in documents in our interview data (Furnas *et al.*, 1987; Svenonius, 2000). For example, when they were looking for information on teaching and learning in multilingual classrooms in Sweden. They considered their preferred system Google Scholar difficult in this example. They had to try very many combinations of search terms, including abbreviations and special words. Their hard work did not pay off since they could not find relevant results on the specific topic of the problems of teaching in homogeneous versus heterogeneous pupils' backgrounds, background referring here to speaking the same language. They were only able to find research on the benefits of multilingual classrooms (P1). When looking for information about the Swedish upper-secondary school system, for publications that describe the system from an international perspective, they demonstrated their search by going to Google Scholar and using the phrase search with quotation marks "Swedish upper-secondary", explaining how the phrase would narrow down the results, although being aware that this could exclude documents not using the specific phrase. The query retrieved 44,300 documents which they then limited to 2012–2020, to exclude publications before the Swedish school reform in 2011. This reduced the result set to 17,200 documents which they then further narrowed down by adding the term "English" to the original search query, resulting in 1,920 documents (P2). In this case the information regarding international perspectives is not well-represented in the search results, which could be enhanced by the use of controlled vocabularies.

However, the perceived usefulness of controlled vocabularies has revealed the well-recognized problem of the exhaustivity and specificity of indexing languages for retrieval purposes (Sparck Jones and van Rijsbergen, 1976; Svenonius, 1986). For instance, the participant went to OneSearch which they also often use to obtain access to full text resources. For the Swedish search term "gymnasieskolan språk" (secondary school language) they got 103 results and then applied a refining function to only retrieve books available in the physical library, retrieving 69 results. They do not normally use the subject filtering option because they would lose texts indexed too narrowly, e.g. a work indexed with "English linguistics" may not be indexed by "education" even if it is about it and relevant for me (P2). This points to the problem of indexing exhaustivity in databases. Additionally, when the most recent search was about linguistic ethnography and related methodology, they resorted to OneSearch, entering the name of a person, "Asif Agha" and directly specified that the term should be found in the author/creator metadata element. This resulted in 52 records which they then sorted by newest publication date. When they opened the metadata record for one relevant publication, the interviewer asked if they would consider using subject keywords from the metadata record in search, and they considered them too broad and said they would

"drown in them"; these included: "kinship", "social history", "social order", "culture and social structure", "social anthropology". This points to the problem of indexing specificity not being addressed to the level the researcher may desire (P6). When they showed how they would interact with SND (http://snd.gu.se) and used the Swedish search term "arkeologi" (archaeology) to retrieve data sets in the discipline, retrieving 540 results, they commented how the metadata would not allow searching by more specific topics or by paradata, i.e. to allow him to find out how data has been collected and why and what hasn't been collected, which would help their search on knowledge production. They also showed how the subject browsing tree for Archaeology at SND database is broad, i.e. there are no further subdivisions such as for periods (e.g. Bronze Age, Iron Age) or types of settlements (P7). Again, humanities researchers would need more specificity from controlled vocabularies.

As our observations of user search behaviour from recent search are similar to controlled search, research issues of search tactics and procedures and search behaviours specific to subject searching are detailed in the following section.

Controlled search

The controlled search task was designed to explore the user search behaviour. It was not a full-scale controlled user experiment study for comparing, for instance, the effect of novel search interfaces, which is widely adopted in interactive information retrieval research (see, e.g. Liu *et al.*, 2022; Liu *et al.*, 2021). In this task, the participants first chose their usual main database: Google Scholar, Google and OneSearch. Three participants started by using broad search terms; this, they commented, in order to learn about the theme and so as not to miss any important resources. A respondent reflected: "I want sort of the broad paradigms before I move it down to subject levels, because I find that if I start sort of from where I am at, I normally miss a bigger overarching concept. And if it's going to be an overview paper, then at least I think it needs to funnel down from the very broad perspective into exactly what it's going to be dealing with". Or: "I try as much as I can to move out of my own discipline, because I feel that it would be limiting if I stay there . . . I'll just make a very broad search 'digital research method'". Our results generally are consistent with the previous finding that professional searchers' database selection was affected by task or context related factors (Kim, 2007). Overall, our participants preferred to select familiar databases that have a broad coverage of research topics to engage with exploratory search activities.

After getting the search results, three participants browsed more than page one. One of them said that they normally go to the 10th page and also would check results under "Related searches" where they sometimes find inspiring search terms.

The participants prefer to use Google Scholar because of citation counts. High citation counts make one participant go to Academia to check the author's other works. From relevant works they also look at other referenced works on Google Scholar – which articles have cited this one? Also, it is important to look for criticism, if any, in the citing works and to see if there is an academic discussion taking place on a certain work. They also like Google as well as Google Scholar because of the keywords being highlighted in the results. In line with search engine studies, our participants pay attention to top ranked search results and the keyword highlighting feature that helps them navigate the search results, and domain expertise affects the time spent on displayed pages in a search session (Savenkov *et al.*, 2011; White *et al.*, 2009). And the humanities scholars' use of Google Scholar reflects their preferred approach of following the bibliographic references from documents and citations for identifying research topics (Green, 2000; Tibbo, 1994).

