Interacting with digitised historical newspapers: understanding the use of digital surrogates as primary sources

Elina Late and Sanna Kumpulainen
Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland

Received 9 April 2021
Revised 1 July 2021
Accepted 4 July 2021

Abstract

Purpose – The paper examines academic historians' information interactions with material from digital historical-newspaper collections as the research process unfolds.

Design/methodology/approach – The study employed qualitative analysis from in-depth interviews with Finnish history scholars who use digitised historical newspapers as primary sources for their research. A model for task-based information interaction guided the collection and analysis of data.

Findings – The study revealed numerous information interactions within activities related to task-planning, the search process, selecting and working with the items and synthesis and reporting. The information interactions differ with the activities involved, which call for system support mechanisms specific to each activity type. Various activities feature information search, which is an essential research method for those using digital collections in the compilation and analysis of data. Furthermore, application of quantitative methods and multidisciplinary collaboration may be shaping culture in history research toward convergence with the research culture of the natural sciences.

Originality/value – For sustainable digital humanities infrastructure and digital collections, it is of great importance that system designers understand how the collections are accessed, why and their use in the real-world context. The study enriches understanding of the collections' utilisation and advances a theoretical framework for explicating task-based information interaction.

Keywords Newspapers, Digital libraries, User studies, Behaviour, Task analysis, History, Task based information interaction

Paper type Research paper

Introduction

If we are to design sustainable digital humanities infrastructure and digital collections, it is highly important to understand how and why these systems are accessed and used, in real-world context. Otherwise, some crucial aspects of work practices that shape the digital tools' and platforms' use could get ignored. To address this issue, we conducted a qualitative research aimed at providing a user-centred picture of digitalised historical newspapers' utilisation that covers all activities involved in the research process.

Digital texts have become integral to research in the humanities and history (Late *et al.*, 2019; Sinn and Soares, 2014). With the digitisation of the materials have come concepts such as the digital historian, digital history (Crymble, 2021; Gregory, 2014) and computational history (Nanetti and Cheong, 2018). They involve applying digital materials and/or

This research was funded by Academy of Finland grant #326612. The authors thank Laura Korkeamäki for providing research data for the purposes of this study.

Journal of Documentation
Emerald Publishing Limited
0022-0418
DOI 10.1108/JD-04-2021-0078

computational methods in research. Multidisciplinary research work involving fields such as computer science characterises many of the new digital research methods. Historians' traditional individual-oriented research practices have given way to more collaborative efforts (Given and Willson, 2018).

The start of the 21st century has seen public and private organisations around the world digitalise historical newspapers (Gooding, 2016a; Mussell, 2012). This digitisation has influenced the work of historians in at least two ways. Firstly, most of the collections display ubiquitous availability: historians can use them from their own devices, at any site. Secondly, many user interfaces serving as a gateway to digital newspaper collections offer advanced techniques for searching and analysing the contents. The standard procedure of optical character recognition (OCR) enables instant text search. In addition to basic full-text and keyword-based search, typical functions include providing metadata, browsing and filtering of results. The most advanced interfaces offer functionality for user interaction (e.g. saving of articles to "Favourites"), content enrichment (e.g. post-OCR correction), connectivity (e.g. links to other repositories) and code extensions and APIs (Ehrmann *et al.*, 2019).

Most studies of historians' information-related practices have focused on the use of archival materials in general (Toms and O'Brien, 2008). There is a dearth of research into the real-world utilisation of digital collections (Allen and Sieczkiewicz, 2010; Bulger *et al.*, 2011; Hughes, 2012a; Meyer *et al.*, 2009; Sinn and Soares, 2014). Among others (e.g. Bulger *et al.*, 2011; Warwick *et al.*, 2008), Gooding (2016a, b) has appealed for qualitative research, to inform our understanding of scholars' practices in the digital world. We set out to address this research gap and study qualitatively history scholars' information interactions in the realm of Finnish digital newspapers and their use as primary research sources – that is, as sources of research data that provide first-hand testimony or direct evidence pertaining to the topic under historical investigation.

We analysed material from qualitative interviews by means of a model for evaluation of task-based information interaction (Järvelin *et al.*, 2015) that enables studying how a given digital collection is used in varying task activities during the research process. It is widely agreed that information behaviours do not occur in a vacuum; they stem from the larger motivating tasks (Toms, 2011).

We begin by discussing previous research focusing on historians' information practices. Then, we present the framework for our examination of task-based information interaction and describe the research setting, along with the data-collection process and the analyses conducted. With these foundations in place, we proceed to the results and lay out some discussion and our conclusions.

Historians' information practices in a digital information environment

The increased usage and ongoing development of technological tools for humanities research has changed history scholars' ways of working over the past decade (Baruchson–Arbib and Bronstein, 2007; Given and Willson, 2018; Toms and O'Brien, 2008). This change has not been entirely smooth. For example, while Burton (2005) noted that mainstream academic historical journals began to accept digital content as a part of their publication in around 2000, work with digital formats has not been rewarded in the same way as that with more traditional forms of history scholarship (Clement and Carter, 2017). Also, some scholars have expressed concern about the impact of digital technologies on reading: printed texts have been seen as more engaging for deep reading than digital ones are (e.g. Carr, 2010). Nevertheless, digital collections are here to stay, and many historians already see them as essential for conducting research (Sinn and Soares, 2014).

Several scholars have delved into how historians seek and locate primary source materials (Anderson, 2004; Duff and Johnson, 2002; Tibbo, 2002). They report on the variety of sources used in history research processes, and they point out a need to assess what the users want and need, and how primary sources are searched for.

Sinn and Soares (2014) conducted a survey of authors with articles in *American Historical Review* to understand how history scholars were using digital archives and what kinds of impacts the digital archival collections have on their research activities. Scholars stated that they found digital collections useful in that they were easy to access and saved time and effort; however, many of them favoured the original format, finding this more authoritative and "special". No correlation was visible between the scholars' technical skills and their use of digital collections, and the findings indicate that history scholars of all stripes consult digital collections throughout the research process. That said, what was meant by "use of digital collections" was unclear, and uses probably varied over the research process. In a further complication, that process was seen as non-linear, looping back and forth and even described as "haphazard". In the course of the research process, historians discovered new materials with potential utility in future studies. Interestingly, Sinn and Soares did not address the role of computational methods in searching and analysing digital materials.

Kumpulainen *et al.* (2020) qualitatively studied historians' desired avenues of access to information and how these manifest themselves in the document collections. Historians focusing on the Second World War have several specific needs that cannot be met via such typical digital-domain methods as entity recognition. Among these are military rank, gender and kinship. The study points to possible ways of speaking to such specific information needs. Moving from the stage of approaching information to that of working with it, Hoekstra and Koolen (2019) examined the interactions in selecting, enriching, connecting between, analysing and evaluating historical data with digital tools. They propose the concept of data scope as an instrument encompassing the interactions and recommend expanding the idea of source criticism toward the digital tools. Finally, in a study of the digital information activities involved in historians' work flows, Koolen *et al.* (2020) found that, for the ability to design better digital history-research infrastructure – and evaluate it – a palette metaphor would be more useful than a pipeline model, on account of the simultaneity and non-linear flows of the activities.

