Received 14 April 2022
Revised 7 October 2022
Accepted 13 October 2022

Task information types related to data gathering in media studies

Laura Korkeamäki, Heikki Keskustalo and Sanna Kumpulainen
Tampere University, Tampere, Finland

Abstract

Purpose – The purpose of this paper is to examine what types of task information media scholars need while gathering research data to create new knowledge.

Design/methodology/approach – The research design is qualitative and user-oriented. A total of 25 media scholars were interviewed about their research processes and interactions with their research data. The interviews were semi-structured, complemented by critical incident interviews. The analysis focused on the activity of gathering research data. A typology of information (task, domain and task-solving information) guided the analysis of information types related to data gathering, with further analysis focusing only on task information types.

Findings – Media scholars needed the following task information types while gathering research data to create new knowledge: (1) information about research data (aboutness of data, characteristics of data, metadata and secondary information about data), (2) information about sources of research data (characteristics of sources, local media landscapes) and (3) information about cases and their contexts (case information, contextual information). All the task information types should be considered when building data services and tools to support media scholars' work.

Originality/value – The paper increases understanding of the concept of task information in the context of gathering research data to create new knowledge and thereby informs the providers of research data services about the task information types that researchers need.

Keywords Information types, Research data, Media studies

Paper type Research paper

1. Introduction

The purpose of this study is to examine what types of information media scholars need in the context of their real-world research processes. We focus on task information, which is information that is specific to the task in question (Byström, 1999, p. 45). Furthermore, we focus on the activity of gathering research data, which we consider as a subtask of knowledge creation. Researchers gather research data to answer specific research questions and formulate new ones. Therefore, data gathering is a key activity in knowledge creation. We define data gathering broadly, involving searching, selecting and collecting research data. Although searching, selecting and collecting are theoretically different activities, they can intertwine and overlap in real-world situations (Järvelin *et al.*, 2015). Especially searching can be difficult to differentiate from other activities because it is sometimes a simultaneous activity or even embedded in other activities (Late and Kumpulainen, 2022).

Media studies is an interdisciplinary field in the intersection of humanities and social sciences (Jensen, 2012). Media scholars gather research data to study media-related phenomena; their research interests range from media history to contemporary media. However, media landscape is increasingly complex and dynamic, which can affect what

This work was supported by The Emil Aaltonen Foundation and Academy of Finland (Grant Numbers 326616 and 351247).

Journal of Documentation
Vol. 78 No. 7, 2022
pp. 528-545
Emerald Publishing Limited
0022-0418
DOI 10.1108/JD-04-2022-0082

information media scholars need while doing research. This makes media studies a particularly interesting domain for information interaction research. Information interaction studies the interaction between people and information regardless of the mediating vehicle, being it information technology, human being or a collection of books (Fidel, 2012, p. 17).

The significance of this research is threefold. Firstly, media landscape is quickly evolving, making data gathering challenging in media studies. This makes media studies an important and timely research area for information studies. Secondly, research data in media studies can be nearly anything from archival sources in paper form to social media data. Therefore, data gathering is complex and requires multiple types of information. It is important to study what information media scholars need for data gathering in order to design tools and data services that meet those needs. Thirdly, according to Byström (1999, pp. 45–46), task information is needed only for the specific task, which makes it inherently different from domain and task-solving information that are needed in several tasks alike. Byström (2009, p. 175) also noted that the definitions of task, domain and task-solving information are general in nature and saw the need for the development of subtypes in future studies. The present study addresses this need by elaborating task information types.

This study is situated within information interaction research in information studies. By studying data gathering and related task information types, we bring researchers' interactions with their research data to the forefront. Our view is that the need for information may arise in the information interaction situation of the subtask in question. Furthermore, we focus on specific information activity (data gathering). This approach is consistent with a task-based information interaction (TBII) model, where information interaction is defined by means of information activities (Järvelin *et al.*, 2015). The perspective of the present study is user-oriented because it focuses on individual media scholar's real-world research processes. This position differs from system-oriented approach, where research might focus on studying the performance of information retrieval systems (Ingwersen and Järvelin, 2005, pp. 114–115). User-oriented perspective is also the opposite of data-centric perspective (e.g. how data travel between organizations? or how data change when moved from one research site to the next?). This perspective helps with transitioning the research focus from studying the data toward helping their users. We analyzed interviews of 25 media scholars who were asked about their research processes and interactions with their research data. Our study addresses the following research question:

RQ1. What task information types do media scholars need while gathering research data to create new knowledge?

2. Literature review

2.1 Research data in media studies

Research data can be called data, primary sources or research materials. A media historian, for example, might prefer using the term primary sources (see Long *et al.*, 2022, pp. 470–506). In this study, we use the term research data. Because research data is a concept that is difficult to define, we capture the concept by describing the various ways it has been approached in the literature.

Research data are something that is used as evidence in research, for example, evidence of phenomena (Borgman, 2015, p. 28) or evidence for knowledge claims (Leonelli, 2015). Research data are also used for discovering new research questions (Gibbs and Owens, 2013). Contextual information is also an important part of the research data. This includes, for example, information about how the research data were created, how the research data have been archived, for what purposes have the research data been used in the past, and what are the terms of use of the research data (Faniel *et al.*, 2019). Sometimes, there is only a fine line between research data and reflective information. Reflective information, like keeping a

research journal, can be thought either as a reflective tool or as an additional data set, for example, if used in action research (Rowley, 2014). To sum up the previous, we define research data as follows: research data are used as evidence in research, they are used for discovering research questions, and research data include all information (i.e. content, context and other information) that are used as inputs to knowledge creation. Furthermore, research data can be gathered from various sources.

Media studies is an interdisciplinary field with roots in social sciences and humanities (Jensen, 2012). Therefore, the data practices and the types of research data that media scholars use vary as well. In social sciences and humanities, it can be difficult to define what exactly are meant by research data because they can be so many things (Borgman, 2007, pp. 204–224). For example, in social sciences, researchers collect research data either themselves (e.g. by doing surveys) or alternatively, they use research data collected by others (e.g. public records). In humanities, there may be difficulties differentiating research data from secondary sources (Borgman, 2007, pp. 204–224). Furthermore, according to a study that provided an overview of communication and media research in nine countries (including Finland), it is very country specific how research in media studies is organized and the focus areas of research and education differ between academic research institutions (e.g. universities, private media research companies and institutions were excluded from the analysis) (Koivisto and Thomas, 2010).