In Google or Google Scholar no participant chose advanced search interface; of the three who went to OneSearch, one chose both simple and advanced interface, commenting it provides more options to a more successful search outcome. One participant went to several

databases as they tend to search several of them (OneSearch, Google). Only one of all participants attempted to use Boolean operators, first in the simple interface of OneSearch, which they admittedly could not remember to use well, and then went to its advanced search interface and used the Boolean logic operator "Not". The black-box approach to searching in OneSearch was commented on by one participant who said they did now know whether all her search words were included in the search or how to influence that. These findings revealed that search expertise makes a difference in user search behaviour and advanced search functions, such as Boolean operators have been rarely used (Liu and Wacholder, 2017; White and Morris, 2007).

Our characterization of search behaviours demonstrated that the participants had problems formulating query terms to represent the information topics of the assigned search task. All the searches proved difficult in their typical searches within selected databases. The search task demanded much more time than predicted in the interview because the topic was new to most of them, especially its interdisciplinary perspective, and writing an overview article would require collecting a good number of relevant publications. But, none have found even a good starting publication on that broad theme of digital research methods. This could be because the name may not be necessarily used as such in the relevant documents; rather, many other specific methods have their own names. The suggested terms from the search function of related searches have been consulted but their search effectiveness has mixed results, similar to previous findings of the significant role of domain expertise in query formulations and reformulations for search success (see, e.g. Liu *et al.*, 2022; Tang *et al.*, 2013). A good search interface with information retrieval thesaurus would support searchers by guiding the searcher to which terms to include in a search, controlling synonyms and disambiguating homonyms (Bates, 1986; Shiri and Revie, 2005; Svenonius, 2000). Please see Appendix 1 for a list of steps conducted in each of their commonly used databases.

As only two participants have ever used disciplinary databases (RILM and ERIC), they were given a suggestion of one by the interviewer. Linguistic databases included LLBA (Linguistics and Language Behaviour Abstracts by ProQuest) and MLA (Modern Language Association by EBSCO), both with an information retrieval thesaurus. In this task the interview focus was on using a thesaurus. The thesaurus is something that neither of the participants have been trained in but they considered it worth exploring. However, the search interface was suboptimal and the thesauri did not have the most relevant terms, although some were helpful to the participants. One respondent commented that they would be happy to use LLBA, but would like it to be simpler. These results are consistent with the finding that domain experts' perceived usefulness of thesaurus terms and the topic familiarity are correlated with search success in controlled user experiment studies (Liu *et al.*, 2022; Tang *et al.*, 2013; Wittek *et al.*, 2016).

The two databases in film and literature, RILM and Film and Television Literature Index, provide subject terms and index terms separately. Although one of the participants had used RILM, they hadn't searched the subject terms. After some attempts, both participants thought that the results were not useful, because they could not find any expected terms and could not help further searches.

The database in education, ERIC, provided thesaurus. It didn't stand out, probably because the participant was not trained and expected to retrieve a tool. The names of emerging tools were hard to get into the thesaurus. In addition, when the participant wants to get the content of the next level of the search term, that is, to get a deeper semantic level of terms, they often only get the related semantic level of terms.

The database from history, Historical Abstracts (from EBSCO), provides index terms. One of the participants made further searches better by using index terms, while the other did not. And they both thought that the interface was difficult to use and not user friendly. For a complete list of search steps using disciplinary databases, please see Appendix 2.

The participants were then asked to try the search in Scopus; most have never used it before ($n = 6$) and one has used it before and would like to use it more. All have used only the simple search mode; one tried the advanced one but considered it too complex. Two have used Boolean AND in the search query. They appreciated the good search interface and features such as highlighting search words in the results, sorting by highest citation and filtering by subject (one participant often searches for adaptation which is a common term in biology and other disciplines and finds it useful to be able to restrict to literature). However, Scopus wasn't very successful either; three participants found the system still offered the irrelevant subject results after using the subject area filter function, because "humanities" also retrieved results in other sciences like medicine, computer science and social sciences.

The final database was DiVA. All participants have used DiVA for searching, mostly to learn more about known authors' works or for student theses as part of their teaching engagements. Most ($n = 6$) chose the simple search mode only, and three chose advanced search. Here the focus of the interview was on subject terms in DiVA, especially the controlled subject terms National Category (described above) which allows hierarchical subject browsing. Only one has used it earlier but considered it too broad because "it just gives you everything within a discipline"; indeed, this is the case and also found unsuitable for end user searching and browsing. Also confusing are categories called "Other" as it is not clear what that would cover. Two considered it useful to get an overview of what kinds of topics are being researched within a discipline or to find out which keywords from identified publications within that discipline to use in searching. One also said how it would be nice to have a category titled "digital research methods", indicating the usefulness and need for browsing on interfaces.

