Received: 3 June 2021 | Revised: 5 November 2021 | Accepted: 11 November 2021

DOI: 10.1002/asi.24608

RESEARCH ARTICLE

JASIST WILEY

Struggling with digitized historical newspapers: Contextual barriers to information interaction in history research activities

Sanna Kumpulainen | **Elina Late**

Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland

Correspondence
Sanna Kumpulainen, Faculty of Information Technology and Communication Sciences, Tampere University, Kalevantie 4, FI33014 Tampere, Finland.
Email: sanna.kumpulainen@tuni.fi

Funding information
Academy of Finland, Grant/Award Number: 326616

Abstract

On account of the complexities related to the use of digitized newspapers, researchers may encounter barriers when interacting with the collections' content. Overcoming barriers that could influence their information interaction should enhance the accessibility and utility of the newspapers. Hence, the study examined the barriers faced in history-research tasks involving interaction with digitized historical newspapers, with focus on the barriers' contexts and the related task-based activities. The analysis employed two datasets, from in-depth interviews and demonstrations of newspaper-use situations. Content analysis from these complementary data showed that barriers arose in multiple contexts, connected with the collection, task, tools, and socio-organizational setting. Most barriers were associated with collection context and occurred in information searching and selection activities and in working with information items. Barriers related to the task or to socio-organizational context arose most often in the planning and monitoring activities and in synthesizing and reporting. Such research-based insight into the barriers faced can aid in illuminating what is required for providing good support to researchers working with digital newspaper content.

1 | INTRODUCTION

The increasing number of digital sources and continuing development in technological tools available for humanities research have together changed the ways in which history scholars work (Given & Willson, 2018; Toms & O'Brien, 2008). Scholars are challenged to learn new skills, collaborate with researchers representing other disciplines (Late & Kumpulainen, 2021), and find new ways to present and publish research results (Clement & Carter, 2017). This change has come at a price, however. Scholars have reported problems with the quality of digitized content, such as poor optical character recognition (OCR) quality (Jarlbrink & Snickars, 2017; Terras et al., 2018) and technical and other problems with the interfaces and tools (Martin-Rodilla & Sánchez, 2020).

To shed light on this landscape, we investigated the information interaction connected with digitized historical newspapers in today's changing work environment. The first of the three main issues is that, although digitization has greatly facilitated ready availability of historical newspapers, their information value is greatly diminished when proper access to the content is lacking. Since digital services are useless if one cannot access the contents, it is vital to develop a better understanding of the barriers to interacting with the information—the obstacles that hinder

the interaction process or even bring it to a halt. Such knowledge can assist in avoiding certain system-design pitfalls, bringing fuller understanding of the reality of system use, and affording information-system evaluation that is better anchored in real-world considerations.

Second, information content is never used in isolation. The use process unfolds during some purposeful action, in this case history research and associated work tasks. Here, the research tasks are purposeful processes targeted at a particular outcome via active use of the content of the newspaper collections. Accordingly, we designed a research setting that follows a framework for evaluating task-based information interaction (Järvelin et al., 2015) in terms of the whole work-task process and all its activities. Conceptualizing information interaction as encompassing information needs, seeking and retrieval, and use behaviors provides analytical power for examining the information activities entailed by research tasks.

Third, many of information science's theories and models apply highly abstract and vague conceptualizations of the context. The aforementioned framework that provided our starting point is no exception. Though there is general agreement that contexts constitute a frame of reference for information-related behavior, information-science scholarship has revealed little of how such a frame is established in practice (Courtright, 2007). We sought to contribute to addressing this issue.

Our study examined barriers in relation to four distinct cognitive activities and embedded contexts, where the former are the actions carried out to reach the goals for the work task (Järvelin et al., 2015) and contexts are defined as the layers of abstraction within which the barriers are encountered (Byström & Kumpulainen, 2020). Because information interaction does not exist in a vacuum, one must account for the contextual origin of information items (Monte-Sano & De La Paz, 2012).

At the heart of our study were in-depth interviews with history scholars and their demonstrations of using digital collections of historical Finnish newspapers as primary sources. With this research setting, we set out to answer the following central questions:

1. In which contexts are the barriers encountered?
2. In which activities during the task-based information interaction do the barriers arise?

2 | BACKGROUND

2.1 | Historical newspapers as research data in digital history

With the research domain of Finnish historical newspapers collected in digital form, we examined how historians work with the contents. Both born-digital and digitized physical collections are essential to digital history research (Kachaluba et al., 2014; Sinn & Soares, 2014), which examines the past by applying new communication technologies and experiments with computational methods for analysis, production, and dissemination of historical knowledge (Salmi, 2021). Since the turn of the millennium, public and private organizations around the world have digitized historical newspapers (Gooding, 2016). This has influenced historians' work in at least two ways: most of the collections are ubiquitous in their availability, granting historians access from their own devices at any site, and, second, many user interfaces serving as a gateway to digital newspaper collections offer advanced computation-based techniques for searching and analyzing their contents. In addition to basic full-text and keyword-based search permitted by the standard OCR technique, the typical set of functions includes provisions for metadata, browsing, and filtering of results. The most advanced interfaces provide greater user-interaction functionality (e.g., saving of articles to "Favorites"), content enrichment (e.g., post-OCR correction), connectivity (e.g., links to other repositories), and application programming interfaces (APIs) for programmatic enrichment (Ehrmann et al., 2019). One problem commonly plaguing digitized newspaper collections is low OCR quality. For Finland specifically, one study of newspaper data found the range of OCR accuracy to be between 50 and 70% on word level and 71–98% at character level (Kettunen & Pääkkönen, 2016). Jarlbrink and Snickars (2017) even characterized the problem thus: in combination, misinterpreted words (OCR errors) and "random" text created by auto-segmentation tools create new newspaper content that was never actually written. They described this machine-interpreted text never printed in the original papers as thereby entering the historical record. Hence, researchers may not be justified in fully trusting the textual representations.

Not many studies focus on the use of digital newspapers in history research. One exception is Late and Kumpulainen's (2021) research into the information interaction that took place when Finnish historians used historical-newspaper collections as primary sources in their research. These historians, often part of multidisciplinary research groups, examined such development as evolution of newspapers, historical use of language, the history of a certain societal phenomenon, or individuals' history by utilizing both qualitative and quantitative methods. Information search was shown to be crucial to various of their activities and to constitute an essential research method for data-collection and analysis work involving the digital collection. The study showed also that manual processing, such as browsing of the materials, was often necessitated

because inconsistency of OCR quality made searching difficult. The literature highlights also that, while many historians are interested in specific types of content, such as only advertisements, editorials, or news articles (see also Allen & Sieczkiewicz, 2010), discrimination among these was difficult since digitization had been performed at page level, not for individual pieces in the original newspapers. The literature has also looked at the output of the process, publication of the results in multiple disciplines' venues (Late & Kumpulainen, 2021). The results' graphical presentation for publication in scholarly journals was especially important for scholars.