A few studies have examined the use of digital newspapers. An interview study by Allen and Sieczkiewicz (2010) showed that scholars in the US employed digitised historical newspapers for such purposes as checking facts (names, dates and locations), collecting information about larger issues (such as elections) and ascertaining public opinion. Scholars also reported using newspapers to fill gaps in research and corroborate information from other sources. This study revealed a tendency toward browsing as opposed to searching. When searching, these scholars used mainly keyword search filtered by topic, date or name. Historians were interested also in specific content types, such as advertisements, editorials, obituaries and death notices and images. Scholars often printed out the results/contents to read later. In other work, a log-based study showed that keyword queries over newspaper collections often include named entities such as personal and place names (Chardonnens *et al.*, 2018) and that some eras are searched more often than others (Gooding, 2016a). With regard to search types, a survey among users of historical-newspaper collections revealed a desire for the availability of image search, named-entity recognition, a tool for excluding certain search elements and visualisation tools. In addition to advanced tools, users desire high-quality content and the ability to save, copy and annotate articles (Oberbichler *et al.*, 2019). Already, some collection providers have started to collaborate with scholars for building virtual laboratories with advanced features and tools (Hauswedell *et al.*, 2020).

Although scholars find digital collections useful, it has been argued that so-called digital surrogates do not, and cannot, replace original paper documents (Conway, 2015; Sinn and Soares, 2014) and that digitised newspapers can even skew historical research (Milligan, 2013). Since digitised contents are not commensurate with the originals Mussell's (2012) calls for historically reflexive media literacy skills and emphasize the need to understand the influence of digitisation on the contents. The poor quality of OCR output is a common

problem in digitised newspaper collections. A report from one study cites 50%–70% accuracy of OCR at word level and levels between 71 and 98% on character level for Finnish historical-newspaper data (Kettunen *et al.*, 2014). Jarlbrink and Snickars (2017) argue that the combination of misinterpreted words (OCR errors) and "random" text created by auto-segmentation tools generates content that was never actually written.

Scholars find it important that technology support their ways of working, irrespective of the choice of tools or whether one is working with print or digital text (Given and Willson, 2018; Hughes, 2012b). Scholars in the humanities actively participate in the development of tools, so much so that Given and Wilson (2018) posited that tool-development and data-preparation are new research practices in their own right. Other practices they identified are text analysis (e.g. employing computational analysis and text-mining), methods such as empirical data-collection via interviews etc., and traditional scholarly work (e.g. reading, thinking and writing).

Task-based information interaction

Information interactions do not arise out of nothing; they are triggered by other tasks, either leisure- or work-related (Toms, 2011; Vakkari, 2001). Research into information interaction, or human–information interaction, focuses on people's cognitive actions and behaviours with information, rather than with technology or librarians (Fidel, 2012). Information needs and activities are rooted in the underlying larger task. Accordingly, Järvelin *et al.* (2015) developed the task-based information interaction (TBII) model to assess how the information interactions contribute to the goals of task performance (Figure 1). This is a theoretical model that encompasses inputs to the process (contextual factors, perceived task and actor); five activities: task-planning, searching information items, selecting them, working with them and synthesis and reporting; and outputs and outcomes of the process. The activities cover a wide range of cognitive behaviours:

(1) *Task-planning* is a meta-level activity present in every stage of performing a task. It includes the performer's understanding related to the task at hand and required procedural knowledge. The activity evolves during the task's performance, leading to a clearer, more structured understanding of the task at the end of the task-performance process.

(2) *Searching information items* entails interactions with a search system and retrieval of the information items.

(3) *Selecting information items* involves making decisions about the relevance or usefulness of the items found. It is very tightly interwoven with the search activity and with the work with information items.

(4) *Working with information items* is about scanning and browsing items, reading and annotating them and comparing and linking them. Annotations provide one possible basis for querying and exploring the materials.

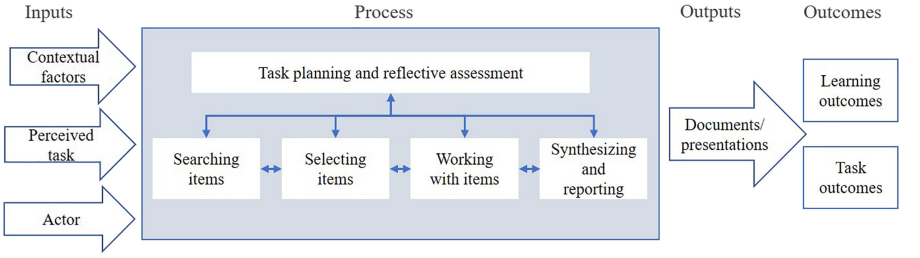

Figure 1.
The Task-Based Information Interaction (TBII) Evaluation framework (adapted from Järvelin *et al.*, 2015)

(5) *Synthesis and reporting* is associated with the task outputs and outcomes. This is an essential part of the knowledge work of history researchers and other academics. Synthesis requires the writer to integrate information from diverse information sources for creation of new knowledge and new information items and objects.

While the model presents the activities as sequential, separate components of the process, they are analytical constructs representing elements that sometimes cannot be separated in the actual task performance. These individual activities are interrelated, simultaneous and – especially in history research – cyclical or iterative rather than pipelined processes (cf. Koolen *et al.*, 2020).

The research setting

The research question

Our study, focused on the real-world research use of a Finnish historical-newspaper collection, was designed to answer the following research question:

RQ1. What kinds of information interactions feature in the activities related to using the historical newspapers as primary research sources?

Our research design took advantage of the model's articulation of the various activities. This enabled covering a wide range of research activities that involve interactions with the contents of the newspaper collection in question.

The framework formed the theoretical backbone for our interview-guide design and the analysis of our findings. Accordingly, its applicability in research of this nature is discussed in detail further (see "Discussion"). By evaluating the use of a prominent digital collection in the context of research work, we were able to form a detailed picture of the material's use as a primary research source.

Collection of qualitative data

To collect the data, we carried out in-depth qualitative interviews with academic history scholars who were using a large Finnish historical-newspaper collection for research purposes. The National Library of Finland (NLF) holds digitised historical newspapers published in Finland from 1771 to 1929, available to citizens and scholars alike. This collection, containing approximately 7.4 million newspaper pages, in the Finnish, Swedish and Russian languages, is offered openly via a Web-based service (digi.kansalliskirjasto.fi). Contents are available as PDF pictures of the pages and as text format (OCR). The online service provides tools for search, collection and analysis of newspaper data, with some advanced functionality available. For example, users can copy and paste selected data from the collection to a "scrapbook" and download the selected data in MS Excel format. All the contents of the historical newspaper collection contents are available also via the Language Bank of Finland as downloadable text files (Kettunen and Pääkkönen, 2016).