In media studies, research often focuses on media texts, media production, media audiences or media history (Long *et al.*, 2022). A media text can be defined as "a unit of meaning for interpretation and understanding" (Gray, 2017). This means more than just written texts like newspaper articles or books. Media texts also include, for example, TV programs, games and podcasts. When working with media texts, media scholars are interested in their meanings, as well as "tracking how context works, and hence in how they connect, to each other, to the outside world, to their producers, and to their audiences" (Gray, 2017.). In humanities, it is typical to close read the texts, gather representations of them (e.g. by taking notes or photographs) and make comparisons (Borgman, 2015, pp. 161–202). Media production research focuses on the business of media, its economy, operation of media organizations and the work of media professionals (Long *et al.*, 2022, pp. 200–248). Media production can be studied using various kinds of research data, for example, using any information released by media companies themselves, interviewing media professionals or observing media production processes all of which can sometimes be challenging to acquire (Lee and Zoellner, 2019). Media audiences are understood and studied, for example, as passive media consumers, or as active listeners, readers, players, bloggers and so on, who consume, interpret and produce media (Long *et al.*, 2022, pp. 304–359). Audience research methods include, for example, surveys, interviews, ethnographic research and action research (Patriarche *et al.*, 2014). Media history is a subfield that studies media from historical perspective "to better understand the nature of contemporary media practices, institutions and cultures" (Long *et al.*, 2022, p. 477). Media historians often rely on archival sources.

Based on prior literature, we know that media scholars use numerous types of research data to create new knowledge. However, in this study, we are interested in what information media scholars need while gathering the data. Uncovering this will help us to understand better what information is important to media scholars when they search, select and collect research data.

2.2 Approaches to task information in information studies

Task information has been approached in different ways in information studies. Our study builds on Byström's (1999, pp. 45–47) typology of information needed in tasks: task information (answers to information requirements of the specific task, often facts such as

names, addresses or courses of events), domain information (a more general type of information such as theories, concepts or known facts) and task-solving information (procedural information such as methods or instructions). In the typology, Byström defined information as an abstract tool that enables the completion of a task, and information need as a need to acquire that tool. Based on previous literature, Byström gave an example of the three information types in journalism: a journalist who is writing about a traffic accident needs facts about the specific accident (task information), information about accidents in general (domain information) and information about journalistic procedures (task-solving information). We use Byström's definition of task information as a starting point to study task information types needed in data gathering in media studies. Domain and task-solving information are outside the scope of this study.

Byström (1999) based her typology (task, domain and task-solving information) on prior typology of problem information, domain information and problem-solving information (Byström and Järvelin, 1995; see also Barr and Feigenbaum, 1981). Byström (1999, pp. 45–47) adjusted and renamed the concepts, emphasizing tasks instead of problems. The biggest difference is between problem information and task information. Problem information includes problem descriptions, which Byström saw as task-solving information and not as task information.

There are several other typologies of information. Crescenzi *et al.* (2019) identified five information types that participants saved while taking notes during exploratory search tasks: background information, facets of the topic (e.g. causal factors), specific details (e.g. statistics), information sources (e.g. URL) and information to help with work tasks (e.g. useful queries). Applying the former, Choi *et al.* (2019) formed four information types to study how search task complexity affects their use: facts (objective statements), concepts (noun phrases denoting ideas, principles or entities), opinions (subjective statements) and insights (advice or tips). Choi and others emphasized that their typology is based on inherent characteristics of information as opposed to Byström's (1999, p. 45) functional role of information, that is, using information as an abstract tool. Li *et al.* (2021) studied cross-session search tasks (i.e. searches made across multiple sessions to achieve a goal) and found four information types that users needed and searched for during those tasks: factual and conceptual information (e.g. names, locations as factual; e.g. descriptions of relationships, distinctions between concepts as conceptual), procedural information (e.g. how to do something), opinions (e.g. book reviews, professional recommendations) and information for helping metacognition (that is, information that helps to understand information already found). All these studies focus on search tasks (individual or cross-session) that make them narrower in terms of task granularity (Byström and Hansen, 2005; Byström and Kumpulainen, 2020). Larger tasks may include not only searching but also span across other kinds of activities (Järvelin *et al.*, 2015).

A study about information types needed in creative projects by Zhang *et al.* (2020) found six types of information: specific information (e.g. about people, locations and products), domain information, information about how to do something, examples of the work of others, recommendations and guidance from other people and motivating information that motivates the task-doers to keep going on their work. Zhang *et al.* compared their findings to the typology of problem information, domain information and problem-solving information in Byström and Järvelin (1995) and did not discuss Byström's (1999) task information specifically. Our view is that specific information in Zhang *et al.* (2020) compares to Byström's (1999) task information.

A study about information types needed in data reuse by Faniel *et al.* (2019) studied context information types needed. They interviewed researchers from quantitative social science, archeology and zoology. They identified three categories containing a total of 12 context information types. The first category was information about data production that included information about research objectives, data collection and analysis methods,

information about missing data, information about specimen and artifacts and information about data producer. The second category was repository information that included information about reputation and history of repository, provenance and curation and digitization. The third category was data reuse information that included information about prior reuse, advice on reuse and information about terms of use. In contrast to their study, ours is not limited to reusing already gathered research data. We also include research data gathered by media scholars themselves.

3. Methods

3.1 Data collection

3.1.1 Recruitment and participants. Suitable participants for this study were academic researchers who were affiliated to media or game studies and were currently conducting their research or had recently finished their research. Participants were recruited from three universities in Finland to cover a variety of research interests in media studies, because media studies is a research field with varying definitions and traditions in different research institutions (Koivisto and Thomas, 2010). A few participants were contacted on the recommendation of other participants. Participants were first contacted via email, and on one occasion, in a face-to-face meeting with a research group.