Overall, our findings confirm that the design of thesaurus-based retrieval system to support query formulation/reformulations needs to further consider enhancing the usability of search interfaces and facilitating the perceived usefulness of suggested terms and tools (Liu *et al.*, 2022; Shiri and Revie, 2005). Since the advanced search option has been designed for search experts, they are perceived as too difficult to use at the interface level, or the participants do not know how to use the subject categories effectively. User characteristics of domain expertise and search expertise and their interactions with the features of search interfaces contribute to better search or learning outcomes (Liu and Wacholder, 2017; O'Brien *et al.*, 2022; Wu and Vakkari, 2018). Nonetheless, the usefulness of thesaurus-based information retrieval systems can be enhanced by designing more usable search interfaces, continuous updates of thesaurus terms and drawing on the best practices of domain and search experts when designing search tools supporting query formulation/reformulation tasks.

Conclusion

This exploratory study of information searching by PhD students in the humanities aimed to determine their typical information channels, ways in which they approach subject searching, the degree to which selected search systems support them in subject searching, as well as their own subject metadata creation. The study was based on a semi-structured interview involving a free and controlled search task. The eight PhD students represented several different age groups, genders and PhD research stages. They conducted their studies in linguistics, comparative literature, archaeology, history and in learning Swedish as a second language.

Six of them have received training in information searching but the training seems to have been too introductory and basic, rarely covering disciplinary databases or advanced search interfaces and has never included any training on controlled vocabularies. None have received any training related to deposit of their research output in the university repository,

which could have included training on metadata creation, including subject keywords, although five have deposited their research outputs. The only obligatory subject element, the national subject category, is considered by some to be too broad and lacking interdisciplinarity. This is not surprising since the main purpose of the categories is to facilitate the collecting of data for official national statistics on scientific publishing. The participants are also never instructed how to best add free keywords, either for the repository, or for journals in which they publish as there are no author guidelines. All this results in the need to rely on purely automatic solutions in search systems that index their work. Therefore, it is not surprising that five out of eight respondents have revealed that they have on occasion missed finding an important resource in their research, only to have found out about it too late, blaming it on the large number of synonyms, too many results or bad search habits (the latter likely connected to the lack of appropriate training).

Most PhD students resort to web search engines and discovery systems for secondary information resources: Google, Google Scholar and the local discovery system (OneSearch). Some also use Web of Science and JSTOR while only two sometimes use disciplinary databases. Primary information resources used are the English Corpora, the IRIS database for language research, SND for Swedish research data and Swedish Fornsök registry of ancient cultural heritage sites. Google and the like are appreciated for their broad coverage and ease of use, but too many results are considered as the main disadvantage. OneSearch is mostly used for known item searching. The role of controlled vocabularies in those search systems is non-existent: the systems do not make use of them at the level of the interface and the users are thus not aware of any support that controlled vocabularies may provide them with. This is, to some degree, also related to the lack of received training appropriate to the PhD level.

Similarly, their recent search pointed to similar advantages and disadvantages of Google, Google Scholar and OneSearch, with the lack of controlled vocabularies proving to be a hindrance as it requires many combinations of search terms, is burdened by problems of polysems and homonyms, as well as inadequate indexing specificity and exhaustivity resulting in too few or too many resources.

The controlled task was also a learning task because most respondents were not aware of what their disciplinary database would be, and they have never received any training related to advanced searching based on controlled vocabularies. However, it also showed how disciplinary databases that do rely on controlled vocabularies do so only in ways that information professionals may use. Also, some terms are missing, and synonyms should be many more.

This research confirms earlier reports in the literature and points to the need to:

(1) Strategic change in conducting training in information searching at the university library to cover disciplinary databases, advanced searching and controlled vocabularies at a PhD level. This should also include repository training, especially on subject keywords, which should come from an established controlled vocabulary.

(2) Introduction of controlled vocabularies in academic journals with ensuing keyword creation guidelines for authors.

(3) Make use of controlled vocabularies from metadata in systems that have them, such as discovery services, and align indexing specificity and exhaustivity policies.

(4) At search interfaces make use of controlled vocabularies by implementing mechanisms for finding relevant index terms, supporting word sense disambiguation, finding narrower and broader terms, etc., in an end-user friendly manner.

Future research should address questions relevant to these four suggestions as well as encouraging constant dialogue between software developers, metadata librarians and researchers to jointly address these complex challenges. Doing so would help fulfil established bibliographic objectives and allow users to come closer to achieving high precision and recall, the ideal of information retrieval, in this way also promoting research where nothing important is missed.