The NLF collects information about the users of the newspaper collection. Via this list, we identified prospective interviewees and contacted them by e-mail. In total, there were 13 unique interviews, conducted by three researchers. These were done in three rounds: the first interviews (PA1–PA4) were in spring 2018, the second set (PB1–PB5) in spring 2020 and the third set (PC1–PC4) in autumn 2020. Triangulation of interviewers enabled us to transcend any possible biases arising from the personality of any single researcher (Kumpulainen, 2017).

While the interviews in 2018 were face-to-face encounters, the pandemic led us to conduct those in 2020 online, using the videoconferencing tool Zoom. All interviewers experienced the

online interviews as both an economical solution and a good way of collecting the research data. In the face-to-face setting, the conversation was audio-taped, and video recordings were made from the online interviews. The video and audio files were transcribed in full for analysis. The average interview length was 60.8 min, and the interviews' audio data run, in total, 13 h and 18 min.

In advance of the interviews, we developed their themes and a list of total 45 interview questions on the basis of the TBII model. Thus, we ensured that each stage of information interaction was covered during every interview. However, we did not necessarily follow the order of questions presented in the interview guide; it functioned as a checklist for keeping track of the interview. Example interview questions are presented in Table 1. Interviewees received a general outline of the interview themes by e-mail before the interview, and informed consent with a permission to publish quotations from the data was obtained.

The interviews began with collection of background information such as current status, fields of research and research experience. After this, interviewees were asked to describe their work at a general level. Conversation then was directed to specific interview themes. We utilised a variation of the critical incident technique (CIT) (Flanagan, 1954) in which interviewees are asked to describe an ongoing or recent research project – a task – in which they used the historical-newspaper collection. The CIT method helps both the interviewee and the interviewer to reflect more profoundly on the subject. It has been widely used in studies of information behaviour (Marcella *et al.*, 2013).

The interviews included requesting a demonstration of the informant's ways of using the newspaper data. Interviewees showed us their data and demonstrated the analysis tools they had used. In face-to-face interviews, the demonstrations were video-recorded; in the online interviews, interviewees could easily "share their screen" with the interviewer. Interviewees also displayed and shared publications and other outputs from the research project. Thereby, the interviewers could gain a more accurate picture of the information interactions of the interviewee.

Table 1.
Examples of initial interview questions related to the various activities in the TBII model

Theoretical construct	Interview questions
CIT anchoring task and task background	- For what type of research project have you used the digital newspaper collection as a source of research data? - Can you differentiate any phases in the research process? - Are you working alone or as part of a research group? - Could you describe your research group?
Task-planning	- Where did you get the idea for the project? - What are the goals for the project? - What are your research questions? How were they formulated?
Search	- How did you become aware of the collection? - How was searching for data in the collection handled? What problems did you face?
Selection	- How did you select the data? - Where did you save the data, and in what format? How were the data organised?
Working with items	- How were the data processed and analysed? Why? - What kinds of research methods did you use? - Did you collect data from other sources?
Synthesis and reporting	- What kind of results did the study produce? - How and where were the results presented or published? - Have you considered opening up your data? How and where?

Analyses

The dataset was subjected to content analysis with ATLAS.ti software. While one researcher handled the analyses, all coding was discussed in detail with another researcher in several rounds over the course of the analysis process, after all researchers read the transcripts several times.

Analysis comprised three steps: iterative readings of the interview transcripts, open coding and selective coding (Strauss and Corbin, 1997). In the first round of coding, we identified the task-specific goals for the newspaper collection's use from the data. During the interviews, the scholars interviewed described ongoing or recent research tasks (such as writing a scholarly article) for which they utilised the digital historical-newspaper collection. In practice, many scholars reported on more than one task as the interview progressed. For example, one described the research processes connected with writing two articles that formed portions of her PhD thesis. In all, participants articulated 19 tasks for which the newspaper content was used for primary research sources. Our dataset encompasses descriptions of other types of task also, such as creating applications, checking facts, teaching and orienting oneself; however, these tasks were excluded from the analyses since they were not explained in detail.

Next, we coded for the activities (task-planning, searching, selecting, working with items and reporting and synthesis) and for the information interactions within them. Sometimes, it proved difficult to distinguish between activities (e.g. searching vs. selection), and the entire research team discussed these problematic situations. After this, the codes were cross-checked and, as necessary, revised. The coding was followed by compilation of the characteristics of tasks and coded activities in an Excel spreadsheet. This afforded examination of the interactions within each task and each activity. We chose illustrative extracts for the various activities and information interactions, loosely translating them from Finnish to English. Quotations were selected the anonymity of individual interviewees in mind.

Results

Descriptive findings

Our data include 19 articulations of research tasks wherein the digital newspaper collection was used for primary source materials. These tasks all were research projects such as descriptions of writing a scholarly article. Characteristics of the tasks are presented in Table 2.

As soon as it was launched, the digital newspaper collection quickly became a popular source of research data recognised for the new opportunities it offers for digital history research in Finland.

> After it was digitised, it changed the field of history research in Finland a lot. [. . .] The question is not what materials have been used but how they are used. (PA2)

Collecting and analysing data via computational methods is a typical feature of the tasks we analysed. The digital newspaper collection has opened new possibilities for researchers and probably encouraged exploiting new research methods. Accordingly, this type of research has gained increasing funding in Finland in recent years. However, our sample is not representative of all uses the collection serves and is likely to be biased toward computational history.

Integrating several types of primary source materials was not very common in our data. In fact, more than half of the tasks relied on digital newspapers alone as primary sources. The newspaper collection is a rich data source, so it is not always necessary to collect data elsewhere. Projects using computational methods relied especially often on a single data

source. Another commonality among the tasks we studied is the research's multidisciplinary nature. This is unsurprising: applying computational methods often requires expertise in multiple disciplines, including technical sciences.

Next, we consider the interactions within each class of activities in the TBII model in the participants' use of the newspaper collection for primary sources.

Task-planning

Task-planning brings in one's understanding of the task at hand and requires procedural knowledge. The activity evolves during the task performance, leading to a more structured, clearer conceptualisation of the task by the end of the research process. When asked in the interviews whether they could distinguish any particular phases in their research projects, the scholars listed collecting secondary and primary sources for their research, analysing the data and writing up the results. This is in line with our chosen framework and the activities therein. However, all scholars emphasised that the process was not linear; rather, it jumped backward and forward. For example, an interviewee working with a research project examining the history of newspapers explained:

> First you need to have an idea, then the data collection, then the analysis and writing. I guess that is it. But, at least for me, they are happily overlapping and muddled. (PB2)

Most scholars described the research process as data-driven. Research topics emerged from the data, and the data available determined what could be studied. Usually, the work started with browsing the contents of the newspaper collection via the NLF Web interface. The historians found it vital to understand how the collection was built and what it contained and did not contain. It was important for the scholars to develop a general sense of the data. Sometimes, they browsed the contents with an open mind, seeking something new to study. The goal was to identify a set of data or an as-yet-unexplored phenomenon. In the same phase, the scholars read the secondary sources. When the interviewees already had a research question in mind, they would try to find out whether and how the newspaper collection could function as a data source. Typically, they already had a research topic in mind, found during some other research project; however, it took a long time to formulate the problem precisely, and problem statements evolved throughout the research project.