A total of 25 media scholars participated in this study. All of them participated in a semi-structured interview, and 12 of them also demonstrated their work in a critical incident interview. The semi-structured and critical incident interviews were conducted in March 2019–April 2020. Informed consent was given by the participants. Nine of the participants were doctoral students, and 16 were post-doctoral researchers, professors, university researchers or university lecturers. Most of the participants identified themselves with one or more of the following research fields: media studies, game studies or communication studies. Some of the participants related themselves to both communication and media studies, whereas one media researcher specifically mentioned "I'm not a communication researcher by any means". In addition, more specific research fields were mentioned, such as film studies, film history, journalism, visual studies, social media research, audience research, critical research or media studies with an emphasis in humanistic, feministic or political research.

3.1.2 Semi-structured interviews. The model of TBII (Järvelin *et al.*, 2015) informed the design of interview questions to cover important information activities from the point of view of interacting with research data. The TBII model is defined by means of five information activities: (1) task planning and reflective assessment (e.g. setting goals, planning how to proceed in the task), (2) searching information items (e.g. using information retrieval systems, browsing interesting material), (3) selecting information items (e.g. selecting relevant materials), (4) working with information items (e.g. reading, organizing or analyzing information) and (5) synthesizing and reporting (e.g. making conclusions, writing research articles). The activities 1–4 were applied in the design of the interview questions. The fifth activity was outside the scope of this study. The TBII model is useful in understanding theoretically the different types of information activities. Therefore, it helps identifying researchers' information activities when collecting empirical data about their research processes. In addition, some of the interview questions concerned research community (e.g. working in a research group) and rules and norms (e.g. research ethics) (see Allen *et al.*, 2011).

Of the 25 semi-structured interviews, 15 were conducted face-to-face and 10 (due to COVID-19 pandemic) by phone. We acknowledge that videoconferencing would have been more similar to face-to-face interviews. However, at the early days of the pandemic, it was unclear whether such a tool that meets GDPR requirements is available on behalf of the research organization. We also had a tight schedule for the interviews. The interviews were conducted over the phone because we wanted to ensure the data security of the participants in

accordance with the GDPR and our data management plan. We also wanted to treat all participants the same for remote semi-structured interviews during the pandemic. When the research organization later updated its guidelines on videoconferencing tools suitable for collecting research data, some participants were offered the opportunity to do a demonstration (critical incident interview) via Microsoft Teams, but only one participant chose this option.

The semi-structured interviews started with background questions (see Appendix). The interview questions also covered participants' research topics, research processes (research goals and questions, phases in the research process) and research data (description of the research data; finding, selecting, collecting, analyzing, managing and archiving of the research data), as well as ethical issues, ownership and licensing related to the research data. Participants were also asked about working in a research group (if applicable). The interviewees were asked to choose some ongoing or recently completed research project that could be discussed in the interview. The goal was to encourage the interviewees to talk about the progress of their research in any order and using their own words. Hence, the interview protocol served as a guide, and the order and wordings of the interview questions varied according to what was suitable for each interview. Also, probes were posed to encourage the interviewees to continue describing their research process, or to ask for clarifications, details or examples (see Roulston, 2010, pp. 9–32). The semi-structured interviews were audio recorded. Each semi-structured interview lasted between 46 min–1 h 16 min, totaling 24 h 26 min. The interviews were transcribed word by word for analysis, totaling 290 pages.

3.1.3 Critical incident interviews. The critical incident interviews were conducted right after the semi-structured interviews ended. As in the critical incident technique (Flanagan, 1954), the participants were asked to demonstrate some recent or ongoing part of their work with their research data. Of the 25 participants, only 12 demonstrated their work (of these, 11 were conducted face-to-face and one via Microsoft Teams). This was partly because of the onset of COVID-19 pandemic that restricted face-to-face meetings. In addition, some of the interviewees felt it difficult to demonstrate one specific part of their work. The critical incident interviews were video recorded (10 interviews), or alternatively, photographs were taken during the demonstrations (2 interviews). Each video recorded interview lasted between 6 min–21 min, totaling 2 h 6 min. The video recordings were transcribed for analysis, including word by word transcriptions of discussions and written descriptions of activities during the demonstrations, totaling 27 pages. Any names of persons or organizations appearing in the data were removed during the transcribing. Ethical approval procedure was not required according to the research organization guidelines, since all the participants were adults and provided informed consents.

3.2 Data analysis

Our analysis focused on the activity of gathering research data. As the interviewees talked about their research processes, they also talked about what information they needed while gathering research data – even though the interview protocol did not include direct questions about the issue. Because of this finding, we focused our analysis on types of information needed while gathering research data. Furthermore, we limited our analysis to the activity of gathering naturally occurring data that exist originally for some other purpose than research and become research data when gathered for research (see Lester and O'Reilly, 2019, pp. 97–122). In total, 20 of the 25 media scholars gathered naturally occurring data including journalistic texts (e.g. newspaper articles), political documents, monographs and social media data, as well as TV programs, films and related material (e.g. PR material). Consequently, the activities of gathering new data (interview, survey or workshop data) were excluded from the

analysis. Gathering new data is an important part of knowledge creation for many researchers. However, because it is inherently different from gathering naturally occurring data (see Lester and O'Reilly, 2019, pp. 97–122), and because there were only five participants who only gathered new data, we decided that it is a subject for future research.

The transcribed interviews were read through several times and coded using ATLAS.ti software. The analysis started as theory-driven, using Byström's (1999, pp. 45–47) concept of task information as a starting point. Byström defined task information as information that is specific to tasks and that is more specific than domain information and different from task-solving information. However, this definition leaves room for interpretation of what it means in the context of different tasks. For the purpose of this study, we defined task information as information that is needed specifically for data gathering in knowledge creation. We excluded domain information (which in our research data was information about theories, models and previous research) and task-solving information (which in our research data was information about research methods and tools) from the analysis. The analysis continued as data-driven using sub-coding (Miles *et al.*, 2020, p. 72) to denote the different qualities of task information. Lastly, we used pattern coding (Miles *et al.*, 2020, pp. 79–83) to group the sub-codes into three main types based on their relation to research data, sources of research data or cases and their contexts that were of interest to media scholars. The coding was done by one researcher in multiple cycles in an iterative manner. The coding process and the decisions made were discussed in a group of three researchers to achieve agreement. In findings, we present quotations that are typical examples of the interview data or that they show variance in the interview data. Because the interviews were in Finnish, the quotations presented in this article were translated in English.