2.2 | Activities and interacting with primary sources

Rather than taking place in isolation, information interaction is derived from a larger, motivating aim—such as a work task. Scholarly work comprises several sub-tasks, referred to as primitives (Palmer et al., 2009; Unsworth, 2000), data-scopes (Hoekstra & Koolen, 2019), or activities (Järvelin et al., 2015; Palmer et al., 2009) in accordance with the taxonomist's focus. Unsworth (2000) proposed that scholarly endeavors comprise the seven primitives of discovering, annotating, comparing, referring, sampling, illustrating, and representing, while Palmer et al. (2009) identified five activities: searching, collecting, reading, writing, and collaborating (along with a separate cross-cutting one). Neither model, however, covers the task itself or how it evolves during the task performance, both of which are important factors in human task performance (Byström & Kumpulainen, 2020; Vakkari, 2001).

One model covering the larger task itself is the task-based information-interaction model (TBII) proposed by Järvelin et al. (2015). Under its evaluative lens, the five activities in a learning task consist of task planning and reflective assessment, information searching, selecting information items, working with information items, and synthesizing and reporting accordingly. The first activity in this list is a meta-cognitive activity defining the perceived task. Intertwined with other activities in the process, it changes dynamically as task performance progresses, thereby leading to a possibly more focused understanding of the task. The searching activity includes all interactions with a search system and the information retrieval, leading up to the third activity: selection of the suitable information items. This includes all decision-making about the relevance or usefulness of the items found. The "working with items" activity involves the interactions with the document content, such as scanning and browsing, reading, and annotating the information items. Finally, synthesizing and reporting is associated with the task outputs and outcomes. This activity requires the writer to integrate information from several information sources to create new knowledge and develop new information items and objects.

3 | CONTEXTS WITHIN INFORMATION-INTERACTION RESEARCH

Scholars agree that context is crucial to studies of information interaction and constitutes a frame of reference for information-related behavior (e.g., Huvila, 2019; Kuhlthau, 1991; Vakkari, 2001). Contexts of interactive information retrieval are recognized as multiple overlapping elements that inform, direct, or shape the interactions (Cool & Spink, 2002). Beyond that, however, there is little understanding of how context is best defined and how the research should address it (Huvila, 2019).

Context constitutes background to something the researcher wishes to understand and explain. In practice, information-interaction studies usually refer to any factors or variables that appear to affect individuals' information-seeking behavior: socio-economic conditions, work roles, tasks, problematic situations, communities and organizations with their structures and cultures, etc. (Talja et al., 1999). Among the contextual factors that Courtright (2007) has cited as studied for their role in shaping information-linked practices are rules, resources and culture, social aspects (such as social networks or social capital), tasks, problems, and situations. According to her, numerous studies regard tasks as the primary contextual shaping influence behind variability in information activities.

Very few task-based studies apply context-oriented frameworks, though. One of them is work in which Kumpulainen and Järvelin (2012) conceptualized barriers to research work as occupying several, nested layers of context: (1) a level reflecting organizational culture, social norms, the organizations' resources, and constraints; (2) within the organization context, a level at which one can examine the factors that affect the work tasks; (3) the integration context, in which the information systems are selected—from among diverse information sources and systems that could be used—in accordance with the actor's preferences/decisions and then used in a concerted manner; and (4) the system context, in which information retrieval and processing tasks take place.

Similarly, Byström and Kumpulainen (2020) developed a model in which the overlapping contexts that influence the task-based information needs are depicted as a construct of vertical relationships in which related information needs are inscribed. According to them, both

the vertical relations and horizontal ones composed of sequences of information-interaction activities inform and guide the information needs. This model acknowledges the context of professions and the workplace, the work task, successive seeking tasks, and a series of search tasks. The last two can be regarded as specific types of information-interaction activities.

3.1 | Barriers to information interaction

Barriers affect seeking, searching, and use of information by hindering information-interaction processes (Kumpulainen & Järvelin, 2012) or impinging on the desired access to information, thereby producing negative effects (Savolainen, 2016). While information-seeking research often explores barriers, common terminology in the domain of interactive information retrieval is lacking. Scholarship typically focuses on problems connected with search formulation, skills, levels of domain knowledge (Wildemuth, 2004) and the search topic (Kelly & Cool, 2002), and mismatches between the user's mental models and the model behind the system (Borgman, 1985).

Kumpulainen and Järvelin (2012) presented an analysis frame intended specifically for studying barriers occupying the system, data-integration, and work-task contexts across various task-complexity levels. In the results, the number of barriers increases with task complexity and complex-task sessions encounter barriers of a more conceptual nature while technological barriers are the most prominent in simple-task settings. The research design was limited to task-level barriers, with socio-organizational ones remaining beyond its scope; however, it is clear that, from a wider, socio-cultural perspective, there exist various relevant barriers related to language skills, social stigma or cultural taboos, so-called small worlds, institution/organization types, or deficits in social and economic capital (Savolainen, 2016).

Importantly, collaboration may impose barriers to information interaction (Aagaard-Hansen, 2007). Many tasks present in scientific collaboration display a high degree of uncertainty, and trial and error is integral to the process (Latour, 1987). An extensive literature review enabled Sonnenwald (2008) to enumerate such problems with collaboration as specifying the goals for the research, articulating realistic tasks, and establishing timeframes. In cross-disciplinary collaboration especially, language, epistemology, and communication-practice differences can hinder collaboration. For instance, technology has been adopted differently in different disciplines (Late et al., 2019). Via efforts to overcome obstacles arising from differences, collaboration can inspire technological innovation.

4 | RESEARCH SETTING

4.1 | Theoretical framework

The TBII model formed the backbone for our theoretical framework. Adapting it as Figure 1 illustrates, we combined its information search and selection activities (cf. Korkeamäki & Kumpulainen, 2019), and we renamed "task-planning and reflective assessment" as "planning and monitoring," for alignment with our focus on research behaviors instead of the original learning-process emphasis. While depicted as phases, these activities are set in often iterative processes that loop back upon themselves in various, complex ways. Additionally, we confined our focus to the activities, excluding the outputs and outcomes from the model, and we reframed its input side. Whereas the original model identifies context as input to the process (without further differentiation), we regarded context as instead surrounding the process and as manifesting itself in multiple ways as the activities proceed. With these underpinnings, we sought to provide insight as to the contexts influencing the information interactions in various activities.