Acquiring the materials is one part of the task-planning. For this, all scholars used the NLF interface for searches of the contents, with some also downloading OCR text files of newspaper content from the Language Bank of Finland. Most scholars worked with their own

Table 2.
Characteristics of tasks connected with using digital newspaper materials as primary research sources (number of cases)

Research topics	Development of newspapers (9)
	History, development and use of language (4)
	History of a certain societal phenomenon (3)
	The history of individuals or a group (2)
Research methods	Digital history (7): applied digital data, collected the data manually, or analysed the data manually
	Computational history (12): used digital data or used computational methods to search, select and analyse data
Source of data	The NLF interface (8)
	The Language Bank of Finland (11)
Integration of data	The digital newspaper collection as the entire set of primary sources (11)
	Retrieval of primary sources from many venues (e.g. other digital collections also) (8)
Research group	Working alone (10)
	Working as a member of a research group (9), possibly multidisciplinary (7)

computer, but some used external services for server space. Setting up the permissions for use of the servers took time.

Some of the scholars worked alone, while others operated in multidisciplinary research groups. When the task was handled in collaboration with others, discussions with the research group formed an important part of the task-planning phase. The scholar's skills too had a large impact on what could or could not be done. A scholar working with a multidisciplinary project on computational history explained the significant role of computing skills thus:

> In the analyses phase, it is very important that everyone is able to participate. Scholars need to have programming skills so that they can handle the data and perform analyses independently. (PC2)

Searching

Searching information items entails interacting with the search system and retrieving the information items of interest. To search for data from the newspaper collection, scholars usually browsed or entered queries for content, with some specific theme in mind. Some topics were not amenable to keyword search; for these, the scholars needed to browse the newspapers page by page. For this, they went through the scanned photos of the newspapers.

> I think it was necessary to browse them thoroughly, page by page, because you cannot find opinion pieces by querying. Of course, I can do searches when I know the titles, but I cannot trust that the OCR search will find them all. (PC3)

Queries employed keywords of various sorts, such as names of authors, place names, terms found in the body text and the titles of relevant articles or columns. Search was often limited to a certain time period. In the simplest case, scholars studied the use of specific words and could use those words as search terms. In most cases, however, the situation was more complex and necessitated using various search techniques offered by the NLF interface. For example, "proximity search" was often mentioned. Scholars needed to experiment with multiple forms of their keywords, to take into account spelling variants and typos in the newspaper data. Because the collection is multilingual, the scholars needed to formulate their searches with several languages in mind.

Sometimes searching started with broad keywords. By browsing a small set of documents from among the search results, the scholars learned what type of material broad searches generated. To narrow the set of results, they performed new searches, with more specific keywords. Simultaneously, they learned about the collection as a whole and then could share their findings with colleagues and apply the information in their future projects. One scholar studying the history of societal phenomena described collecting the data in "sections", year by year, and at the same time honing the research setting.

> I do a lot of testing, in different ways. Firstly, I go through a small set of search results with broad keywords [. . .]. This manual reading makes it slow, but the good side is that I keep track of what works. If the search is too broad, I can specify it. It also helps me to specify the research setting. (PB4)

In one task tackled by a research group, scholars distributed the retrieval tasks among the members of the group. They had to keep track of which keywords were used and how the queries were formulated, so they took notes for later use.

> I use these auxiliary files. I take notes in Word or Notepad such as "make this query" or "try with this keyword" if I don't have time to make the query at the moment. (PC3)

Selecting

Selecting material is a process of assessing the relevance or usefulness of the information items found. Search and selection of data were often done in parallel. Scholars explained

about concentrating on, for example, a specific theme to give focus to the search and thereby create their own corpora. Data search and selection was laborious, taking weeks or months of work.

The scholars collected the selected research data on their own computers or other tools external to the NLF interface (e.g. in MS Excel or Google spreadsheets). The NLF interface allows users to paste from the articles and create their own scrapbook from the newspapers' contents. Scholars who utilised the scrapbook approach could download the search results in an Excel spreadsheet providing selected metadata for the articles and add their own notes in Excel.

To become familiar with the search results, scholars often needed to skim or read the text in the image-based PDF files from the newspapers, for a more accurate picture of the contents. These files rendered it easier to detect article boundaries – because the newspapers are scanned page by page, scholars had to check whether the article continued on another page. Also, low-quality OCR data sometimes proved unreadable, so scholars jumped between the OCR and PDF files.

> It is hard to make sense of the machine-read texts, so it is easier to read it from the newspapers with the Gothic font. So the work is done back and forth. We read and work with both, the OCR materials and the digital newspapers from the National Library's portal. (PB3)

The data collected typically consisted of articles from the newspapers and/or metadata for the selected articles. Scholars saved mainly the OCR text of the articles, not the PDF pictures. The resulting corpus varied in size with the research topic but comprised thousands of articles in most cases. Describing her data, one scholar explained the need to browse the contents in these terms:

> At the moment, I have six or seven hundred articles and my aim is to reach over a thousand. [. . .] I am selecting them by hand because I think that, although you would be able to save them automatically, from the point of view of a historian we still don't know well enough what the 19th-century newspapers contain. Because of that, this manual work is very useful. (PB4)

As they were selecting the data, the scholars often performed analyses at the same time and took notes about the data. Also, they collected statistical details related to the contents, such as the number of certain types of articles found. Interviewees sometimes organised the data in Excel spreadsheets by category and calculated frequencies for certain categories. In addition, they were interested in the metadata available for the newspapers.

Working with items

The fourth activity in the TBII model, working with items, encompasses such actions as scanning and browsing the documents, reading and annotating them and comparing and linking materials. Scholars analysed the newspaper data by various qualitative and quantitative means. The most common tactic was to read the content and compare the information obtained with that from prior literature, in what is known as close reading. One type of this involved using certain synonyms and antonyms to study concepts' life cycle. Scholars utilised further sources – for example, from other archives – to enrich the view produced from the newspaper data. In most cases, the interviewees read the material online, but sometimes they printed articles out for reading. Scholars used colour coding and wrote marginal notes on the printouts.