4. Findings

The findings are summarized in Table 1. We found three task information types that included a total of eight subtypes. Firstly, media scholars needed information about research data. It included aboutness and characteristics of data, metadata and secondary information about

Task information types	Subtypes	Definitions
Information about research data	Aboutness of data	What the data items are about, what topics they relate to, what themes they represent or what meanings they convey
	Characteristics of data	Properties and aspects that describe the data (provenance, authenticity, completeness, type, granularity, temporal continuity and overview)
	Metadata	Bibliographic and catalog information, self-created descriptive annotations
	Secondary information about data	Information from other sources that tells something about data items that cannot be found or no longer exist
Information about sources of research data	Characteristics of sources	Properties and aspects that describe the sources (geographical area, genre, format, age and publication frequency, media reach and publicity, ownership, conditions for discussion and content creation, accessibility and usability)
	Local media landscapes	Information about media landscapes of a particular country or countries that the sources are a part of
Information about cases and their contexts	Case information	Information about an example of something occurring that is of interest
	Contextual information	Information that is conceptually connected to circumstances of a case

Table 1.
Task information types and their subtypes with definitions

data. Secondly, media scholars needed information about data sources. It included characteristics of sources and information about local media landscapes the sources are a part of. Thirdly, media scholars needed information about cases and their contexts. It included case information (information about an example of something occurring that is of interest) and contextual information (information that is conceptually connected to circumstances of a case). Next, we present the findings in more detail.

4.1 Information about research data

Information about research data concerns information that is related to the subject matter, provenance and form or format of the data set. We analyzed the types of information needed about research data while gathering the research data.

4.1.1 Aboutness of data. Many of the participants gathered media texts (e.g. blogs, films, social media posts and news articles) as research data. They needed information about what the media texts are about, what topics they relate to, what themes they represent or what meanings they convey. This can be called aboutness of the texts.

Some participants searched for media texts (or pieces within the texts) that were about or related to specific topics. They expressed their information needs as topical search terms and made queries into various web interfaces, as in the following example:

> Then I searched each newspaper's online archive for [a topical search term] and asterisk, that is, everything related to [the topic] [. . .] every single news from each newspaper. [P16]

Sometimes aboutness was beyond the apparent and was based on the interpretations that participants associated with the media texts during close reading. A participant, who studied films, watched all potentially relevant films to select the ones that best represent certain themes of how people are portrayed in films:

> I wanted to watch the films that I hadn't seen [. . .] and somehow catalogue them [. . .] and then classify them by their themes [. . .] and to all of them [themes]. I selected [. . .] the most representative films [. . .]. [P24]

Participants also talked about selecting media texts based on how normative (or not) they seemed based on their content or what meanings (e.g. irony) they conveyed. When the emphasis was on the various meanings (rather than on the topics) of the texts, it was sometimes difficult to search them systematically. Media texts were also evaluated in relation to each other, in which case participants' understanding of aboutness and relevance of the texts gradually increased.

4.1.2 Characteristics of data. Characteristics of data are the properties and aspects that describe the data. Media scholars expressed that they paid attention to various characteristics of data when gathering research data. *Provenance* means information about how the research data were created and information about other data items in proximity of the research data. A participant recounted that, in a research group, descriptions of how the research data were gathered were shared with others:

> [The research data] can be found there on [a shared file location] [. . .] and then there are the descriptions from each researcher of how this data were gathered. Just because now each of us have access [. . .] to all the data. [P11]

Another participant, who gathered social media data, included the topically relevant posts as well as the topically nonrelevant posts in their proximity to the research data to preserve the context.

> I have taken all the posts from that period although not all of them [are topically relevant], but I want the context, what is the context in which the posts appear. [P6]

Authenticity means that participants wanted to save the data items as they were originally. Sometimes this was difficult because of technical issues. For example, one participant talked about the hardship on gathering and saving webpages in authentic form without losing information:

> [. . .] the downside of [data collecting software] is that [. . .] for example, here was a video that was also a picture of this [news article], so because the link is no longer active, the picture also disappears. [P8]

Some authenticity was also preserved by taking extracts from the media texts used as research data. For example, one participant wrote down quotes from TV programs to have the exact words and conversation to help with the analysis. Another participant was taking pictures of TV programs to capture their visual style:

> [. . .] of those TV programs [. . .] I take some pictures so that [I] can later view the visual style as well, because otherwise, describing it afterwards would be quite hard. [P25]

Participants also talked about *completeness* of a data set. Some participants wanted to be sure that the data sets they gathered are complete, meaning that they would include all media texts of interest (e.g. all films) that can be found (with a reasonable amount of work) from a certain period of time. One way of finding out whether they had all the media texts they wanted was verifying this from some separate source of information (e.g. by following release lists of films). However, sometimes, it can be difficult to achieve full certainty as to whether every single relevant item is included in the research data. One participant described ways of trying to make sure that the search (for news articles, for example) has been thorough:

> I always have a feeling that have I missed something, and then I do searches with multiple search terms [. . .] and then I do kind of double-checking [. . .] just trying to make sure that there isn't anything relevant left that could be found, and that way [. . .] starting to build a sense of certainty. [P14]

Type of the data was an important characteristic. For example, one participant desired news content but wanted to filter out the editorials of the newspapers. Another participant limited the research data to feature films (instead of short films, for example). Furthermore, *granularity* of the data played a role in their data gathering. This means the level of details in the data set, which affected the potential analysis methods. For example, the objects that share the same identity number were an issue when a social media platform was updated:

> At one point, my intention was to analyze it [. . .] per actor, which [actors] are doing what [. . .]. But then [. . .] the anonymization changed [. . .]. Then you could no longer know who updated how many things for example. And it was perhaps not so essential in my research after all, that it would have hindered it significantly. [P21]

Participants paid attention to *temporal continuity* of the data. One media scholar gathered a data set of journalistic articles that covered "a long time series". Another media scholar used a data archive to search for multiple data sets of media texts covering roughly the same, "longer period of time".