4.2 | Data collection

Our research examined use of a historical-newspaper collection held by the National Library of Finland (NLF), where newspapers published in Finland from 1771 to 1929 are available in digitized form to scholars and all citizens. This collection, containing approximately 7.4 million newspaper pages, in the Finnish, Swedish, and Russian languages, is accessible via a Web-based service (digi.kansalliskirjasto.fi), with the content presented in text format (via OCR) and as PDF files of pictures of the pages. The online service provides tools for search, gathering, and analysis of newspaper data. Some advanced functionality is available. For example, users can copy selected data from the collection, paste to a "scrapbook," and later download the selected data in MS Excel files. All content of the historical-newspaper collection is available also via the Language Bank of Finland, formatted as downloadable text files (Kettunen & Pääkkönen, 2016).

We gathered the interview data in 2018 and 2020 via in-depth qualitative interviews with history scholars who were using the digitized newspaper collection for their research, and we complemented these data with the aforementioned demonstrations of their use of the newspapers for research data. We identified potential participants through a list of research projects that had utilized the NLF's collection, supplementing this source with a snowballing method: the scholars interviewed were asked

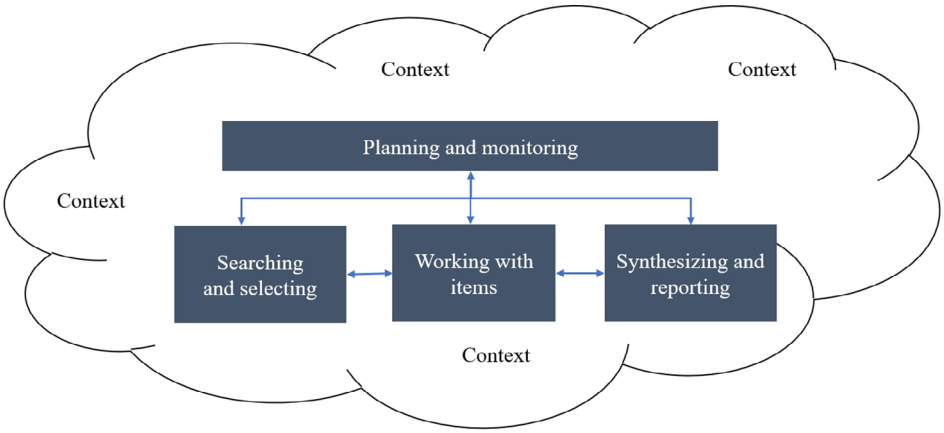

FIGURE 1 Our task-based information-interaction model, adapted from the work of Järvelin et al. (2015)

TABLE 1 Profile of the research participants

Attribute	Breakdown by number of interviews
Affiliation	University A: 5, university B: 4, university C: 4
Position	Full professor: 1, associate professor: 1, postdoctoral researcher: 8: doctoral student: 3
Research field	History of books: 1, conceptual history: 3, cultural history: 6, political history: 1, societal history: 2
Collaboration	Solo work: 6, research group: 7

about potential other interviewees. Our sampling gave us 13 interviewees, with a range of scholarly backgrounds (see Table 1). Triangulation involving three rounds of interviews, conducted by three researchers, aided in rising above the possible biases stemming from the individual researchers' personality (Kumpulainen, 2017). All researchers were experienced interviewers who were familiar with the digital newspaper collection. The first round of interviews (PA1–PA4) was conducted in spring 2018 face-to-face, while the second and third, in spring 2020 (PB1–PB5) and fall 2020 (PC1–PC4), respectively, were completed online by means of the videoconferencing tool Zoom. All interviews were conducted in Finnish. The average interview length was 60.8 min. The face-to-face interviews were audio-recorded and the online interviews video-recorded, and the full body of material (with a duration of 13 hr and 18 min) was transcribed for analysis. Data were collected until saturation was reached.

We should note that some data from the interviews were used in two previous studies, which had other purposes. The first of these, using material from three interviews, is Korkeamäki and Kumpulainen's (2019) research into the use of various sources, rather than interactions specifically with the newspaper collection. Second, Late and Kumpulainen (2021) identified the information interactions that feature in the activities, whereas the work reported upon here elaborates on the kinds of barriers the researchers faced during their interactions.

To guarantee coverage of the entire information-interaction process, we based the interviews' themes and list of questions on the task-based framework (Table 2 provides examples). The interviewees were e-mailed the general interview themes beforehand, at which time informed consent was obtained. Some interview questions solicited background information such as the interviewee's current work status, academic discipline, and research experience. Further questions applied a variation of the critical-incident technique, or CIT (Flanagan, 1954), to anchor the experiences to an ongoing or recent research task in which the interviewee had used the newspaper collection. The interview guide did not include direct questions about the barriers faced—as the interviewees described their information interactions, they brought up difficulties they experienced.

The demonstration portion of the data-gathering entailed asking interviewees to show their ways of using the newspapers for research data and share research outputs related to the tasks. This enabled increasing the accuracy of our data related to their work tasks. Face-to-face demonstrations were video-recorded, and online demonstrations' "shared screens" were recorded.

The interviews mirrored the scholars' recollections, so this method could reveal only those barriers that the interviewees remembered and articulated. Hence, demonstrations served as an excellent technique for facilitating their recall of particular phases of their projects and the choices they made in pursuit of their research goals. During the demonstrations, the interviewer could ask, for instance, why and how the action described was carried out. The resulting data also gave insight as to what the scholars wanted to do with the data but, on account of major obstacles (e.g., lack of access to key data), could not. This material yielded vital information for further

TABLE 2 Example interview questions related to the information interaction activities

Theoretical construct	Interview questions	Demonstration
Planning and monitoring	Where did you get the idea for the project? What are/were the goals for the project? What are/were your research questions? How were they formulated? Did any new criteria or insights for [the activity discussed] occur during the research? How was this manifested?	Not applicable
Searching and selecting	How did you become aware of the collection? How was searching for data in the collection handled? What problems did you face? How did you select the data? Where did you save the data, and in what format? How were the data organized?	Demonstrating the means of data search and selection
Working with items	How were the data processed and analyzed? Why? What kinds of research methods did you use? Why? Did you collect data from other sources?	Demonstrating the use of various tools for research
Synthesizing and reporting	What kind of results has the study produced? How and where did/will you present or publish the results? Have you considered providing open access to your data? How and where? If not, why?	Sharing the end products of the research

development of information services and for a better understanding of history scholars' research-related needs. Such knowledge cannot be obtained by such means as direct observations or log data alone.

