Data-Seeking Behaviour in the Social Sciences

Krämer, Thomas; Papenmeier, Andrea; Carevic, Zeljko; Kern, Dagmar; Mathiak, Brigitte

International Journal on Digital Libraries
https://doi.org/10.1007/s00799-021-00303-0

Data-Seeking Behaviour in the Social Sciences

Thomas Krämer[1] · Andrea Papenmeier[1] · Zeljko Carevic[1] · Dagmar Kern[1] · Brigitte Mathiak[1]

Received: 23 July 2020 / Revised: 30 March 2021 / Accepted: 2 April 2021

Abstract
Purpose Publishing research data for reuse has become good practice in recent years. However, not much is known on how researchers actually find said data. In this exploratory study, we observe the information-seeking behaviour of social scientists searching for research data to reveal impediments and identify opportunities for data search infrastructure.*Methods* We asked 12 participants to search for research data and observed them in their natural environment. The sessions were recorded. Afterwards, we conducted semi-structured interviews to get a thorough understanding of their way of searching. From the recordings, we extracted the interaction behaviour of the participants and analysed the spoken words both during the search task and the interview by creating affinity diagrams.*Results* We found that literature search is more closely intertwined with dataset search than previous literature suggests. Both the search itself and the relevance assessment are very complex, and many different strategies are employed, including the creatively "misuse" of existing tools, since no appropriate tools exist or are unknown to the participants.*Conclusion* Many of the issues we found relate directly or indirectly to the application of the FAIR principles, but some, like a greater need for dataset search literacy, go beyond that. Both infrastructure and tools offered for dataset search could be tailored more tightly to the observed work processes, particularly by offering more interconnectivity between datasets, literature, and other relevant materials.

Keywords Data search · Dataset retrieval · Social science · User behaviour · User study · Information-seeking behaviour

1 Introduction

Publishing research data along with the primary publication is becoming good practice or even mandatory[1] in academic research. Making research replicable and discovering information that remained unmentioned or uncovered in the primary analysis is a driver for publishing data [33]. Aggregating, replicating, or applying different methods to existing data lead to new insights and increase the quality of the underlying research results. However, frequencies of data reuse are widely unknown or assumed to be low [8,10]. The GO FAIR Initiative[2] coordinates efforts to support data reuse by increasing the findability, accessibility, interoperability, and reusability of digital assets in science. Finding datasets that fit a research question at hand is essential for data reuse. Darby et al. [11] describe drivers and barriers to scientific data sharing and reuse. For data discovery, in particular, they identify a lack of infrastructure to support international

✉ Thomas Krämer
thomas.kraemer@gesis.org

Andrea Papenmeier
andrea.papenmeier@gesis.org

Zeljko Carevic
zeljko.carevic@gesis.org

Dagmar Kern
dagmar.kern@gesis.org

Brigitte Mathiak
brigitte.mathiak@gesis.org

[1] GESIS - Leibniz-Institute for the Social Sciences, Cologne, Germany

[1] Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.

[2] https://www.go-fair.org/go-fair-initiative/.

Published online: 18 April 2021

cross-disciplinary data discovery. Although data search portals aiming to provide easy access to datasets launched in recent years[3], little is known about the role of dedicated data search portals in the researchers' information-seeking process. Only a few studies exist that solely focus on users' data-seeking practices. Often, this research is conducted in the context of data reuse [16]. The predominant research methods to examine data-seeking behaviour are online surveys [13], semi-structured or in-depth interviews either in person or via telephone [16,21], as well as analysing log files [23]. The studies to date offer insights into the process of data retrieval from a retrospective viewpoint based on what participants reported. In a think-aloud study, Murillo [30] examined what factors influence decisions regarding data reuse in the context of data retrieval processes. However, this study was limited to the interaction with a specific data archive, the DataONE system, and did not focus on the whole data-seeking process.