Lastly, some expressed a need to have an *overview* of their data set in order to understand it. Here, data gathering intertwined with data analysis. For example, overview of the data set was needed to make further decisions how to select the data. Some media scholars used analysis tools to make visualizations of their research data and used this information to select certain parts of the data for further analysis.

4.1.3 Metadata. Metadata included bibliographic and catalog information, and self-created descriptive annotations. Bibliographic and catalog information were used in finding, organizing and re-finding research data. In the absence of such information (as for uncatalogued material), one option was to travel to the archive to go through folders. A participant said about searching for administrative documents of media institutions:

> Not everything is in databases that are easy to search [. . .] we went there [to the archive] to look at the shelves [. . .] to find which folders to look more closely. [P25]

The participant added that materials might lack important bibliographic information altogether, like:

> [. . .] what year some materials are from [. . .] they are in folders, and not all of those publications necessarily have any year [of publication], so then it might not be easy to time them accurately. [P25]

Besides using metadata provided by others, participants created their own notes of such information to make their research data more manageable. This included dealing with missing information. One participant compiled metadata of documents such as news articles to organize them and make them more manageable in later stages of the research process.

> [. . .] when collecting data, you never really know what metadata you might need someday [. . .] like, the original URL and dates. And, for example, if the name of the author is missing, then you try to remember to get it somewhere [. . .] and [. . .] the date and title that they go in the same order in the files [. . .] to speed up the reading when you do the analysis. [. . .] it is often laborious, the kind of collecting and processing data [. . .] that it becomes easy to analyze. [P14]

In addition to metadata such as "the publication date and the title of the [magazine article] and [. . .] the section [of the magazine]", one participant wrote "a kind of synopsis" of each article to organize and become more familiar with the research data. This kind of self-created, descriptive metadata may prove particularly important if the researcher loses access to the original data. This can happen, for example, when terms of use prohibit saving the material for research use. Another participant talked about such challenges in gathering social media data:

> [. . .] the [social media] channels are closed all the time [. . .] we no longer have access to these [contents] themselves, although we have the URLs [. . .] and the descriptions [of the contents] and all the metadata. [P19]

4.1.4 Secondary information about data. Sometimes, the research data that media scholars were interested in gathering could not be found or no longer existed. A participant described challenges of studying historical webpages that are poorly documented:

> One cannot go back to what [a website] looked like 5 years ago [. . .] it affects directly how a website can be analyzed, because more often than not one would like to see what it was like in previous years and, roughly, what has changed. [P23]

However, one possibility is to rely on secondary information. This means searching for other sources that would tell something about the lost items, which is like detective work according to two participants. A participant, who used films as research data, described how secondary sources might be the only way to find any information about some old films that seem to be lost forever.

> All release prints and all original negatives [of the films] are lost at least according to current knowledge [. . .] one must turn to all other material that can be found, photographs, or manuscripts, or drafts [. . .] so, that has now been perhaps the clearest challenge here. [P22]

4.2 Information about sources of research data

Participants gathered research data from various sources. Examples of these are newspapers as sources of news articles, discussion forums as sources of discussion forum data and social media platforms as sources of social media data. Participants needed information about characteristics of the sources and about local media landscapes the sources were a part of.

4.2.1 Characteristics of sources. Characteristics of sources refer to the properties and aspects that describe the sources. Characteristics of sources were used as selection criteria for the sources of the research data. Characteristics were also mentioned in the sense that they were acknowledged but not used as selection criteria for the sources. Information about the characteristics of the sources was also needed to describe the sources in a research publication.

Media scholars selected sources that focused on or came from a certain *geographical area*. For example: one participant selected a local newspaper as a data source because of its local perspective and its ties to the local area; some participants selected social media accounts of people from specific countries as sources of social media data. *Genre* (e.g. daily newspapers vs. tabloids) and *format* (e.g. digital vs. paper format) were also mentioned as characteristics of sources. *Age and publication frequency* of sources describe the length of time the sources have existed and how actively content is published in them. For example, one participant selected long-running newspapers as sources of journalistic articles because this made it possible to gather articles from several years. Another participant, who searched for social media accounts as sources for social media posts, needed to determine how active the accounts were (e.g. were there any posts from a certain period of time). Sources were also described in terms of their *media reach and publicity*, in other words, how many people the sources reached and were they freely available and public to everyone. For newspapers, participants mentioned circulation of the papers and readers' access to online content (e.g. was the content free or only available to paying subscribers). For social media platforms, participants needed to differentiate public user accounts from private ones for ethical reasons. Participants also mentioned information that they did not have, such as the number of users in social media platforms.

> These platforms themselves are causing the problem [. . .] they do not reveal their certain logics or even necessarily how many people there are. [P3]

Ownership refers to information about who owned the sources (e.g. which media group a newspaper belonged to) and were the owners commercial or public service companies. Information about possible political ties of newspapers (historically or today) was also mentioned. Participants also talked about *conditions for discussion and content creation* in discussion forums, social media platforms and comment sections of online newspapers. They mentioned characteristics such as were the discussions conducted anonymously, was user registration required, was it possible to identify which posts were from the same users (e.g. based on stable usernames) and were the discussions moderated. Some participants also commented that the platforms' terms of use affected what kind of content the users were allowed to post in the first place.