4.3 | Data analysis

Content analyses were conducted with the ATLAS.ti software workbench and quantitative analyses via SPSS. The content analysis comprised iterative readings of the interview transcripts, open coding, and selective coding (Strauss & Corbin, 1997). All coding was done by two researchers; after each researcher completed separate coding, they discussed it to find consensus. We employed this procedure to avoid biased analyses.

The first step involved reading through the dataset several times, for familiarity with it, and open coding of the material. The first round of coding traced the barriers' expressions from the research data. In total, 202 instances of barriers were identified. Citations related to the barriers identified were imported from ATLAS.ti to SPSS for further coding and analyses. Links to the original data were retained in case reviewing them was required. Second, coding of the information-interaction activities was applied, per the typology in Figure 1. Through a mix of deductive and inductive coding, we articulated divergences from the TBII activities originally conceptualized. We identified the following activities: planning and monitoring, searching and selecting, working with items, and synthesizing and reporting.

Next, we pinpointed and categorized the contexts of the various barriers, using a modification of a set of overlapping contexts identified earlier (Byström & Kumpulainen, 2020; Kumpulainen & Järvelin, 2012). The initial coding of the contexts was deductively oriented, but we switched to inductive, data-driven analysis to alter some categories from the starting model. We chose interview extracts that illustrate each type of barrier in particular contexts and loosely translated them from Finnish to English. Fourth, we coded for individuals' vs. collaborative tasks since our data included tasks in both classes.

In the quantitative analyses, the activities and the barriers' context were cross-tabulated, to reveal the contexts wherein the barriers were encountered during specific phases of the research process. Chi-squared testing in SPSS assessed the categorical variables' mutual statistical independence.

5 | RESULTS

The presentation of results begins with the contexts analyzed and how the instances of barriers were categorized under these. Then, we address the distribution of barriers across the contexts and the TBII activities. These results emerged from the interviews and demonstrations.

5.1 | Contexts of barriers

We found, in all, 202 instances of barriers observed from the interview data. Figure 2 depicts our categorization of

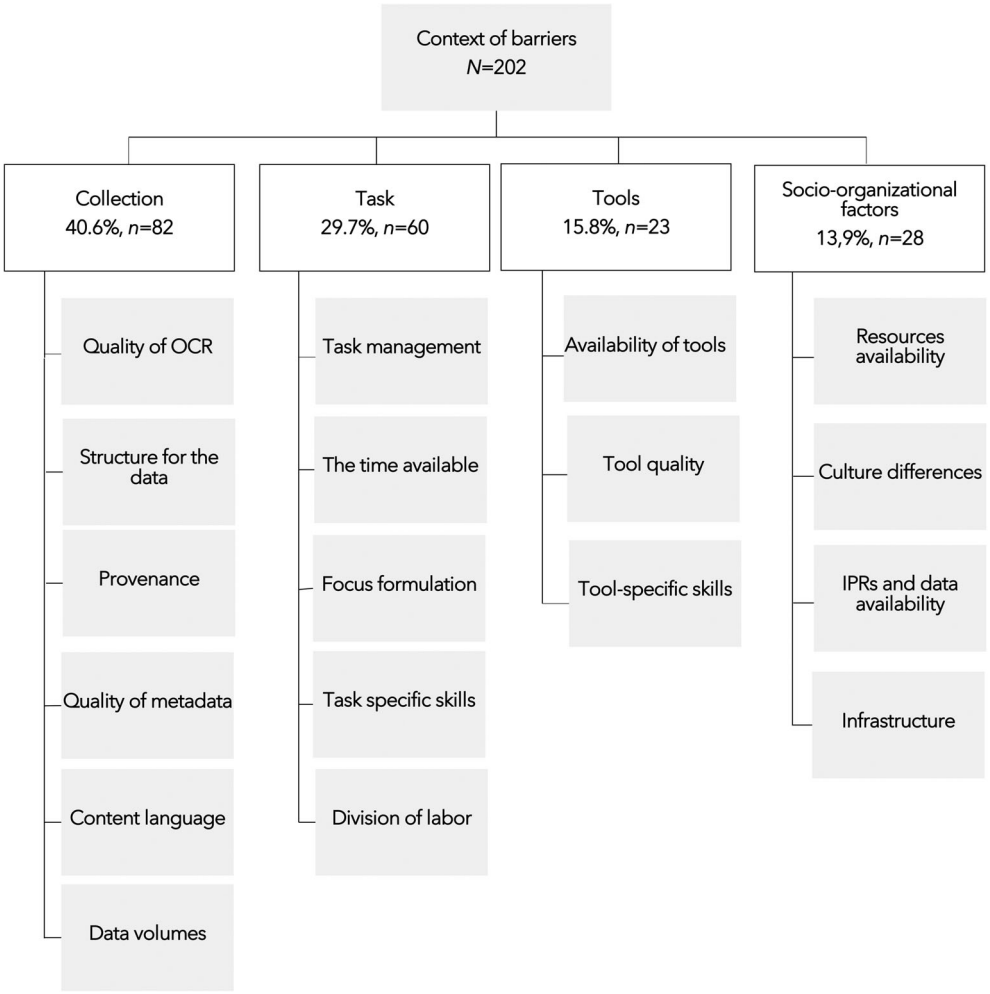

FIGURE 2 The barriers' contexts and the instances of barriers involved

these instances in accordance with the context with which they were associated.

The second most common class of barrier, accounting for 29.7% of instances, was "Task." The barriers faced in the task context were typically related to ways of working (e.g., lack of time or various issues with task management and division of labor) or to actor-internal elements (e.g., procrastination and averseness to continuing the task, poor focus related to tasks' goals, and lack of specific skills/procedural knowledge). One scholar studying the history of newspapers described the manual work that compiling specific types of articles from the collection demanded:

> There are some insecurities with the data, and a lot needs to be done manually. There is so much work, and it takes time (PB2).