Nothing is known so far about the actual and system-independent procedure of data-seeking sessions. Keeping the first role of usability from Jakob Nielson in mind "Don't listen to Users. To design the best UX, pay attention to what users do, not what they say"[4], we address this gap by conducting an observational study, in which 12 participants searched for data they need. There were no instructions on how to search for data. The participants were free to search online, contact personal networks, or search personal archives on their computers or university servers. The whole session was recorded and followed by semi-structured interviews to get a fine-grained understanding of what they did.

Our research is guided by the following questions:

- How is dataset search performed in the "wild" and what are the challenges that researchers encounter during dataset search?
- How effective are observations for studying data-seeking behaviour?
- How do our findings relate to higher level discourses, such as FAIR[43] and data literacy?

As in the social sciences, data sharing and reusing have a long tradition [17] and we experienced that scientists still perceive that a considerable effort is required to discover data for potential reuse [21], we choose social sciences as our research context. As social sciences are a broad research field with disciplines such as sociology, demography, ethnology, political science, education, psychology, communication studies, economics, social policy, interdisciplinary and applied social sciences, and a larger number of related areas, we narrowed it further down to empirical social research in which researchers mainly work with big surveys, statistical data, and similar data sources. The restriction is necessary to ensure comparability between the participants and to be able to form meaningful results rather than isolated observations.

With our research, we contribute to the field of dataset-seeking behaviour by providing insights based on real dataset-seeking processes. We observed and interviewed 12 social scientists in their usual working environment. We analysed the participants' interaction steps and used affinity diagrams to cluster the findings from the observations and the answers given in the semi-structured interview. Four clusters emerged: (1) the *context* in which the data search process takes place, (2) the *resources* used to find datasets, (3) the applied *working methods*, and (4) information on the *data assessment*. One further cluster summarised *general feedback*. From these results, we derived recommendations on how to improve the dataset-seeking process.

2 Background and related work

In this section, we contextualise our work to the field of information seeking in general and more precisely to the field of data-seeking behaviour. We provide an overview of other user studies in this context and describe how we further contribute to the research field.

2.1 Information seeking and search activities

Research on information seeking in web search and digital libraries has a long tradition [7]. The theoretical foundations of information seeking evolved throughout the past decades: Today, numerous models of the information-seeking process exist, which, among others, emphasise the process-oriented [12,24,29] and strategic [2,3] aspects of the information-seeking process. Traditional models of the information-seeking process are usually rooted in the area of document retrieval or web search in general. Hence, the applicability of traditional models of the information-seeking process in the area of dataset retrieval is uncertain. For instance, the stages of the information-seeking process [12] described by Ellis have been widely accepted throughout the literature, but whether these stages are representative for the process of dataset retrieval is unknown. A more fine grained view on different stages of the information-seeking process during dataset retrieval is so far not available.

From a broad perspective, one can distinguish two types of search activities: lookup and exploratory. Lookup activities cover tasks such as known-item search and question answer-

[3] EUDAT B2Find https://eudat.eu/services/b2find, Elsevier DataSearch, https://datasearch.elsevier.com, Google Dataset Search https://toolbox.google.com/datasetsearch, Socrata https://dev.socrata.com.

[4] https://www.nngroup.com/articles/first-rule-of-usability-dont-listen-to-users/.

ing that are well supported by current systems. Exploratory search, on the other hand, comprises activities such as learning and investigating [27]. These searches aim to acquire new knowledge, require more time, and involve cognitive processing (e.g. comparing and making qualitative judgments). The information needs motivating an exploratory search are generally open-ended, persistent, and multifaceted. Open-endedness leads to affective and cognitive uncertainties which are observable during an exploratory search. Exploratory searchers experience uncertainties about the information available or incomplete information about the nature of the tasks [41]. Exploratory search tasks, on the other hand, are suitable to describe and emphasise characteristics of complex search tasks and, thus, certainly relevant for dataset retrieval tasks. An explicit focus on the exploratory nature of the dataset-seeking processes, however, is missing so far.