Lastly, *accessibility and usability* of sources affected data gathering. Many participants appreciated digital archives and their search functions in finding research data, although sources in physical form were also used. A participant, who searched for data sets from digital archives, put a strong emphasis on immediate availability of data:

> I want something that can be used immediately, or like easily. That I first wouldn't have to negotiate something for a year or collect some data laboriously myself. So, I have been thinking even on those terms how this research question could be studied. [P18]

4.2.2 Local media landscapes. Participants needed information about the local media landscapes the sources were a part of, for example, about the types of newspapers or media organizations in different countries. One participant elaborated that local knowledge is essential when studying foreign newspapers. The participant had local research partners who had a good understanding about the local politics and the media field of their country:

> I thought that because the research focuses on [a foreign country] [. . .] it is good to have this local partner [. . .] who knows the context better and the language. [P9]

The need for understanding media landscapes was also expressed by another participant, who wanted to gather comparable data from different countries:

> These [. . .] countries have different media systems [. . .] we decided on [particular media channels] because they belonged to the same journalistic genre, in which case the way of doing journalism would not affect that much [to the news content]. [P21]

4.3 Information about cases and their contexts

Some participants gathered research data related to a case (e.g. to some example of an event or a phenomenon) that were of interest in their research. One participant said, "it is terribly typical in media studies that there is a case". Participants used information about those cases (and their contexts) to determine what research data to gather and to contextualize the objects of study. In this study, we call these case information and contextual information.

4.3.1 Case information. Case information means information about an example of something occurring that is of interest. This included information about course of events, timelines and lifespans of the cases. The information was used in data gathering. For example, some participants, who gathered news articles or online discussions related to a case, identified critical periods or turning points of the events which helped them determine the time frame for data collection.

> I tried to cut it to somewhere like, from the perspective of the process [of the case], to a turning point [. . .] I ended the data collection there that the process [of the case] ended in a certain way. And then the same in these others, we looked for these kinds of corresponding turning points in the process, and in them we included all the articles that met the search criteria [. . .] [P2]

One participant explained that the idea was to capture the entire lifespan of a particular case by collecting a series of articles from specific time periods throughout its lifespan. Another participant anticipated that making a timeline of a case would help seeing the big picture and would also help in organizing research materials related to the case.

> Because I think that if I continue making this timeline [about the case], then, in a way, it also supports my thing with the ethnographic research diary, which is then in a more systematic form. [P3]

4.3.2 Contextual information. Contextual information means information that is conceptually connected to circumstances of a case. Sometimes, the need for contextual information was intrinsic to media scholars' research interests and methods, as illustrated in the following comments.

> [. . .] interpreting a film alone didn't feel meaningful, it needed [. . .] historical and production context. [P22]

> [. . .] methods in media history, for example, archival work and [. . .] contextualizing [. . .] where [you] try to understand the object by combining different kinds of materials. [P25]

Media historians, for example, emphasized the importance of gathering and comparing various types of historical documents and contextualizing the objects of study. Doing media history involved cross-reading different kinds of materials that complemented each other. Media products (e.g. films) were studied in their broader contexts such as how the media products were marketed, reviewed, or discussed in the media. In these cases, media scholars gathered, for example, reviews, PR materials, interviews or other works (in text, audio, visual or audiovisual form) as research data. The media products worked as conceptual starting points when gathering such data. Media products were also studied in the context of their production. This information could be found, for example, in documents produced by organizations (e.g. action plans, reports of meetings).

5. Discussion

In this study, we aimed at answering the research question: What task information types do media scholars need while gathering research data to create new knowledge? We found three

task information types containing a total of eight subtypes. The first task information type was information about research data that included aboutness of data, characteristics of data, metadata and secondary information about data. The second task information type was information about data sources that included characteristics of sources and information about local media landscapes the sources were a part of. The third task information type was information about cases and their contexts. It included case information (information about an example of something occurring that is of interest) and contextual information (information that is conceptually connected to circumstances of a case).

Our research revealed that the data were collected from various sources, and that users desire specific content and appreciated the ability to manage and annotate the data. The research shows new typology for assessing the usefulness of research data as criteria for data gathering. Typically, usefulness is defined as some sort of topicality of documents. Here, we show that several sub-types of information about the data, information about the sources and information about the cases are critical to the data gathering activity and should be considered in evaluating usefulness of information. Keeping the recent efforts in providing open data sets for data reuse and ensuring research reliability in mind, it is important to know what the requirements for data uses in real research work are.

Concerning the first task information type, information about research data, Koesten *et al.* (2017) found similarly that people needed information about provenance, completeness or granularity when selecting data sets. They also found that people tried to understand a new data set by exploring it, for example, by scrolling through it, looking for errors, filtering relevant data or visualizing to see trends or peaks. Data gathering activity may encompass use of visualizations and other data exploration methods for revealing the aboutness of the data such as organizing them using topic modeling. Thematic and topical aspects are inherent particularly to self-created annotations (cf. Melgar *et al.*, 2017). The exhaustivity or timely coverage of the data, especially in terms of temporal aspects, could be revealed by showing the data in a timeline. Further, metadata has been found as important to accountability of research and data as evidence (Mayernik, 2019). According to them, the production may be invisible and informal, especially in the case of self-created annotations. The processes affect the outcomes of the research. If its production is properly supported, the metadata should be interoperable with other data sets and analysis tools. Mayernik (2019) also considers metadata processes and products important in supporting day-to-day research tasks and meeting the requirements of funders, journals and data repositories. Kumpulainen and Late (2022) found that lacking or improper metadata was one obstacle to successful uses of newspaper contents in the history research domain.

The second type, information about sources of research data, included characteristics of the source. Newspaper ownership was one such characteristic that had an effect on the contents. Similarly, social media moderating processes fundamentally shape the digital social and political spheres, and the moderators been shown to have power over the contents (Malinen, 2021). We found also that it was important to the media scholars to understand how the local media landscapes shape the media contents. The quality standards of media and how they change vary across nations (c.f., Kuipers, 2011).

Regarding the last type, information about cases and their contexts, we found that the need for contextual information was an integral part of the media scholars' data gathering. Similarly, Bron *et al.* (2016) identified a contextualization phase in media scholars' research cycle, referring to targeted data gathering, where media scholars looked for additional data sources to provide context for their research. Finding contextual information could be supported by showing conceptual relationships across media texts, other documents and data sets.

Previous research about task information has not been considering the activity of research data gathering. Byström (1999) was interested in municipal officials' work, and it has

changed significantly during the last decades (Saastamoinen *et al.*, 2012). Further, there was no actual report about the types in the empirical research, but rather theoretical definition of the concept and discussion about the sources of information that were characterized as information types.