Barriers in the context of tools (15.8% of instances) arose in relation to the diverse methods and procedures (some involving software applications), both internal and external to the collection-access Web service. Certain barriers to interaction with the newspaper data were directly tool-related: some tools did not function as desired (i.e., there were conceptual mismatches) or were ill suited to such a large dataset, with interfaces that did not support the task. Also, scholars pointed to barriers caused by their lack of skills in using technological tools, which develop rapidly. One scholar using the NLF interface for research and teaching described the need to maintain his skills thus:

> The development of the tools and applications is sometimes so sudden that I am not always on track. (PC4)

Barriers related to socio-organizational context were not reported as frequently, at 13.9% of instances. However, when they did appear, they could well derail the entire research process. Likewise, they might be rare in our data partly because they prevent some projects from even leaving the brainstorming stage. This category included lack of resources such as funding and other support (e.g., a suitable work space) from the organization,

a clash of research cultures in interdisciplinary settings, and problems wrought by transformation/changes in humanities research culture or the legislative environment (affecting access rights, interpretation of intellectual property rights connected with the collection, etc.). Lack of research infrastructure and of data access for humanities researchers were observed, and scholars emphasized being under-resourced relative to other disciplines. One digital-humanities scholar characterized the issue thus:

> Creating and supporting infrastructure for the humanities [...] there are practical things. If we did not collaborate with people from other disciplines, we wouldn't be able to operate. In essence, for many things we are on our own. (PC2)

5.2 | Barriers across the activities

Proceeding to identify the activities wherein the barriers were encountered, we cross-tabulated the barriers across contexts (see Table 3). A Pearson chi-squared test of independence afforded examining the relation between the activities and their context. That association was significant (X^2 [9, $N = 202$] = 77.464, $p < .001$).

Of the barriers, 27.2% ($n = 55$) of instances were expressed in planning and monitoring activity. In the course of these activities, scholars formulate their research setting and become familiar with the data sources. In this activity area, nearly half of the barriers (45.5%) arose in the task context. The scholars encountered these in task planning, procedure planning, and division of labor such as the collaboration arrangements within participants' research projects. One interviewee with a multidisciplinary research team described the importance of such arrangements thus:

> It would be important to be in the same room as the rest of the research group because sometimes the IT people can solve the problems easily if we just know to ask. (PB3)

Those planning-and-monitoring-related barriers arising in socio-organizational context (25.5% of instances) involved availability of data and infrastructure. In the "Collection" context, poor-quality OCR, the enormous body of data, lack of information about the collection, and data-availability issues created 20.0% of the barriers to planning and monitoring.

Next, the "searching and selecting" activity class was linked to almost a third of the barriers specified (28.7%, $n = 58$). Of the instances of barriers in these activities, related to interacting with the search system and to browsing and queries, two thirds (65.5%) appeared in the context of the collection, with OCR data quality being clearly the most common issue. The other barriers were visible mainly in the context of tools (24.1%). For instance, sometimes no tools were available for seeking and selecting the desired content, or using the tools supplied proved too complicated. A scholar working in the field of conceptual history considered the possibility of advanced searches:

> People from computer science have told me that I could work around [an issue] if I have API access, but my skills aren't good enough for that. (PA2)

Working with items was the activity group that occasioned the largest number of barriers (31.2%, $n = 63$). This activity class covers interactions with the contents of the documents, such as scanning and browsing, reading, and annotating the information items. Nearly half (47.6%) of the barriers in this class emerged in the collection context. When working with the items, scholars experienced barriers rooted in lack of knowledge about the contents of the collection, poor-quality OCR data, structure issues with the data (e.g., absence of article-level organization), and missing metadata.

TABLE 3 Percentages of the barriers' presence in TBII activities, by context (%, $p < .001$)

TBII activities					
Context	**Planning and monitoring**	**Searching and selecting**	**Working with items**	**Synthesizing and reporting**	**All**
Collection	20.0	65.5	47.6	11.3	40.6
Task	45.5	8.6	28.6	46.2	29.7
Tools	9.1	24.1	20.6	0	15.8
Soc-org.	25.5	1.7	3.2	42.3	13.9
Total	100.0 ($n = 55$)	100.0 ($n = 58$)	100.0 ($n = 63$)	100.0 ($n = 26$)	100.0 ($n = 202$)

> In an ideal world, I could analyze the results on article level, and they would contain more metadata. This would make it easier. (PB1)

More than a quarter (28.6%) of the barriers in this category showed links to the task context. These were related to such matters as excessive manual work and multidisciplinary research groups. The tool context occasioned a fifth of the barriers (20.6%). For instance, there were no high-quality tools that suited the research group's purposes, or the tools were hard to use. Also, scholars using computational methods cited losing their grasp of the provenance of the newspaper items.

Finally, least frequently reported were barriers in synthesizing and reporting activities (12.9%, $n = 26$). When characterizing this activity, the participants described how they produced their outputs and what hindered or otherwise affected their efforts, such as properties of the manuscripts or databases (in cases of data-sharing), the work load and available resources, cross-disciplinary work, and varying publication practices. Most barriers cited in this activity class were connected with the task context (46.3%), but almost as many were visible in the socio-organizational context (42.3%). One common barrier related to the former is length: the texts produced were too long for the target journals. As for socio-organizational barriers, these involved resources, publishing and disciplines' culture, and sharing of data. For instance, an experienced researcher described problems he had faced when attempting to publish digital-humanities research in traditional history journals:

> Some of our papers were desk rejected because the editors thought they were too much about the methods, although we disagreed. (PC2)

Scholars who worked alone ($n = 6$) articulated 36% ($n = 73$) of the instances of barriers, with members of a research group ($n = 7$) describing the remaining 64% ($n = 129$). These two groups differed in the barriers' breakdown by context class (X^2 [3, $N = 202$] = 9.739, $p = .021$). The prevalence of collection- and tool-related barriers was emphasized among those working alone. In contrast, barriers related to the task and socio-organizational issues were expressed more frequently in the data from members of collaborative groups. Multidisciplinary collaboration and finding a shared focus within the research group often posed difficulties. In some cases, the research group's dispersion over several time zones rendered collaboration even more difficult.

The distribution of the barriers by activity category too shows variation between solo and collaborative scholars (X^2 [3, $N = 202$] = 24.088, $p < .001$). The share of barriers related to planning and monitoring and to synthesizing and reporting were greater for scholars working in collaborative projects, while those working alone encountered barriers more in their searching and selection.

6 | DISCUSSION

The research showed that barriers appear at various levels of nested contexts and in all types of activities. The contextual analysis shed light on the origins of the barriers, thereby answering the first research question. Barriers connected with the collection context, including all barriers related to the contents of the historical-newspaper collection, were the most frequent in our data. These pertain to the digitized image surrogates of the newspapers, OCR-based textual data, and metadata connected with provenance and the original newspapers. This finding highlights the importance of the items' quality, genre, and information architecture as inputs to the research process (Monte-Sano & De La Paz, 2012). Also, as Leonelli (2019) pointed out, changes in data format can have a significant impact on who uses the data, when and where. It has even been argued that the "digital surrogates" do not, and cannot, replace original paper documents (Conway, 2015; Sinn & Soares, 2014) and that digitized collections can skew humanities research (Milligan, 2013). Cultural-heritage data such as historical newspapers are provided by organizations that are frequently beyond the influence of the humanities scholars who utilize them. The production and choice of the data that they use lie outside their control. Since history scholars' research process is often data-driven, what is available determines what can be studied and how (Late & Kumpulainen, 2021).