2.2 Studying exploratory search

Designing a study that induces an exploratory search is challenging due to open-ended and highly subjective information needs [25]. Usually, researchers employ exploratory studies in such situations. The usage of simulated work tasks [5] in exploratory studies helps to frame the experimental condition [42]. Exploratory studies often combine different research methods to study the effects or phenomena of information-seeking behaviour [20]. The methods include, among others: questionnaire surveys, semi-structured interviews, transaction logs, and (unobtrusive) observations of search behaviour. Kules et al. [26] even applied eye-tracking combined with stimulated recall interviews to learn what parts of a faceted interface searchers attend to, for how long, and in what order. Research questions posed during exploratory studies are often broad and open-ended. The present work is designed as an exploratory study given our goal of obtaining a fine-grained understanding of the dataset-seeking process.

2.3 Data-seeking behaviour

Research of information-seeking behaviour commonly focuses on the web in general [6,22,40]. Many activities can also be observed in the field of academic literature search [31,34,35]. Some of the findings from literature search might also be relevant or transferable to dataset search. Keeping up-to-date, exploring new and unfamiliar topics, reviewing literature, collaborating with other researchers, preparing lectures, and recommending material for students are purposes that motivate literature search [1] that might also be applicable for dataset search. Furthermore, the sources for literature search like academic journals, web pages, and online databases [18] might also be starting points or sources for finding datasets. Not least, the effect of topic familiarity on resources and relevance criteria that plays a role in literature search [39] might be well transferable to dataset search. However, recent research emphasises the differences between literature search and dataset search [8,9,13,21]. Kern et al. [21], for example, found that browsing through results in a dataset search is more complex and thus more time-consuming than in a literature search. This complexity originates from the large number of materials related to a dataset (e.g. in the social sciences, a dataset can comprise data records, questionnaires, method reports, and codebooks).

As a result of the GO FAIR Initiative, sharing and reusing datasets have become more popular, and research activities in the field of dataset retrieval have increased. However, so far, only a few studies focus on dataset retrieval. Instead, existing research concentrates on data sharing, reuse practices, or user studies for particular repositories [15]. Recently, Chapman et al. [8] published a comprehensive state-of-the-art survey on dataset search including an overview of current dataset search implementations, research activities in this field, and open problems. They identified open issues and challenges in the context of query languages (suggesting going beyond keyword search), differentiated access to datasets, linking between datasets and external knowledge as well as result presentations. While their survey is very comprehensive concerning the technical aspects of dataset search, the users with their needs and seeking behaviours are only marginally considered.

In contrast, Gregory et al. [15] focused more on the user side and provided a review of observation data retrieval practices. They reviewed nearly 400 publications on data search and data discovery. They focus on data search practices in different disciplines (astronomy, earth, and environmental sciences, biomedicine, field archaeology, and social science). For each discipline, they assess the user needs, the user actions, and the dataset evaluation criteria. They question that current information retrieval models sufficiently describe data retrieval practices at the moment. Identifying the importance of personal connections and networks leads them to the conclusion that "it is not enough to understand data retrieval as a series of interactions between users and search systems; rather, data retrieval is, in fact, a complex socio-technical process" [15]. The importance of personal networks, collaboration, and connections to main contributors is also supported by findings of [13,28,44]. Having this in mind, we designed our study accordingly to be able also to observe these interpersonal aspects in the dataset-seeking process.

Only a few user studies are reported in the literature that focus on data-seeking behaviour, albeit none that use observation as their primary methodology. Friedrich [13] followed a mixed-method approach and developed an online questionnaire based on expert interviews. 1388 social scientists answered this questionnaire, and the results showed that common strategies are forward chaining from journal articles