The data were very rich and provided extensive accounts of the data gathering activities. According to Mason (2018, p. 112), qualitative interviewing can be used as a research method when "instead of asking abstract questions, or taking a 'one-size-fits-all' structured approach, you may want to give maximum opportunity for the construction of contextual knowledge by focusing on relevant specifics in each interview". By conducting interviews instead of, for example, questionnaires with closed-ended questions, we were able to avoid limiting the participants' narratives too much in advance. Furthermore, because Byström's (1999) definition of task information was quite general, we wanted to use qualitative methods to find out what it could mean in the context of data gathering. In analysis, differentiating the activity of gathering research data (including searching, selecting and collecting) from other activities was sometimes difficult because of the iterative and intertwined nature of the real-world research activities. However, this was overcome by scrutinizing the codes and discussing them within the research group in each round of coding.

The typology is not necessarily domain specific, but this should be tested in future studies. Especially the three main types (information about research data, information about sources of research data, information about cases and their contexts) could be generalizable to other situations. They could be understood more broadly as information about any information items that are gathered for some specific purpose, information about the sources of the information items and information about the matter – what it is about and what is the context – for which the items are gathered. On the other hand, the typology may in some respects be specific to gathering research data and to media studies. For example, characteristics of data are the properties and aspects that participants wanted to know about their research data, and the characteristics may be different to other kinds of information items.

The information types analyzed are based on real-world research tasks, which allow investigating the phenomena holistically. This is important to avoid too narrow research settings (Järvelin *et al.*, 2015). In future studies, the research perspective could be broadened to task information needed in other research activities (other than data gathering), as well as to study what domain and task-solving information are needed in different research activities.

6. Conclusion

Radical changes in media landscape have affected the media scholars' work. These developments require a holistic view to understand how to support knowledge creation processes. We examined what task information media scholars needed while gathering research data to create new knowledge by using information interaction approach. The key finding was that media scholars needed information about research data, sources of research data and cases and their contexts. All the task information types should be considered when building data services and tools to support media scholars' work. The typology created in this study can be used for assessing the usefulness of research data. Furthermore, the research elaborated the concept of task information by characterizing its subtypes.

References

Allen, D., Karanasios, S. and Slavova, M. (2011), "Working with activity theory: context, technology, and information behavior", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 4, pp. 776-788, doi: 10.1002/asi.21441.

Barr, A. and Feigenbaum, E. (Eds) (1981), *Handbook of Artificial Intelligence*, Vol. I, Pitman, London.

Borgman, C.L. (2007), *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, The MIT Press, Cambridge, MA.

Borgman, C.L. (2015), *Big Data, Little Data, No Data: Scholarship in the Networked World*, The MIT Press, Cambridge, MA.

Bron, M., Van Gorp, J. and de Rijke, M. (2016), "Media studies research in the data-driven age: how research questions evolve", *Journal of the Association for Information Science and Technology*, Vol. 67 No. 7, pp. 1535-1554, doi: 10.1002/asi.23458.

Byström, K. (1999), *Task complexity, information types and information sources: examination of relationships*, Academic dissertation, University of Tampere, Tampere, ISBN: 978-952-03-1893-2, available at: https://urn.fi/URN:ISBN:978-952-03-1893-2

Byström, K. (2009), "Information activities in work tasks", in Fisher, K.E., Erdelez, S. and McKechnie, L. (Eds), *Theories of Information Behavior*, Information Today, NJ, pp. 174-178.

Byström, K. and Hansen, P. (2005), "Conceptual framework for tasks in information studies", *Journal of the American Society for Information Science and Technology*, Vol. 56 No. 10, pp. 1050-1061, doi: 10.1002/asi.20197.

Byström, K. and Järvelin, K. (1995), "Task complexity affects information seeking and use", *Information Processing and Management*, Vol. 31 No. 2, pp. 191-213, doi: 10.1016/0306-4573(95)80035-R.

Byström, K. and Kumpulainen, S. (2020), "Vertical and horizontal relationships amongst task-based information needs", *Information Processing and Management*, Vol. 57 No. 2, 102065, doi: 10.1016/j.ipm.2019.102065.

Choi, B., Ward, A., Li, Y., Arguello, J. and Capra, R. (2019), "The effects of task complexity on the use of different types of information in a search assistance tool", *ACM Transactions on Information Systems*, Vol. 38 No. 1, p. 28, Article 9, doi: 10.1145/3371707.

Crescenzi, A., Li, Y., Zhang, Y. and Capra, R. (2019), "Towards better support for exploratory search through an investigation of notes-to-self and notes-to-share", *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*, Association for Computing Machinery, New York, NY, pp. 1093-1096, doi: 10.1145/3331184.3331309.

Faniel, I.M., Frank, R.D. and Yakel, E. (2019), "Context from the data reuser's point of view", *Journal of Documentation*, Vol. 75 No. 6, pp. 1274-1297, doi: 10.1108/JD-08-2018-0133.

Fidel, R. (2012), *Human Information Interaction: An Ecological Approach to Information Behavior*, Mit Press, Cambridge, Massachusetts.

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

Gibbs, F. and Owens, T. (2013), "The hermeneutics of data and historical writing", in Dougherty, J. and Nawrotzki, K. (Eds), *Writing History in the Digital Age*, University of Michigan Press, pp. 159-170, doi: 10.2307/j.ctv65sx57.18.

Gray, J. (2017), "Text", in Ouellette, L. and Gray, J. (Eds), *Keywords for Media Studies*, NYU Press, New York, pp. 196-200.

Ingwersen, P. and Järvelin, K. (2005), *The Turn: Integration of Information Seeking and Retrieval in Context*, Springer Netherlands, Dordrecht.

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, p. 30, Article 3, doi: 10.1145/2699660.

Jensen, K.B. (Ed.) (2012), *A Handbook of Media and Communication Research: Qualitative and Quantitative Methodologies*, 2nd ed., Routledge, London.

Koesten, L.M., Kacprzak, E., Tennison, J.F.A. and Simperl, E. (2017), "The trials and tribulations of working with structured data – a study on information seeking behaviour", *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*, Association for Computing Machinery, New York, NY, pp. 1277-1289, doi: 10.1145/3025453.3025838.