In the second context bucket, "Task," are barriers related to ways of working, the task process, and collaboration. Some problems may be rooted in task complexity—as it rises, information use becomes more complex and grows more difficult to plan (cf. Saastamoinen et al., 2013). Tasks embedded in scientific collaboration are especially prone to uncertainty, with trial and error often being inherent to them. Moreover, history scholars applying computational methods need technology that supports their ways of working (Given & Willson, 2018) rather than pipeline-style techniques adopted from computer science and the life sciences (cf. Koolen et al., 2020).

Barriers associated with systems, protocols, and computation methods fall into our "Tools" context class. Scholars' information needs cannot always be met with existing methods such as entity recognition (cf. Kumpulainen et al., 2020). That said, scholars in the humanities participate actively in developing tools and infrastructure (Given & Willson, 2018; Late & Kumpulainen, 2021). Partly bound

up with such rapid development, inadequate skills in using the tools were brought up too by the participants. Since using digital tools is time-consuming and requires understanding of specific technologies and analysis processes (Given & Willson, 2018), using existing tools and creating new methods and techniques challenges scholars to learn new types of interdisciplinary skills.

Finally, socio-organizational context (encompassing phenomena related to organizations' structure and surrounding fields' culture), while having the smallest share of barrier instances in our dataset, is entangled with the obstacles that are probably the most difficult to overcome, for reason of their culturally pervasive nature. Clashes of cultures were visible at both macro level, between national cultures, and at micro level, between research cultures. Barriers in the latter may emerge from language problems (most typically "engineering language" vs. "historian language"). In moving between one culture and another, individuals' standards must be calibrated, and tensions among various discipline-specific approaches must be carefully managed in balancing acts that require negotiation and compromise (Poole & Garwood, 2018).

In addition, each community exhibits communal motives that may be difficult to articulate, and each community's distinct rules and norms guide behavior (Engeström, 2000). Breaking the rules for what is "appropriate" may feel too risky. One way to resolve this matter is to introduce cross-disciplinary work methods early in the research process (cf. the deep dialogues of Holton, 2001) or implement stage-aware collaborative information systems (Kumpulainen et al., 2018). When a new cross-disciplinary collaborative community is established with its own clear rules and norms, barriers of this sort may be lower. Also in this class are resource barriers, such as lack of time, which can bring the information interaction to a standstill. Interviewees expressed a sense of inequality too (e.g., feeling that some fields are more generously funded than others). This affective barrier could be overcome via more open information-sharing, to build trust (Swift & Hwang, 2013). Our findings are consistent with earlier discussion of lack of resources in building and maintaining data infrastructure; Poole and Garwood (2020) found sustainability to be a major challenge in digital humanities.

To answer the second research question, we examined the barriers' breakdown by activity group Barriers were encountered most frequently in the "working with items" activity class and least commonly in activities related to synthesis and reporting. This is not surprising, since information interactions involving the collection are less frequent toward the end of the research process (Late & Kumpulainen, 2021). Nonetheless, the results highlight the importance of studying the full information-interaction process, not just search. In addition to searching/selecting and working with items, quite a few barriers arose in the planning activity. If the goals are unclear, the task's perceived complexity increases, and information interactions become more complex and difficult (cf. Byström & Järvelin, 1995).

Our analyses show that the importance of barriers at particular layers of context hinges on the type of activity. For example, barriers in the collection and tools contexts were most common in activities that involve search and selection or working with items. In contrast, task-context and socio-organizational barriers were emphasized in planning and monitoring and in synthesis and reporting activities.

6.1 | Limitations

There were some limitations to the methods employed. First, we collected the history scholars' experiences via interviews, which are open to biases caused by personal chemistry and possible reticence to describe what actually happened (cf., Kumpulainen et al., 2009). However, rather than collect data on real activities, we investigated the history scholars perceived interactions and, thereby, could collect their insights related to the phenomenon at the heart of our study. Second, we used a variation of the CIT method. The incidents recalled may not be representative of all possible uses of the historical newspapers' content and might over-represent the simplest incidents. Still, we were able to collect in-depth information at lower cost than observations would permit and overcame several issues brought by the pandemic situation. Furthermore, the demonstrations helped to deepen our understanding of the information interactions, thus increasing the work's validity. A third limitation is related to the number of participants, which was low, with only 13 researchers. We found only a few historians who had applied at least some digital method while working with the newspaper collection. That said, the analysis satisfied us by demonstrating that the barrier types were repeated in our data, which indicates saturation.

Finally, the TBII model steered the research toward particular granularity levels envisioned for the activities. As a meta-activity, planning and monitoring might be expected to hide certain barriers from participants. However, the fact that this activity class occasioned almost as many barriers as searching and selecting attests that we found enough barriers related to even this class.

6.2 | Implications and future research

There are clear practical implications for newspaper-collection development and for tool design. Some workarounds employed for overcoming barriers in the collection

and tools contexts some workarounds may be such that the information tools/systems could incorporate them as scripts or work flows. More specific suggestions would require further work, for example, with full description derived via observational research with log data or video recordings of the steps taken. This is a topic for future research.

Addressing the poor OCR quality that caused most of the barriers requires improving the data. There are already ongoing efforts in this direction (Kettunen et al., 2020), but its documentation lacks details. Detailing the quality of the OCR-based content (e.g., related to decade-specific quality variations) would help scholars make informed decisions when planning their research projects. Further, often there is no in-depth documentation of the whole collection and what it includes or lacks. The information about the collection should also address possibly ongoing digitization projects.

Also, scholars expressed a desire for "segmentation" of content such that they can search at individual-article granularity. While attempts exist to provide automated clippings from today's full-page entities (Kettunen et al., 2019), doing this is not easy. The individual articles are seldom enough in any case; scholars need their metadata. These data are now provided mainly at newspaper-title level, but data specific to the individual issues and the articles within would assist researchers greatly.

On the level of tasks, the barriers require better communication tools to facilitate managing the tasks and handling division of labor. Collaborative work modes for dealing with the clippings would aid collaborators working on a given topic in parallel and help them follow the progress of others in the research project. One workaround for joint work entails all parties using the same NLF user account, but this precludes simultaneous work since only one person can be logged in as that user at a time.