References

Alani, H., Jones, C. and Tudhope, D. (2000), "Associative and spatial relationships in thesaurus-based retrieval", in Borbinha, J. and Baker, T. (Eds), *Research and Advanced Technology for Digital Libraries. ECDL 2000. Lecture Notes in Computer Science*, Springer, Vol. 1923, pp. 45-58, available at: https://link.springer.com/chapter/10.1007/3-540-45268-0_5

Bates, M.J. (1986), "Subject access in online catalogs: a design model", *Journal of the American Society for Information Science*, Vol. 37 No. 6, pp. 357-376, doi: 10.1002/(SICI)1097-4571(198611)37:63.0.CO;2-H.

Bergenmar, J. and Golub, K. (2020), "Subject indexing: the challenge of LGBTQI literature", in Reinsone, S., Skadiņa, I., Baklāne, A. and Daugavietis, J. (Eds), *DHN 2020: Digital Humanities in the Nordic Countries: Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, CEUR-WS, pp. 203-210, available at: https://ceur-ws.org/Vol-2612/short4.pdf

Borgman, C.L. (2007), *Scholarship in the Digital Age*, The MIT Press, Cambridge, MA.

Bouazza, A. (1989), "Information user studies", in Kent, A. (Ed.), *Encyclopedia of Library and Information Science*, Marcel Dekker, Vol. 44, pp. 144-164.

Case, D.O. and Given, L.M. (2016), *Looking for Information: A Survey of Research on Information Seeking, Needs and Behavior*, Emerald, Bingley.

Catalano, A. (2013), "Patterns of graduate students' information seeking behavior: a meta-synthesis of the literature", *Journal of Documentation*, Vol. 69 No. 2, pp. 243-274, doi: 10.1108/00220411311300066.

Choi, Y. and Syn, S.Y. (2015), "Characteristics of tagging behavior in digitized humanities online collections", *Journal of the Association for Information Science and Technology*, Vol. 67 No. 5, pp. 1089-1104, doi: 10.1002/asi.23472.

Chu, C.M. (1999), "Literary critics at work and their information needs: a research-phases model", *Library and Information Science Research*, Vol. 21 No. 2, pp. 247-273, doi: 10.1016/S0740-8188(99)00002-X.

Costa, R., Salgado, A. and Almeida, B. (2021), "SKOS as a key element for linking lexicography to digital humanities", *Information and Knowledge Organisation in Digital Humanities*, Routledge, London, pp. 178-204, doi: 10.4324/9781003131816-9.

Dempsey, L. (2012), *Thirteen Ways of Looking at Libraries, Discovery, and the Catalog: Scale, Workflow, Attention*, EDUCAUSE Review, available at: https://er.educause.edu/articles/2012/12/thirteen-ways-of-looking-at-libraries-discovery-and-the-catalog-scale-workflow-attention

East, J. (2007), "Subject retrieval from full-text databases in the humanities", *Portal*, Vol. 7, doi: 10.1353/pla.2007.0018.

Encheva, M., Zlatkova, P., Tammaro, A.M. and Brenner, M. (2019), "Information behavior of humanities students in Bulgaria, Italy and Sweden: planning a game-based learning approach for avoiding fake content", in Kurbanoğlu, S., Špiranec, S., Ünal, Y., Boustany, J., Huotari, M.L., Grassian, E., Mizrachi, D., *et al.* (Eds), *Information Literacy in Everyday Life*, Springer International Publishing, Cham, pp. 295-306, doi: 10.1007/978-3-030-13472-3_28.

Ferro, N. and Peters, C. (Eds) (2019), *Information Retrieval Evaluation in a Changing World: Lessons Learned From 20 Years of CLEF*, Springer, Cham.

Flanagan, J.C. (1954), "The critical incident technique", *Psychological Bulletin*, Vol. 51 No. 4, pp. 327-358, doi: 10.1037/h0061470.

Furnas, G.W., Landauer, T.K., Gomez, L.M. and Dumais, S.T. (1987), "The vocabulary problem in human-system communication", *Communications of the ACM*, Vol. 30 No. 11, pp. 964-971, doi: 10.1145/32206.32212.

Ge, X. (2010), "Information-seeking behavior in the digital age: a multidisciplinary study of academic researchers", *College and Research Libraries*, Vol. 71 No. 5, pp. 435-455, doi: 10.5860/crl-34r2.

Golub, K. (2018), "Subject access in Swedish discovery services", *Knowledge Organization*, Vol. 45 No. 4, pp. 297-309, doi: 10.5771/0943-7444-2018-4-297.

Golub, K., Tyrkkö, J., Hansson, J. and Ahlström, I. (2020), "Subject indexing in humanities: a comparison between a local university repository and an international bibliographic service", *Journal of Documentation*, Vol. 76 No. 6, pp. 1193-1214, doi: 10.1108/JD-12-2019-0231.