Koivisto, J. and Thomas, P.D. (2010), *Mapping Communication and Media Research: Conjunctures, Institutions, Challenges*, Tampere University Press, Tampere.

Kuipers, G. (2011), "Cultural globalization as the emergence of a transnational cultural field: transnational television and national media landscapes in four European countries", *American Behavioral Scientist*, Vol. 55 No. 5, pp. 541-557, doi: 10.1177/0002764211398078.

Kumpulainen, S. and Late, E. (2022), "Struggling with digitized historical newspapers: contextual barriers to information interaction in history research activities", *Journal of the Association for Information Science and Technology*, Vol. 73 No. 7, pp. 1012-1024, doi: 10.1002/asi.24608.

Late, E. and Kumpulainen, S. (2022), "Interacting with digitised historical newspapers: understanding the use of digital surrogates as primary sources", *Journal of Documentation*, Vol. 78 No. 7, pp. 106-124, doi: 10.1108/JD-04-2021-0078.

Lee, D. and Zoellner, A. (2019), "Media production research and the challenge of normativity", in Deuze, M. and Prenger, M. (Eds), *Making Media: Production, Practices and Professions*, Amsterdam University Press, Amsterdam, pp. 45-59, doi: 10.5117/9789462988118.

Leonelli, S. (2015), "What counts as scientific data? A relational framework", *Philosophy of Science*, Vol. 82 No. 5, pp. 810-821.

Lester, J.N. and O'Reilly, M. (2019), *Applied Conversation Analysis: Social Interaction in Institutional Settings*, SAGE Publications, Inc., Thousand Oaks, CA, doi: 10.4135/9781071802663.

Li, Y., Ward, A.R. and Capra, R. (2021), "An analysis of information types and cognitive activities involved in cross-session search", *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR '21)*, Association for Computing Machinery, New York, NY, pp. 313-317, doi: 10.1145/3406522.3446044.

Long, P., Johnson, B., MacDonald, S., Rogerson Bader, S. and Wall, T. (2022), *Media Studies: Texts, Production, Context*, 3rd ed., Routledge, London.

Malinen, S. (2021), "The owners of information: Content curation practices of middle-level gatekeepers in political Facebook groups", *New Media and Society*, Advance online publication, doi: 10.1177/14614448211062123.

Mason, J. (2018), *Qualitative Researching*, 3rd ed., SAGE, Los Angeles.

Mayernik, M.S. (2019), "Metadata accounts: achieving data and evidence in scientific research", *Social Studies of Science*, Vol. 49 No. 5, pp. 732-757, doi: 10.1177/0306312719863494.

Melgar, L., Koolen, M., Huurdeman, H. and Blom, J. (2017), "A process model of scholarly media annotation", *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*, Association for Computing Machinery, New York, NY, pp. 305-308, doi: 10.1145/3020165.3022139.

Miles, M.B., Huberman, A.M. and Saldaña, J. (2020), *Qualitative Data Analysis: A Methods Sourcebook*, 4th ed., International student edition, SAGE, Los Angeles.

Patriarche, G., Bilandzic, H., Jensen, J.L. and Jurišić, J. (Eds) (2014), *Audience Research Methodologies: Between Innovation and Consolidation*, Routledge, New York, London.

Roulston, K. (2010), *Reflective Interviewing: A Guide to Theory and Practice*, SAGE Publications, London, doi: 10.4135/9781446288009.

Rowley, J. (2014), "Data analysis", in Coghlan, D. and Brydon-Miller, M. (Eds), *The SAGE Encyclopedia of Action Research*, Vols 1-2, SAGE Publications, pp. 239-242, doi: 10.4135/9781446294406.n102.

Saastamoinen, M., Kumpulainen, S. and Järvelin, K. (2012), "Task complexity and information searching in administrative tasks revisited", *Proceedings of the 4th Information Interaction in Context Symposium (IIIX '12)*, Association for Computing Machinery, New York, NY, pp. 204-213, doi: 10.1145/2362724.2362759.

Zhang, Y., Capra, R. and Li, Y. (2020), "An in-situ study of information needs in design-related creative projects", *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, Association for Computing Machinery, New York, NY, pp. 113-123, doi: 10.1145/3343413.3377973.

Appendix
Interview guide for semi-structured interview

A. Background information

(1) What is your educational background?

(2) What is your current job title?

(3) How long have you worked as a researcher?

(4) What is your field of research?

B. Research topic

(1) What is your research topic? If possible, select an ongoing or recently completed research that you can still remember well.

C. Research process

(1) Where did you get the idea for the research topic?

(2) What are your research goals?

(3) What are your research questions?

(4) Can you distinguish phases from your research process? What are they? Where are you now in this continuum?

D. Working in a research group (if applicable)

(1) What is your research group like?

(2) What is your role in the research group?

(3) What is the division of labor in the group?

(4) Do you have common research data?

(5) Do you have common tools?

E. Research data

Description of the research data

(1) What is your research data like?

(2) In what form is your research data?

Collecting the research data

(1) What are your research methods?

(2) How did you collect the research data?

(3) Did you collect the research data in one or several sessions?

(4) Where do you keep your research data?

(5) How do you organize your research data?

(6) How do you know when you have enough data?

(7) Did you have all the necessary information you needed to be able to use the research data in your research?

(8) Have there been any difficulties getting the research data for research purposes?

Finding the research data

(1) Where did you find the research data?

(2) How did you know where to look for the research data/participants for the research?

Selecting the research data

(1) Why (and how) did you choose the research data?

Analyzing the research data

(1) How do you analyze the research data?

(2) Did you start analyzing the data before all was collected?

(3) Did the data require preprocessing for the analysis?

Archiving the research data

(1) Have you thought about what to do with the research data after the research is completed?

Managing the research data

(1) Have you planned your data management beforehand?

(2) How about during the research?

Research ethics, ownership and licensing

(1) What ethical questions did you need to think about concerning your research data?

(2) Were there any issues related to ownership and licensing of the research data?

Corresponding author
Laura Korkeamäki can be contacted at: laura.korkeamaki@tuni.fi