The lack of resources that participants identified as creating some barriers in socio-organizational context may be partly resolved through research infrastructure for history scholars. There is already a national effort in Finland aimed at creating digital-humanities infrastructure. This initiative, FIN-CLARIAH, could reduce research projects' costs via such mechanisms as sharing of digital methods and researcher training. Further, support for open data and resolution of intellectual-property issues could help tackle several barriers by making the personal digital collections more readily reusable.

Several improvements are already being made to the newspaper collection itself, with corresponding changes to the ways of interacting with it. Some of the barriers described here may evaporate as this progress unfolds. However, there are many other collections out there. Also, scholarly methods and digital collections are developing constantly, and these advances call for better understanding of how to provide improved information environments for scholars. Collaboration, which is only growing, seems to be of particular import in this regard, since it appeared to multiply the barriers experienced. Apart from the practical implications, this adds to the task-based information interaction model by differentiating the contextual levels. A framework such as ours should prove instrumental to examining and addressing such issues methodically.

7 | CONCLUSION

For enhancing the information interactions entailed by information-intensive tasks, solid understanding of those interactions throughout the task process is pivotal to avoiding an overly narrow view of human information activities. Our analysis of the barriers to information interaction that faced history scholars in research tasks wherein historical newspapers served as primary sources revealed connections between the perceived barriers and both context level and activity type. Barriers in search and selection activities and in working with information items were clear in the context of the newspaper collection, but a wider work-process perspective revealed that, with task-planning and monitoring and with synthesis and reporting activities, the number of barriers was greater in the context of the task. Studying the barriers in a task-based setting contributes to understanding human information interaction and overcoming the obstacles to it. Our work in this direction points to definite implications for improved access to digitized collections and for designing systems that better support the task activities and related information interactions.

ACKNOWLEDGMENTS

This research was funded by Academy of Finland grant #326616. We thank Laura Korkeamäki for providing research data for this study.

ORCID

Sanna Kumpulainen https://orcid.org/0000-0002-7016-257X

Elina Late https://orcid.org/0000-0002-3232-1365

REFERENCES

Aagaard-Hansen, J. (2007). The Challenges of Cross-disciplinary Research. *Social Epistemology*, *21*(4), 425–438. https://doi.org/10.1080/02691720701746540

Allen, R. B., & Sieczkiewicz, R. (2010). How historians use historical newspapers. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–4. https://doi.org/10.1002/meet.14504701131

Borgman, C. L. (1985). The user's mental model of an information retrieval system. *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Montreal, Quebec, Canada. pp. 268–273. https://doi.org/10.1145/253495.253533

Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing & Management*, *31*(2), 191–213. https://doi.org/10.1016/0306-4573(95)80035-R

Byström, K., & Kumpulainen, S. (2020). Vertical and horizontal relationships amongst task-based information needs. *Information Processing & Management*, *57*(2), 102065. https://doi.org/10.1016/j.ipm.2019.102065

Clement, T. E., & Carter, D. (2017). Connecting theory and practice in digital humanities information work. *Journal of the Association for Information Science and Technology*, *68*, 1385–1396. https://doi.org/10.1002/asi.23732

Conway, P. (2015). Digital transformations and the archival nature of surrogates. *Archival Science*, *15*(1), 51–69. https://doi.org/10.1007/s10502-014-9219-z

Cool, C., & Spink, A. (2002). Issues of context in information retrieval (IR): An introduction to the special issue. *Information Processing & Management*, *38*(5), 605–611. https://doi.org/10.1016/S0306-4573(01)00054-1

Courtright, C. (2007). Context in information behavior research. *Annual Review of Information Science and Technology*, *41*(1), 273–306. https://doi.org/10.1002/aris.2007.1440410113

Ehrmann, M., Bunout, E., & Düring, M. (2019). Historical newspaper user interfaces: A review. *IFLA WLIC 2019—Athens, Greece—Libraries: Dialogue for change*. http://library.ifla.org/2578/

Engeström, Y. (2000). Activity theory as a framework for analyzing and redesigning work. *Ergonomics*, *43*(7), 960–974. https://doi.org/10.1080/001401300409143

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, *51*(4), 327–358. https://doi.org/10.1037/h0061470

Given, L. M., & Willson, R. (2018). Information technology and the humanities scholar: Documenting digital research practices. *Journal of the Association for Information Science and Technology*, *69*(6), 807–819. https://doi.org/10.1002/asi.24008

Gooding, P. (2016). Exploring the information behaviour of users of Welsh Newspapers Online through web log analysis. *Journal of Documentation*, *72*(2), 232–246. https://doi.org/10.1108/JD-10-2014-0149

Hoekstra, R., & Koolen, M. (2019). Data scopes for digital history research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *52*(2), 79–94. https://doi.org/10.1080/01615440.2018.1484676

Holton, J. A. (2001). Building trust and collaboration in a virtual team. *Team Performance Management*, *7*(3/4), 36–47. https://doi.org/10.1108/13527590110395621

Huvila, I. (2019). Rethinking context in information research: Bounded versus centred sets. *Information Research*, *24*(4), paper colis1912. http://informationr.net/ir/24-4/colis/colis1912.html

Jarlbrink, J., & Snickars, P. (2017). Cultural heritage as digital noise: Nineteenth century newspapers in the digital archive. *Journal of Documentation*, *73*(6), 1228–1243. https://doi.org/10.1108/JD-09-2016-0106

Järvelin, K., Vakkari, P., Arvola, P., Baskaya, F., Järvelin, A., Kekäläinen, J., ... Sormunen, E. (2015). Task-based information interaction evaluation: The viewpoint of program theory. *ACM Transactions on Information Systems*, *33*(1), 1–30. https://doi.org/10.1145/2699660

Kachaluba, S. B., Brady, J. E., & Critten, J. (2014). Developing humanities collections in the digital age: Exploring humanities faculty engagement with electronic and print resources. *College & Research Libraries*, *75*(1), 91–108. https://doi.org/10.5860/crl12-393

Kelly, D., & Cool, C. (2002). The effects of topic familiarity on information search behavior. *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, Portland, OR. pp. 74–75. https://doi.org/10.1145/544220.544232

Kettunen, K., Koistinen, M., & Kervinen, J. (2020). Ground truth OCR sample data of Finnish historical newspapers and journals in data improvement validation of a re-OCRing process. *LIBER Quarterly*, *30*(1), 1–20. https://doi.org/10.18352/lq.10322

Kettunen, K. & 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*, pp. 956–961.