Scholars analysed the metadata too. For instance, some interviewees counted the number of items of certain types published, such as reader letters, to examine how newspaper content developed. Some were interested in material aspects of the newspapers and considered the number of pages, headlines and columns used for headlines, along with how these features

developed with time. A scholar studying historical language use explained his research methods:

> I often search for relevant materials from the newspapers, and then I read and analyse them by close reading. Lately, I have started to do more computer-assisted research, with a quantitative approach. I have calculated word frequencies, for example – relative and absolute word frequencies – and that way I've tried to understand single news items. (PA2)

Interviewees chose various computational methods for their data analysis. Some employed only qualitative methods, while others utilised both. An important aspect of this was reading the pieces' text, which was always a part of the analysis: the scholars never based their studies on quantitative methods alone. When involving computational methods, the research process was not linear: but interviewees used several tools and approaches (searching, browsing, etc.) to obtain a more precise picture of the phenomenon studied. The sheer size of the body of data available encouraged the use of computational methods, with the Language Bank of Finland corpus often subject to computational analysis. In one case, a research group developed their own interface for their data.

> For us, it has been important that we have taken the whole newspaper collection as metadata ALTO files. We have indexed the data, and we have our own interface. We have our own API that we can use for exploiting the data. (PC2)

Because of the historical language and the poor-quality OCR output, considerable data-cleaning work preceded the actual analyses. Some errors in the data were systematic, so processing to rectify them was straightforward. For example, scholars harmonised the orthography by changing certain letters (for example, the historical "*w*" was replaced with the modern Finnish "*v*"). Also, scholars used stemming applications to convert words into their base form. This helped address the challenges posed by automatic processing of Finnish, which is a highly inflected language (e.g. Kumpulainen *et al.*, 2020).

The scholars varied in their tools for computational analysis, and many described participating in tool development as a member of a multidisciplinary research group. The computational analyses were based on queries over the corpus (either one generated via their earlier search and selection or the full Language Bank corpus). In one case, the interviewee had collaborated with a computer-science scholar to develop a system for named-entity recognition: this scholar manually annotated newspaper content by identifying place names in the data, to contribute to the training of the system, and participated in evaluation of the result quality by comparing the manual annotations with the system's automatically generated ones.

> I have 500 cases categorised manually, and the computer provides the same categorisation for the next 500. (PB4)

Another case involves scholars utilising a search interface built for analysing how material has been copied between newspapers. This search interface, freely available online, made it possible to search for clusters of appearances of the same text snippets. This affords research into, for example, how rapidly, in what way and where news items were spread.

> There are different variables for what can be studied and analysed. For example, the temporal continuum [. . .]. It would have been impossible to find the copied text manually, but it can be done with a computer-assisted method. (PB3)

Many scholars utilised methods from corpus linguistics to study the language used by counting the number of words in the text. Scholars compared, for example, the number of words or the context of the words between types of articles or between corpora (e.g. newspapers produced by different political parties). Other work involved studying how

the use of a particular concept has developed over time in the newspapers. No matter the type or purpose of the analysis, reading was always present in the analysis of the texts.

Visualisation of the data and results was important. Most commonly, data were presented via a map for depicting the movement of news between cities and countries. A tool called Palladio, developed at Stanford University, was cited especially often. Data could be shown on a map with recognised place names and their co-ordinates. One scholar stated that, via the graphical presentation, scholars learn about new features of the collection. For example, it came as a surprise that the Finnish newspaper collection contained papers published outside Finland. This was detected when metadata were used for creating a map showing the newspapers' place of publication.

Some scholars mentioned having ideas for using computational methods but experienced development of new techniques as difficult, slow and uncertainty-ridden, especially for someone working alone. One scholar described the future needs of historians:

> The research in digital humanities has been biased so far. There are these multidisciplinary teams and these individuals who have programming skills. In reality, most of the history scholars are left outside, and I think the National Library could develop it in a direction that would serve the scholars. You don't need to know what is under the bonnet but get tools that you can exploit and increase the quality of your own research. (PC1)

Synthesis and reporting

The final activity under the TBII model, synthesis and reporting is associated with the task outputs and outcomes. Scholars started to think about the outputs and publishing of the results at the very beginning of the research process, and they made oral presentations on their findings, research methods and tools at conferences and seminars before publishing the findings, with visualisation of results being an important part of their oral presentations. Publications on research results and tool/method development were often written for several forums, sometimes even ones of different disciplines. Some of the pieces were written in Finnish and some in English.

> Let's say that methods are published in IT forums but the results we publish [are] mainly in history-research forums. The writing style is different, the length of the text varies, so there is a massive difference. (PC1)

Our data did not feature many descriptions of the actual writing process. However, one scholar described formulating the main argument for the publication first and building an outline for the article by, for example, selecting citations from the data; the body text was written around the citations to build a solid story. Especially with regard to computational methods, the scholars found that they need to describe their collection and analysis of data carefully, to convince the referees and final readers. One scholar said that she even reported the keywords used for collecting the data.

Some scholars prepared the write-ups alone, but most articles were co-authored. Scholars who worked in multidisciplinary research groups had experienced a clear shift in the publishing culture from articles written mostly alone toward co-authored ones. Also, the individual authors had different roles in the writing process: one or two scholars did the actual writing, while others participated in data-collection and analysis work and/or commented on the manuscript.

> They were written mainly by large teams, because we have to acknowledge that in digital humanities the data preparation and processing is an essential part of the authorship, and writing algorithms is part of it. (PC1)

In one case, the principal investigator was listed as the last author for the articles, in line with the culture of the natural sciences. When writing collaboratively, scholars utilised different digital tools and applications for the writing itself versus discussing the content, publication forums, timetables, etc.

> I have mostly used Google Docs, and it works well enough. We have our shared Slack channel, where we discuss [things] before the writing, timetables and such. (PA1)

Scholars discussed releasing their data for use by other scholars. However, it was not always clear where the data could be archived and who would use them. Also, making the data freely available was seen as laborious when this necessitates formatting the material such that others could use it, while some scholars considered the body of data they had collected too small for archiving. While interviewees mentioned Zenodo [1], a Finnish data archive for the social sciences [2], Language Bank [3] and GitHub [4] as familiar data archives, it was clear that open research data has not yet become an established element of the research process in the field of history research.

> I guess that isn't so common yet, but in digital humanities projects it is becoming so [. . .]. You already see in international journals citations to open data. (PC4)

Discussion

Our research focused on what kinds of information interactions are involved in the activities related to using historical newspapers for primary research sources in research tasks. The task-based information interaction framework, designed for evaluating the entire cycle of task-relevant information interactions (Järvelin *et al.*, 2015), helped us evaluate the interactions with the historical-newspaper collection. Our examination of research tasks in which the digital newspaper collection was used for acquiring primary research sources revealed that interactions with the collection varied across the activities. Table 3 summarises the findings by showing the various activities and the information interactions that occurred during them.