Green, R. (2000), "Locating sources in humanities scholarship: the efficacy of following bibliographic references", *The Library Quarterly*, Vol. 70 No. 2, pp. 201-229, doi: 10.1086/630018.

Huvila, I., Enwald, H., Eriksson-Backa, K., Liu, Y.-H. and Hirvonen, N. (2022), "Information behavior and practices research informing information systems design", *Journal of the Association for Information Science and Technology*, Vol. 73 No. 7, pp. 1043-1057, doi: 10.1002/asi.24611.

Kim, S. (2007), *The effect of users' work tasks on librarians' database selection*, University of Maryland, College Park, available at: http://hdl.handle.net/1903/7315

Kumar, A. (2013), "Assessing the information need and information seeking behavior of research scholars of MBPG College: a case study", *International Journal of Digital Library Services*, Vol. 3 No. 3, pp. 1-12.

Langridge, D.W. (1976), *Classification and Indexing in the Humanities*, Butterworths.

Li, Y. and Belkin, N.J. (2008), "A faceted approach to conceptualizing tasks in information seeking", *Information Processing and Management*, Vol. 44 No. 6, pp. 1822-1837, doi: 10.1016/j.ipm.2008.07.005.

Liu, Y.-H. and Wacholder, N. (2017), "Evaluating the impact of MeSH (Medical Subject Headings) terms on different types of searchers", *Information Processing and Management*, Vol. 53 No. 4, pp. 851-870, doi: 10.1016/j.ipm.2017.03.004.

Liu, C., Liu, Y.-H., Liu, J. and Bierig, R. (2021), "Search interface design and evaluation", *Foundations and Trends in Information Retrieval*, Vol. 15 Nos 3-4, pp. 243-416, doi: 10.1561/1500000073.

Liu, Y.-H., Thomas, P., Gedeon, T. and Rusnachenko, N. (2022), "Search interfaces for biomedical searching: how do gaze, user perception, search behaviour and search performance relate?", *ACM SIGIR Conference on Human Information Interaction and Retrieval*, New York, NY, Association for Computing Machinery, pp. 78-89, doi: 10.1145/3498366.3505769.

Liu, Y.-H., Wu, M., Power, M. and Burton, A. (2023), "Elicitation of contexts for discovering clinical trials and related health data: an interview study", *Zenodo*, doi: 10.5281/zenodo.7839282.

Madden, R. (2014), "Information behaviour of humanities PhDs on an information literacy course", *Reference Services Review*, Vol. 42 No. 1, pp. 90-107, doi: 10.1108/RSR-07-2013-0034.

Marchionini, G. (2006), "Exploratory search: from finding to understanding", *Communications of the ACM*, Vol. 49 No. 4, pp. 41-46, doi: 10.1145/1121949.1121979.

Markey, K. (2007), "The online library catalogue: paradise lost and paradise regained?", *D-Lib Magazine*, Vol. 13 Nos 1/2, doi: 10.1045/january2007-markey.

Mierzecka, A. (2015), "Information behavior within the humanities: searching or browsing, recall or precision? Researching the information needs of academics: the case study of the faculty of history of the university of Warsaw", *Zagadnienia Informacji Naukowej – Studia Informacyjne, Wydawnictwa Uniwersytetu Warszawskiego*, Vol. 53 No. 1 (105), pp. 82-95, doi: 10.36702/zin.319.

Musgrave, S. and Haugh, M. (2019), *The Australian National Corpus (and beyond)*, Australian English Reimagined, Routledge, London.

O'Brien, H., Cole, A., Kampen, A. and Brennan, K. (2022), "The effects of domain and search expertise on learning outcomes in digital library use", *ACM SIGIR Conference on Human Information Interaction and Retrieval*, New York, NY, Association for Computing Machinery, pp. 202-210, doi: 10.1145/3498366.3505761.

Organisation for Economic Co-operation and Development (2007), "Revised field of science and technology (FOS) classification in the Frascati manual", available at: http://www.oecd.org/science/inno/38235147.pdf

Savenkov, D., Braslavski, P. and Lebedev, M. (2011), "Search snippet evaluation at Yandex: lessons learned and future directions", in Forner, P., Gonzalo, J., Kekäläinen, J., Lalmas, M. and Rijke, M.D. (Eds), *Multilingual and Multimodal Information Access Evaluation*, pp. 14-25, doi: 10.1007/978-3-642-23708-9_4.

Shboul, M.K.A., Obeidat, O.A. and Sundar, N. (2019), "Humanities scholar information-seeking behavior: quantitative approach", in Al-Masri, A. and Curran, K. (Eds), *Smart Technologies and Innovation for a Sustainable Future*, pp. 177-185, doi: 10.1007/978-3-030-01659-3_20.