Kettunen, K., Pääkkönen, T., & Liukkonen, E. (2019). Clipping the page—Automatic article detection and marking software in production of newspaper clippings of a digitized historical journalistic collection. In A. Doucet, A. Isaac, K. Golub, T. Aalberg, & A. Jatowt (Eds.), *Digital libraries for open knowledge. TPDL 2019. Lecture notes in computer science* (Vol. 11799). Springer. https://doi.org/10.1007/978-3-030-30760-8_33

Koolen, M., Kumpulainen, S., & 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. https://doi.org/10.1145/3343413.3377969

Korkeamäki, L., & Kumpulainen, S. (2019). Interacting with digital documents: A real life study of historians' task processes, actions and goals. *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pp. 35–43. https://doi.org/10.1145/3295750.3298931

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, *42*(5), 361–371.

Kumpulainen, S. (2017). Task-based information searching: Research methods. In *Encyclopedia of library and information sciences* (pp. 4526–4536). CRC Press.

Kumpulainen, S., Huurdeman, H., & Keskustalo, H. (2018). Personalization needs extension towards task stages in collaborative research work tasks. In Jones, G.J.F, Belkin NJ, Lawless S, Pasi G, (Eds.) *WEPIR 2018: Workshop on Evaluation of Personalisation in Information Retrieval*. https://wepir.adaptcentre.ie/papers/WEPIR_2018_paper_2.pdf

Kumpulainen, S., & Järvelin, K. (2012). Barriers to task-based information access in molecular medicine. *Journal of the American Society for Information Science and Technology*, *63*(1), 86–97. https://doi.org/10.1002/asi.21672

Kumpulainen, S., Järvelin, K., Serola, S., Doherty, A., Byrne, D., Smeaton, A. F., & Jones, G. F. J. (2009). Data collection methods for analyzing task-based information access in molecular medicine. P*roceedings of the 1st International Workshop on Mobilizing Health Information to Support Healthcare-Related Knowledge Work*, pp. 49–58.

Kumpulainen, S., Keskustalo, H., Zhang, B., & Stefanidis, K. (2020). Historical reasoning in authentic research tasks: Mapping cognitive and document spaces. *Journal of the Association for Information Science and Technology*, *71*(2), 230–241. https://doi.org/10.1002/asi.24216

Late, E., & Kumpulainen, S. (2021). Interacting with digitised historical newspapers: Understanding the use of digital surrogates as primary sources. *Journal of Documentation*. https://doi.org/10.1108/JD-04-2021-0078

Late, E., Tenopir, C., Talja, S., & Christian, L. (2019). Reading practices in scholarly work: From articles and books to blogs. *Journal of Documentation*, *75*(3), 478–499. https://doi.org/10.1108/JD-11-2018-0178

Latour, B. (1987). *Science in action*. Harvard University Press.

Leonelli, S. (2019). Data governance is key to interpretation: Reconceptualizing data in data science. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.17405bb6

Martin-Rodilla, P., & Sánchez, M. (2020). Software support for discourse-based textual information analysis: A systematic literature review and software guidelines in practice. *Information*, *11*(5), 256. https://doi.org/10.3390/info11050256

Milligan, I. (2013). Illusionary order: Online databases, optical character recognition, and Canadian history, 1997–2010. *Canadian Historical Review*, *94*(4), 540–569. https://doi.org/10.3138/chr.694

Monte-Sano, C., & De La Paz, S. (2012). Using writing tasks to elicit adolescents' historical reasoning. *Journal of Literacy Research*, *44*(3), 273–299. https://doi.org/10.1177/1086296X12450445

Palmer, P., Teffeau, L., & Pirmann, C. (2009). Scholarly information practices in the online environment. Report commissioned by OCLC research.

Poole, A. H., & Garwood, D. A. (2018). Interdisciplinary scholarly collaboration in data-intensive, public-funded, international digital humanities project work. *Library & Information Science Research*, *40*(3–4), 184–193. https://doi.org/10.1016/j.lisr.2018.08.003

Poole, A. H., & Garwood, D. A. (2020). Digging into data management in public-funded, international research in digital humanities. *Journal of the Association for Information Science and Technology*, *71*(1), 84–97. https://doi.org/10.1002/asi.24213

Saastamoinen, M., Kumpulainen, S., Vakkari, P., & Järvelin, K. (2013). Task complexity affects information use: A questionnaire study in city administration. *Information Research*, *19*(4), paper592. http://www.informationr.net/ir/18-4/paper592.html#.YZUXC7rhW1s

Salmi, H. (2021). *What is digital history?*. Polity Press.

Savolainen, R. (2016). Approaches to socio-cultural barriers to information seeking. *Library & Information Science Research*, *38*(1), 52–59. https://doi.org/10.1016/j.lisr.2016.01.007

Sinn, D., & 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*, *65*(9), 1794–1809. https://doi.org/10.1002/asi.23091

Sonnenwald, D. H. (2008). Scientific collaboration. *Annual Review of Information Science and Technology*, *41*(1), 643–681.

Strauss, A., & Corbin, J. M. (1997). *Grounded theory in practice*. Sage.

Swift, P. E., & Hwang, A. (2013). The impact of affective and cognitive trust on knowledge sharing and organizational learning. *The Learning Organization*, *20*(1), 20–37. https://doi.org/10.1108/09696471311288500

Talja, S., Keso, H., & Pietiläinen, T. (1999). The production of 'context' in information seeking research: A metatheoretical view. *Information Processing & Management*, *35*(6), 751–763. https://doi.org/10.1016/S0306-4573(99)00024-2

Terras, M., Baker, J., Hetherington, J., Beavan, D., Zaltz Austwick, M., Welsh, A., … Farquhar, A. (2018). Enabling complex analysis of large-scale digital collections: Humanities research, high-performance computing, and transforming access to British Library digital collections. *Digital Scholarship in the Humanities*, *33*(2), 456–466. https://doi.org/10.1093/llc/fqx020

Toms, E. G., & O'Brien, H. L. (2008). Understanding the information and communication technology needs of the e-humanist. *Journal of Documentation*, *64*(1), 102–130. https://doi.org/10.1108/00220410810844178

Unsworth, J. (2000). Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this. *Symposium on Humanities Computing: Formal Methods, Experimental Practice*. King's College, London, England, p. 13.

Vakkari, P. (2001). A theory of the task-based information retrieval process: A summary and generalisation of a longitudinal study. *Journal of Documentation*, *57*(1), 44–60. https://doi.org/10.1108/EUM0000000007075

Wildemuth, B. M. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology*, *55*(3), 246–258. https://doi.org/10.1002/asi.10367

How to cite this article: Kumpulainen, S., & Late, E. (2021). Struggling with digitized historical newspapers: Contextual barriers to information interaction in history research activities. *Journal of the Association for Information Science and Technology*, 1–13. https://doi.org/10.1002/asi.24608