Task-planning is a meta-activity that occurs throughout the research process and intertwines with other activities. New research topics and ideas emerge during the search, selection and working with items as the historians are exposed to the digital materials. The research topic's creation and formulation of focus, which represent understanding and narrowing of the research topic (Vakkari, 2001), were particularly evident during searching in the early stages of the research process. During all the activities, interactions with the collection contributed to learning about the topic and the task at hand, as Vakkari (2016) has noted. The task planning activity was particularly intertwined with the searching activity, in which the historians discovered new content, to be used in future studies. This corroborates the findings of Sinn and Soares (2014). Furthermore, task-planning actions were visible in the selection activity when scholars sought an understanding of which research topics are suited to examination via the selected research data. The new digital information environment affords large-scale data-driven discovery of a kind that was not previously possible.

Researchers have characterised historians who use historical newspapers as "browsers" (Allen and Sieczkiewicz, 2010). However, the scale of browsing has been assumed to be due to the newspaper collections' print nature. Some have speculated as to whether historians would search more if this were more feasible. In our case, during their *searching* activity, the historians were retrieving newspaper articles/issues ("items"), and they also were searching for information content from their personal collections, in the activity related to *working with information items*. The two are distinct: the first is aimed at locating an item that contains the information needed, and the second entails searching

Task planning	Searching	Selecting	Working with items	Synthetizing and reporting
Data-driven discovery	Creating personal corpus:	Cutting out useful articles from the pages	Close reading	Writing articles
Orienting to the newspaper contents	by browsing (PDF)	Downloading metadata	Analysing metadata	Writing monographs
Discussing the contents with research group	by querying (OCR): *Keywords Filtering Proximity search Backtracking search*	Creating a corpus	Quantifying data	Oral talks
Consulting earlier knowledge and secondary sources	*Forward tracking: keeping list of possible queries future*	Skimming and reading the articles (PDF/OCR)	Data preparation	Data visualisation
Acquiring access to data		Refining their own collections	Searching *within* personal corpus for analyses: *Annotating Using named entity recognition Evaluating search Using corpus linguistics Using via own API*	Writing alone
Deciding if server space is required			Data visualisation	Co-authoring Opening the data
Discovering new topics for research				

Table 3. Information interactions during the TBII activities

from the information contents directly. The former, locating items, may precede the latter, but the boundaries of individual documents as conceptual entities are stretched in today's digital information environments. The whole collection – with all of its information content – is directly searchable. The digital newspaper collection that our study focused on accelerates progress with this by the way it arranges the documents. At the same time, most of the historians in our study needed to "cut out" newspaper text on article level, whereas the newspapers are digitised as pages of the original newspapers, not as articles. Hence, this stage included large amounts of manual processing of the materials, although finding the relevant articles already was handled partly through searching the contents directly. Still, the varying quality of the OCRed text made searching difficult because the matching of strings was not straightforward.

Selecting information items required skimming and assessing the relevance of the articles found. The process of selection is closely connected with the search activity, and they were described as occurring simultaneously. The selection made the search goals visible: in most cases; the aim was to create a sub-collection for a particular topic to analyse further. This is akin to the activities described by Huistra and Mellink (2016). It entails recall-oriented searching and only rarely (e.g. in fact-finding-type tasks) precision-oriented searching. This activity includes content-based filtering that utilises both the content's OCR-rendered text and metadata.

We found data preparation to be an inherent part of the activity of *working with items*. This is consistent with the findings of Given and Wilson (2018). Hoekstra and Koolen (2019), who found "data preparation" too broad a term, deemed this to include several decisions that can be called "data scopes". There were few mentions of data preparation in our data. The methods selected by the scholars may have influenced this activity. Also, we collected *ex post* accounts of the behaviours, and the activities recalled in the interviews may have been characterised as more straightforward than they were during the work. Working with items was among the activities detected as involving collaboration. This activity could be

augmented with digital collaboration-oriented workspaces, which were not available to the scholars. The services provided were designed for individual users and were more suitable for researchers working alone.

As for the *synthesis and reporting* activity, the historians described their research outputs in light of the findings and how the digital newspapers were used in creating these. For the most part, this activity did not require direct interactions with the original newspaper collection, but indirect use was evident in the derivative collections and analysis. Visualisation served as a means of analysis and sense-making but also as an excellent method for showing the findings in oral presentations at conferences. The presentations and sharing of research findings are preceded in the process by analysis and processing of the research data. For large collections, computational and quantitative methods are used alongside more interpretive qualitative analysis. It is noteworthy that articles on history research seldom report or discuss the research processes and the decisions made during the data analysis in depth (Hoekstra and Koolen, 2019). Nevertheless, insightful interpretations do not arise solely from the newspaper content; these decisions contribute to them. Therefore, it is important to describe both the methods and the modelling processes affecting the outcomes of the research's processes. Finally, the historians interviewed actively made their research data available, but the research infrastructure in place does not seem to support doing so in a sustainable way.

A change in research culture, leading toward multidisciplinary ways of working, was obvious. The history scholars referred several times to a change in the work culture and to an increasing resemblance to the natural sciences. As in digital humanities (see Wang, 2018) publishing practices in the history field have grown more collaboration-based. In addition, authorship practices show a shift toward including people with other roles, such as programmers with their contributions (see also Bradley, 2012). However, increasing collaboration is not just due to the digitisation but other drivers such as research assessment may foster collaboration (Hockey, 2012). Changes in the field's research methods were discussed, with a mix of quantitative statistical and qualitative interpretive methods being typical.

The literature features much discussion surrounding digitisation and its effects on history research. Digitisation is making access easier and less expensive, eliminating much of the need for travel to archives, but at the same time there have been fears that this digital progress is going to compromise research quality by making the researchers "lazy" through encouraging the use of low-quality sources (Holm *et al.*, 2015). Developments may necessitate cross-checking and validation of findings against the original sources, but none of the interviewees spoke of reading print newspapers. This is in sharp contrast to the finding of Sinn and Soares (2014) that researchers rely heavily on these.

Our work attests that the improved access to newspaper materials is considered more important that the downsides. Digitisation is meant not purely for preservation but also for increased accessibility (cf. Jarlbrink and Snickers, 2017). Both preservation (to ensure quality) and access (to ensure use) are important, but they are two sides of the same coin. All materials that are preserved are preserved to be used later. Therefore, access is important and both sides should be considered if one wishes to guarantee reliable and useable digital heritage collections. Thus, when developing the collections, the needs of different types of users should be taken into account (Hughes, 2012a). Users desire information about the selection criteria that shape the collection (see also Hauswedell *et al.*, 2020). Opening the policies of the archives may help users to develop historically reflexive media literacy skills (Mussell, 2012).