Shiri, A. and Revie, C. (2005), "Usability and user perceptions of a thesaurus-enhanced search interface", *Journal of Documentation*, Vol. 61 No. 5, pp. 640-656, doi: 10.1108/00220410510625840.

Sparck Jones, K. and van Rijsbergen, C.J. (1976), "Information retrieval test collections", *Journal of Documentation*, Vol. 32 No. 1, pp. 59-75, doi: 10.1108/eb026616.

Spezi, V. (2016), "Is information-seeking behavior of doctoral students changing?: a review of the literature (2010-2015)", *New Review of Academic Librarianship*, Vol. 22 No. 1, pp. 78-106, doi: 10.1080/13614533.2015.1127831.

Statistics Sweden. (2016). "Standard för svensk indelning av forskningsämnen 2011: Uppdaterad augusti 2016", available at: https://www.scb.se/dokumentation/klassifikationer- och-standarder/standard-for-svensk-indelning-av-forskningsamnen/

Svenonius, E. (1986), "Unanswered questions in the design of controlled vocabularies", *Journal of the American Society for Information Science*, Vol. 37 No. 5, pp. 331-340, doi: 10.1002/(SICI)1097-4571(198609)37:53.0.CO;2-E.

Svenonius, E. (2000), *The Intellectual Foundation of Information Organization*, The MIT Press, Cambridge, MA.

Tang, M.C., Liu, Y.H. and Wu, W.C. (2013), "A study of the influence of task familiarity on user behaviors and performance with a MeSH term suggestion interface for PubMed bibliographic search", *International Journal of Medical Informatics*, Vol. 82 No. 9, pp. 832-843, doi: 10.1016/j.ijmedinf.2013.04.005.

Tibbo, H.R. (1994), "Indexing for the humanities", *Journal of the American Society for Information Science*, Vol. 45 No. 8, pp. 607-619, doi: 10.1002/(SICI)1097-4571(199409)45:83.0.CO;2-X.

Tudhope, D., Binding, C., Blocks, D. and Cunliffe, D. (2006), "Query expansion via conceptual distance in thesaurus indexed collections", *Journal of Documentation*, Vol. 62 No. 4, pp. 509-533, doi: 10.1108/00220410610673873.

Vakkari, P. (2003), "Task-based information searching", *Annual Review of Information Science and Technology*, Vol. 37, pp. 413-464, doi: 10.1002/aris.1440370110.

Voorhees, E.M. and Harman, D.K. (2005), *TREC: Experiment and Evaluation in Information Retrieval*, The MIT Press, Cambridge, MA.

White, R.W. and Morris, D. (2007), "Investigating the querying and browsing behavior of advanced search engine users", *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, Association for Computing Machinery, pp. 255-262, doi: 10.1145/1277741.1277787.

Walsh, J.A., Cobb, P.J., de Fremery, W., Golub, K., Keah, H., Kim, J., Kiplang'at, J., Liu, Y.-H., Mahony, S., Oh, S.G., Sula, C.A., Underwood, T. and Wang, X. (2021), "Digital humanities in the iSchool", *Journal of the Association for Information Science and Technology*, Vol. 73 No. 2, pp. 188-203, doi: 10.1002/asi.24535.

White, R.W., Dumais, S.T. and Teevan, J. (2009), "Characterizing the influence of domain expertise on web search behavior", *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, New York, NY, Association for Computing Machinery, pp. 132-141, doi: 10.1145/1498759.1498819.

Wiberley, S.E. Jr. (1983), "Subject access in the humanities and the precision of the humanist's vocabulary", *The Library Quarterly*, Vol. 53 No. 4, pp. 420-433, doi: 10.1086/601405.

Wiberley, S.E. Jr. (1988), "Names in space and time: the indexing vocabulary of the humanities", *The Library Quarterly*, Vol. 58 No. 1, pp. 1-28, doi: 10.1086/601949.

Wittek, P., Liu, Y.-H., Darányi, S., Gedeon, T. and Lim, I.S. (2016), "Risk and ambiguity in information seeking: eye gaze patterns reveal contextual behavior in dealing with uncertainty", *Frontiers in Psychology*, Vol. 7, doi: 10.3389/fpsyg.2016.01790.

Wu, M. and Chen, S. (2014), "Graduate students appreciate Google Scholar, but still find use for libraries", *The Electronic Library*, Vol. 32 No. 3, pp. 375-389, doi: 10.1108/EL-08-2012-0102.

Wu, I.-C. and Vakkari, P. (2018), "Effects of subject-oriented visualization tools on search by novices and intermediates", *Journal of the Association for Information Science and Technology*, Vol. 69 No. 12, pp. 1428-1445, doi: 10.1002/asi.24070.

Yi, K., Beheshti, J., Cole, C., Leide, J.E. and Large, A. (2006), "User search behavior of domain-specific information retrieval systems: An analysis of the query logs from PsycINFO and ABC-Clio's Historical Abstracts/America: History and Life", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 9, pp. 1208-1220, doi: 10.1002/asi.20401.

Appendix 1
Controlled search task in their typical databases

Google Scholar search sequences (search terms are given in quotation marks) with results:

(1) "digital research methods in language learning" → the results considered too broad → narrowing down to "digital research methods in second language learning sla" → results also very wide, one would need to narrow it down further, e.g. "research instruments online, language testing resources"

(2) "'digital methods' 'DH methods' meta-study" → zero hits → "'DH methods' meta-study" → zero hits → "digital research methods" → 5,080,000 results → browsed through the results and identified a few publications to see if they are relevant and use citing works to find further publications

(3) "digital close reading literature" → 1,660,000 results and considered some "not exactly related, maybe, but they're interesting" → opened one book on the Google Book website and commented that she would look for it in OneSearch to obtain a copy from the university library → "digital analysis literature" (she considered "close reading" too limited and decided to add "analysis") → 3,810,000 results which she considered too many → "digital humanities textual analysis literature" → 120,000 results → "digital textual analysis literary methods" where she found some good articles and said she would consider consulting "Related searches" as well as browse all the way to the 10th results page.

Google search sequences (search terms are given in quotation marks) with results:

(1) "digital methods in English linguistics" → 17,900,000 results → "digital research methods" → 5,080,000 results → he discovered a book "The Routledge Handbook of English Language and Digital Humanities" that he considered a good book to start with although he was wondering whether digital humanities is the same as digital methods

(2) "critical perspectives digitalization archaeology" → 90,700 results out of which he identified one interesting publication from which he would check the bibliography and move on through that

(3) "digital research methods" → 1,110,000,000 results → "digital research methods gender studies" → 18,000,000 → "digital research methods history archives" → 87,200,000 results of which she founds two articles she considered would be worth examining further → one of them turned out to be irrelevant, for high school pupils.

OneSearch search sequences with results:

(1) At simple search interface: "digital research methods in qualitative interviews" → 52,973 results → narrowed it down to "review digital research methods in qualitative interviews" → 44,139 results → chose refining filter to select Qualitative Research, Interview, and Research methodology → 7,497 results with irrelevant ones on sex and from medicine → "review digital research methods in qualitative interviews -sex -medicine" to exclude sex and medicine → results included more publications on sex and medicine → "review digital research methods in qualitative interviews -sex* -medicine*" → 0 hits, upon which the participant said how they forgot how to use the search syntax accurately → advanced search interface → "review digital research methods in qualitative interviews INTE sex INTE medicine" (INTE means NOT in Swedish) as well as limiting publication date to past five years → 14,890 results including works in sports management, nursing research → "review digital research methods in qualitative interviews audience INTE sex INTE medicine" → 3,595 results and found a few relevant works on social science research methods, media audience research, big data and digital research and considered them to serve as a good starting point.

(2) "English linguistics digital methods" → 12,837 results and commented that it was too many and then limited to works that are available in the physical library → 1 result. He was not satisfied and went to Google (see above).

(3) "antconc sfl ideational metafunction" → 1 result, commenting that she does not know whether all the keywords were included in this search; she could find "metafunction", "sfl" but not "antconc" and "ideational" → she would consult the article to identify any relevant keywords for further searching. She also commented how she feels inexperienced in the topic, and this is an example where the search system should help inexperienced users, non experts, to find their way, but fail to do so.

Appendix 2

Controlled search task in disciplinary databases

LLBA (Linguistics and Language Behaviour Abstracts):

(1) Starting at the advanced search interface, as this is its home interface: "digital research methods in language learning" filtering on scholarly journals, articles and English language → too broad results → the interviewer reminded he could use Thesaurus to find good keywords → opened the Thesaurus search interface → "research methods in SLA computer aided" → the Thesaurus suggested related terms of which he chose "Aptitude Tests" but couldn't find other interesting terms → he added that to the search query that now automatically read "su (Language Aptitude Tests)" → found three books but the ones he knew about did not seem to have been covered by the database → "language attitude AND test validity" → found directly 5 relevant → opened a metadata record of one and clicked on a relevant subject to retrieve more relevant ones. He appreciated it and commented that although the learning curve may be high, it might be worthwhile to learn the system.