The theoretical model we selected seems highly suitable for examining data such as ours and assists in organising the data, but at the same time it does limit analysis somewhat. In some cases, the scholars' various analysis activities were hard to distinguish from each other, on account of their intertwined and simultaneous nature, a characteristic more prominent in the humanities' history domain than in the sciences (Koolen *et al.*, 2020). Search in particular, while a

separate activity in the model, was present in several of the other activities. Furthermore, we could distinguish between two ways of searching: locating information items (by means of both metadata and content elements) and going through the content (primarily the text itself, from OCR output). One way of integrating the two was content-browsing that encompassed items located prior to the browsing. This embedded activity that occurred in several activities calls for revising the overall model – when search is a separate activity at the same level as all the other activities in the model, analysis is complicated by the activities' simultaneous and embedded nature. The model could be refined by including the embedded sub-activities ("activity in another activity") and some way to delineate the simultaneity.

Further, the model itself was developed to analyse uses of multiple secondary sources while we applied it to use situations with a historical newspaper collection that was used as a primary source. The model allowed us to design and cover the whole research cycle, but it focuses heavily on interaction and is emphasising the searching and selecting information items. The intertwined and multi-layered nature of the activities made the analysis difficult and the model lacked power in analysis of contextual aspects. Therefore, adding various contextual layers as explicit structures in the model would be useful.

Conclusions

Our investigation of academic historians' use of a digital historical-newspaper collection and of the information interactions involved in their research-process activities filled an important research gap by both analysing in-depth interviews with history scholars who used digitised historical newspapers as primary research sources and exploiting the TBII model as a tool for the data's collection and analysis. Among the key findings are that the purposes behind using digitised historical newspapers vary and that there are numerous information interactions within activities related to task-planning, search, selection, working with items and synthesis and reporting. There are important implications of the finding that information search occurs during various activities (where these overlap, since the history-research process is not linear). Our work highlights its centrality to collection and analysis of material in such use of digital collections.

The study points to several new research topics. One of the most striking phenomena for examination is related to how the use of digital collections has shaped the way scholars work, with sole-authorship-based work becoming transformed, in a shift toward more collaborative efforts. With work often being conducted in multidisciplinary research teams, there is a clear need for greater knowledge of how the multidisciplinary collaboration unfolds in digital history work and how information technologies might best support this collaboration. Likewise, little is known about the barriers scholars face and how they overcome these when using digital collections. Most of the studies on the research use of digitised newspapers have been dominated by studies of English language resources. This article contributes offering results of the Finnish newspaper collection and in the future more research on other language areas is needed.

In the course of our study of the newspaper collection's use in relation to primary sources, it emerged that the collection often gets used for many other purposes, such as teaching and fact checks. While we call for in-depth study of these, for richer understanding of the various uses of the numerous digital collections at scholars' disposal, we have already found evidence that information interactions vary across the activities involved. The finding implies that awareness of this would improve endeavours to design better information tools for history scholars.

Notes

1. https://zenodo.org/
2. https://www.fsd.tuni.fi/en/

3. https://www.kielipankki.fi/language-bank/
4. https://github.com/

References

Allen, R.B. and Sieczkiewicz, R. (2010), "How historians use historical newspapers", *Proceedings of the American Society for Information Science and Technology*, Vol. 47 No. 1, pp. 1-4, doi: 10.1002/meet.14504701131.

Anderson, I.G. (2004), "Are you being served? Historians and the search for primary sources", *Archivaria*, Vol. 58, pp. 81-129, available at: https://archivaria.ca/index.php/archivaria/article/view/12479.

Baruchson–Arbib, S. and Bronstein, J. (2007), "Humanists as information users in the digital age: the case of Jewish studies scholars in Israel", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 14, pp. 2269-2279, doi: 10.1002/asi.20699.

Bradley, J. (2012), "No job for techies: technical contributions to research in the digital humanities", in Deegan, M. and McCarty, W. (Eds), *Collaborative Research in the Digital Humanities, a Volume in Honour of Harold Short, on the Occasion of His 65th Birthday and His Retirement*, Ashgate publishing, Farnham, pp. 11-26.

Bulger, M., Meyer, E., De la Flor, G., Terras, M., Wyatt, S., Jirotka, M., Eccles, K. and Madsen McCarthy, C. (2011), "Reinventing research? Information practices in the humanities", *A Research Information Network Report*. doi: 10.2139/ssrn.1859267.

Burton, O.V. (2005), "American digital history", *Social Science Computer Review*, Vol. 23 No. 2, pp. 206-220, doi: 10.1177/0894439304273317.

Carr, N. (2010), *The Shallows: How the Internet Is Changing the Way We Think, Read and Remember*, Atlantic Books, London.

Chardonnens, A., Rizza, E., Coeckelbergs, M. and van Hooland, S. (2018), "Mining user queries with information extraction methods and linked data", *Journal of Documentation*, Vol. 74 No. 5, pp. 936-950, doi: 10.1108/JD-09-2017-0133.

Clement, T.E. and Carter, D. (2017), "Connecting theory and practice in digital humanities information work", *Journal of the Association for Information Science and Technology*, Vol. 68 No. 6, pp. 1385-1396, doi: 10.1002/asi.23732.

Conway, P. (2015), "Digital transformations and the archival nature of surrogates", *Archival Science*, Vol. 15 No. 1, pp. 51-69.

Crymble, A. (2021), *Technology and the Historian: Transformations in the Digital Age*, University of Illinois Press, Champaign.

Duff, W.M. and Johnson, C.A. (2002), "Accidentally found on purpose: information-seeking behavior of historians in archives", *The Library Quarterly*, Vol. 72 No. 4, pp. 472-496.

Ehrmann, M., Bunout, E. and Düring, M. (2019), "Historical newspaper user interfaces: a review", *IFLA WLIC 2019 - Athens, Greece - Libraries: Dialogue for Change*, available at: http://library.ifla.org/2578/.

Fidel, R. (2012), *Human Information Interaction: an Ecological Approach to Information Behavior*, MIT Press, Cambridge.

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

Given, L.M. and Willson, R. (2018), "Information technology and the humanities scholar: documenting digital research practices", *Journal of the Association for Information Science and Technology*, Vol. 69 No. 6, pp. 807-819, doi: 10.1002/asi.24008.

Gooding, P. (2016a), "Exploring the information behaviour of users of Welsh Newspapers Online through web log analysis", *Journal of Documentation*, Vol. 72 No. 2, pp. 232-246, doi: 10.1108/JD-10-2014-0149.

Gooding, P. (2016b), *Historic Newspapers in the Digital Age: Search All about it!*, Routledge, London.

Gregory, I. (2014), "Challenges and opportunities for digital history", *Frontiers in Digital Humanities*, Vol. 1 No. 1, doi: 10.3389/fdigh.2014.00001.

Hauswedell, T., Nyhan, J., Beals, M.H., Terras, M. and Bell, E. (2020), "Of global reach yet of situated contexts: an examination of the implicit and explicit selection criteria that shape digital archives of historical newspapers", *Archival Science*, Vol. 20, pp. 139-165, doi: 10.1007/s10502-020-09332-1.

Hockey, S. (2012), "Digital humanities in the age of the internet: reaching out to other communities", in Deegan, M. and McCarty, W. (Eds), *Collaborative Research in the Digital Humanities, a Volume in Honour of Harold Short, on the Occasion of His 65th Birthday and His Retirement*, Ashgate Publishing, Farnham, pp. 81-92.