(2) "digital research method AND digital humanities" and applied the filter function for publications from last three years → 6 hits which he considered relevant → opened one article's metadata and clicked on its subject "Digital Humanities" → 36 results and he very much liked this feature → chose Thesaurus browsing list under the guidance of the interviewer → clicked the S in the "Browse terms" option, and selected the term "Second Language Learning" → some troubles adding the narrower search terms to the search query (not end user friendly) so he had to paste it himself "subject ('second dialect learning')" → 7 results → went back to the Thesaurus interface to find term "Sociolinguistics" and added to the search query → "subject

('second dialect learning') AND subject ('Sociolinguistics')", published in last 3 years → 0 results → the interviewer described the difference between AND and OR → "subject ('Second Language Learning') OR subject ('sociolinguistics')" → 116 results which he considered was very practical → he used these steps to learn about the interface and now he went to look for terms for the search task → "digital" in the Thesaurus search box and chose "Begin with" → one term retrieved, "Digital Literacy" → "computational" in the Thesaurus search box and chose "Begin with" → two terms retrieved, "Computational Linguistics" and "Tagging (Computational Linguistics)" → added them to the document search interface → 8 documents retrieved but none were about research methods → "methods AND subject ('Computational Linguistics')" → 254 results.

RILM (Répertoire International de Littérature Musicale):

(1) "digital research method" (explaining that in disciplinary database one does not need to specify the discipline) → 8 results, some of which she considered relevant to examine further → in one article she chose "Music and related disciplines – social sciences, media studies and public culture" from "Major topics" → 398 results but these were no longer about research methods → chose the Subject interface under the guidance of the interviewer which she explored to find the right term and learned that the best way is to be specific about naming a method (e.g. Schenkerian theory) while the general "method" does not find any terms. The Subject browsing interface is hard to find; one needs to go to "More" → "Indexes" → "Subjects".

Film and Television Literature Index:

(1) "literary methods AND digital humanities" → 0 hits → "literary methods AND digital" → 11 results but not judged relevant → "close reading AND digital AND techniques" → 18 results which she also did not consider relevant and commented that the reason she did not like disciplinary databases is that one needs broader searches to learn about the theme first to then be able to use right keywords → "literature AND computer AND textual analysis" → 11 results → "literature AND textual analysis AND digital" → 5 results → opening a metadata record upon suggestion by the interviewer to consider subjects added there but she did not find any relevant ones → went to the Subject index under the guidance of the interviewer → it was hard to find appropriate terms, e.g. "literary methodology" was not in the list of terms and mechanisms like broader terms were not there to support the user.

MLA (Modern Language Association):

(1) "English linguistics digital methods" → 5,658 results among which he identified one book considered worthwhile exploring further → interviewer suggesting to look at the "Browse Thesaurus" function → "digital methods" → 0 results → "digital humanities" → found the term with that name but considered it not relevant enough → "digital linguistics" → 0 results → he commented that the question is whether linguistics is not already digital in its use of research methods and that there may not be a need to specify the methods as digital → played with some more terms but considered the interface complicated → went to Google books to find more relevant keywords → chose from the Thesaurus "quantitative methods AND quantitative methods" → 7,229 results but the system wrote "Note: your initial search query did not yield any results. However, using Smart Text Searching, results were found based on your keywords" → he found some relevant publications. But what is the value of thesaurus here?

ERIC (Educational Resource Information Center):

(1) "antconc AND sfl AND ideational AND metafunction" → 0 results → "antconc AND sfl" → went to Google to make sure her spelling of "antconc" was accurate as it was → "antconc AND grammar" → this broadening retrieved 4 results that were not too relevant → "antconc AND grammar AND functional AND systemic" → 0 results → the interviewer suggest to try out the Thesaurus → "antconc" was too narrow and had no thesaurus term → "tool" → retrieved "Research tools" terms which she found relevant, but could include also non-digital tools, so this was not most relevant after all.

Historical Abstracts:

(1) "archaeology AND digitalization OR digitalization AND critical" → 0 results, commenting how maybe this database is not good for this theme and he would normally not continue exploring it but go to Google → the interviewer suggested to go to "Browse an Index: Subject Terms" → he did not find the interface friendly but found "archaeology" → searched for "digital" but only found "digital preservation" → added the two terms to search query with OR → 3154 → the interviewer explained the meaning of OR → changed it into AND → 272 results but the system showed "Note: your initial search query did not yield any results. However, using Smart Text Searching, results were found based on your keywords" → he considered some of the results relevant → opened metadata of one article upon interviewer suggestion → discovered an additional relevant search term, "digital technology".

(2) Felt intimidated with all the search options → "digital research methods" → 12 results, some of which they considered relevant, but would not cover parts of the world outside of Europe and the United States that the participant would also be interested in → the interviewer suggested to use Subject Terms in Browse an Index → "digital research methods" resulted in 0 results because it is an alphabetical listing → in Browse for: but s/he got no results. The system showed "The terms(s) you entered could not be found. The list below is in alphabetical order" → they wanted to give up here and go to Google again → "colonia philippines" in Browse for → no results on Philippines, only other colonies.

Corresponding author
Koraljka Golub can be contacted at: koraljka.golub@lnu.se