Hoekstra, R. and Koolen, M. (2019), "Data scopes for digital history research", *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, Vol. 52 No. 2, pp. 79-94, doi: 10.1080/01615440.2018.1484676.

Holm, P., Jarrick, A. and Scott, D. (2015), *Humanities World Report 2015*, Palgrave Macmillan, Hampshire.

Hughes, L.M. (Ed.), (2012a), *Evaluating and Measuring the Value, Use and Impact of Digital Collections*, Facet Publishing.

Hughes, L.M. (Ed.) (2012b), "Using ICT methods and tools in arts and humanities research", in Hughes (Ed.), *Evaluating and Measuring the Value, Use and Impact of Digital Collections*, Facet Publishing, London, pp. 123-134.

Huistra, H. and Mellink, B. (2016), "Phrasing history: selecting sources in digital repositories", *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, Vol. 49 No. 4, pp. 220-229, doi: 10.1080/01615440.2016.1205964.

Järvelin, K., Vakkari, P., Arvola, P., Baskaya, F., Järvelin, A., Kekäläinen, J., Keskustalo, H., Kumpulainen, S., Saastamoinen, M., Savolainen, R. and Sormunen, E. (2015), "Task-based information interaction evaluation: the viewpoint of program theory", *ACM Transactions on Information Systems (TOIS)*, Vol. 33 No. 1, pp. 1-30, doi: 10.1145/2699660.

Jarlbrink, J. and Snickars, P. (2017), "Cultural heritage as digital noise: nineteenth century newspapers in the digital archive", *Journal of Documentation*, Vol. 73 No. 6, pp. 1228-1243, doi: 10.1108/JD-09-2016-0106.

Kettunen, K. and Pääkkönen, T. (2016), "Measuring lexical quality of a historical Finnish newspaper collection–analysis of garbled OCR data with basic language technology tools and means", *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 956-961, available at: https://www.aclweb.org/anthology/L16-1152.pdf.

Kettunen, K., Honkela, T., Lindén, K., Kauppinen, P., Pääkkönen, T. and Kervinen, J. (2014), "Analyzing and improving the quality of a historical news collection using language technology and statistical machine learning methods", *IFLA World Library and Information Congress Proceedings 80th IFLA General Conference and Assembly*, IFLA, available at: http://hdl.handle.net/10138/136269.

Koolen, M., Kumpulainen, S. and Melgar-Estrada, L. (2020), "A workflow analysis perspective to scholarly research tasks", *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pp. 183-192, doi: 10.1145/3343413.3377969.

Kumpulainen, S. (2017), "Task-based information searching: research methods", *Encyclopedia of Library and Information Sciences*, CRC Press, Boca Raton, pp. 4526-4536.

Kumpulainen, S., Keskustalo, H., Zhang, B. and Stefanidis, K. (2020), "Historical reasoning in authentic research tasks: mapping cognitive and document spaces", *Journal of the Association for Information Science and Technology*, Vol. 71 No. 2, pp. 230-241, doi: 10.1002/asi.24216.

Late, E., Tenopir, C., Talja, S. and Christian, L. (2019), "Reading practices in scholarly work: from articles and books to blogs", *Journal of Documentation*, Vol. 75 No. 3, pp. 478-499, doi: 10.1108/JD-11-2018-0178.

Marcella, R., Rowlands, H. and Baxter, G. (2013), "The critical incident technique as a tool for gathering data as part of a qualitative study of information seeking behaviour", in Mendy and Geringer (Eds), *Leading Issues in Business Research Methods*, Academic Conferences and Publishing, Vol. 2.

Meyer, E., Eccles, K., Thelwall, M. and Madsen, C. (2009), "Final report to JISC on the usage and impact study of JISC-funded phase 1 digitisation projects and the toolkit for the impact of digitised scholarly resources (TIDSR)", available at: https://ora.ox.ac.uk/objects/uuid:a2dcf9d1-89ed-44c0-b47e-2523c1fbb704.

Milligan, I. (2013), "Illusionary order: online databases, optical character recognition, and Canadian history, 1997-2010", *Canadian Historical Review*, Vol. 94 No. 4, pp. 540-569, doi: 10.3138/chr.694.

Mussell, J. (2012), *The Nineteenth-Century Press in the Digital Age*, Palgrave Macmillan, London.

Nanetti, A. and Cheong, S.A. (2018), "Computational history: from big data to big simulations", in Chen (Ed.), *Big Data in Computational Social Science and Humanities. Computational Social Sciences*, Springer, Cham. doi: 10.1007/978-3-319-95465-3_18.

Oberbichler, S., Hechl, S., Klaus, B., Kaukonen, M., Pääkkönen, T. and Ansel, M. (2019), "Online research of digital newspapers of three national libraries: a survey", available at: https://www.newseye.eu/fi/blogi/news/online-research-of-digital-newspapers-of-three-national-libraries-a-survey-by-sarah-oberbichler-stef/.

Sinn, D. and Soares, N. (2014), "Historians' use of digital archival collections: the web, historical scholarship, and archival research", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 9, pp. 1794-1809, doi: 10.1002/asi.23091.

Strauss, A. and Corbin, J.M. (1997), *Grounded Theory in Practice*, Sage.

Tibbo, H.R. (2002), "Primarily history: historians and the search for primary source materials", *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pp. 1-10, doi: 10.1145/544220.544222.

Toms, E.G. (2011), "Task-based information searching and retrieval", in Ruthven and Kelly (Eds), *Interactive Information Seeking, Behaviour and Retrieval*, Facet Publishing, London, pp. 43-75.

Toms, E.G. and O'Brien, H.L. (2008), "Understanding the information and communication technology needs of the e-humanist", *Journal of Documentation*, Vol. 64 No. 1, pp. 102-130, doi: 10.1108/00220410810844178.

Vakkari, P. (2001), "A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study", *Journal of Documentation*, Vol. 57 No. 1, pp. 44-60, doi: 10.1108/EUM0000000007075.

Vakkari, P. (2016), "Searching as learning: a systematization based on literature", *Journal of Information Science*, Vol. 42 No. 1, pp. 7-18, doi: 10.1177/0165551515615833.

Wang, Q. (2018), "Distribution features and intellectual structures of digital humanities: a bibliometric analysis", *Journal of Documentation*, Vol. 74 No. 1, pp. 223-246, doi: 10.1108/JD-05-2017-0076.

Warwick, C., Galina, I., Terras, M., Huntington, P. and Pappa, N. (2008), "The master builders: LAIRAH research on good practice in the construction of digital humanities projects", *Literary and Linguistic Computing*, Vol. 23 No. 3, pp. 383-396, doi: 10.1093/llc/fqn017.

Corresponding author
Elina Late can be contacted at: elina.late@tuni.fi