and personal contacts. She also found that individual characteristics, community involvement, and experience play a role. This confirmed us to recruit participants with different levels of experience. Semi-structured interviews with 22 participants were conducted by Gregory et al. [16]. They categorised their findings along three dimensions: (1) user contexts and data needs (including the purpose), (2) search strategies (including resources for seeking data, methods to locate data and success), and (3) evaluation criteria (including social interaction, information need to evaluate data, and information used to establish trust and data quality) and pointed out that the interplay between technology and social practice makes data search a complex phenomenon. Furthermore, based on their results, they offer some suggestions for designing retrieval systems, e.g. improving metadata standardisation in general or means to enable social interactions with the data authors or others who use the same data. Koesten et al. [23] also conducted semi-structured in-depth interviews with 20 participants and analysed user logs of data.gov.uk. One main issue raised by many participants is the difficulty finding the right data, either because no dataset could be found that matched the researcher's demand or because existing tool support was not sufficient. Most participants use Google to find data online and ask colleagues or other people working in the respective field for dataset suggestions. Since interviews can only elicit information that participants can recall easily, we aim with our study on uncovering information of data-seeking processes that may not have been mentioned before.

Murillo [30] applied a similar approach to ours and conducted a pilot think-aloud study in the context of the DataONE system[5]. She observed users while using the system and focused mainly on aspects in the context of data reuse, e.g. what information is needed to decide whether a dataset is reusable or not. She found that the participants made their decisions based on the metadata snippets provided by the system. system. As her results are based on a specific system, the applicability to understand the data-seeking process in general is limited.

2.4 Dataset search in the context of dataset reuse

As dataset search is a part of the dataset reuse process, research in this field also provides insights into dataset search. For example, Pasquetto et al. [32] researched how scientists find reusable data, how they reuse those data, and how they interpret this data. Their results base on several observational studies, interviews, and literature reviews over a period of two decades including different domains like sensor networks, environmental or earth science. They conclude that data reuse is a process in which researchers consider more datasets than they finally reuse. Expertise and trust play important roles in this process, and here again, researchers prefer to collaborate with the data creators.

To investigate the different factors that influence data reuse, Curty [10] conducted interviews (face-to-face, Skype calls, phone) with 13 social scientists. She introduces a conceptual model including six theoretical variables: perceived benefits, perceived risks, perceived effort, social influence, facilitating conditions, and perceived reusability. Regarding data discovery, she emphasises the effort researchers must invest to find the right data, not only in the discovery process but also in requesting access to data.

2.5 Summary

The vast majority of studies presented above draw their results from retrospective data. Participants report their experience with dataset search in interviews or online questionnaires. Our exploratory study, instead, focuses on the actual search process. We observe users while searching for datasets to uncover aspects that cannot be revealed by interviews. We analyse the interaction paths, thoughts, and verbal feedback recorded while searching for data.

3 Methodology

We conducted an explorative study with 12 participants to gain insights into social scientists' information-seeking behaviour during dataset search.

3.1 Procedure

We visited the participants at their workplace or at a dedicated room at their university to ensure they conduct the search task with their usual means, including their own technical equipment.

Participants were informed verbally about the purpose of the study, the experiment, the procedure and data collection. Then, the description of data collection and privacy regulations was handed out and participants were requested to sign a consent form after reading it.

As searching for data is not a daily activity of social scientists, we used a simulated work task inspired by Borlund [5] to create a simulated but realistic scenario. We confronted the participants with the following task description: "In the context of your research, you need research data. For today, you decide to start with the search for research data"[6]. We clarified our understanding of "research data" by provid-

[5] https://www.dataone.org/.

[6] All studies and interviews were conducted in German. The authors provided all translations for instructions, survey questions, and direct quotes from the interviews.

ing the following definition: "Research data is data that is generated in the course of scientific projects, e.g. through digitisation, source research, experiments, measurements, surveys or questionnaires". We asked participants to think aloud during the task. We let them know that they could take as much time as needed, leave the room if needed, and contact other persons by email or phone. After the search task, participants described their demographic background and their knowledge of research data based on their previous research activities. Furthermore, participants used a 5-point Likert scale (ranging from 1 "very easy" to 5 "very difficult") to rate the perceived difficulties of both dataset search and literature search. Subsequently, in a semi-structured interview, we further discussed observations from the search session with the participant. We were especially interested in getting more insights into their usual data search workflow, criteria to select a dataset, and aspects they miss in the current data search process.

We recorded the audio of the entire sessions and captured the screencasts of the participants' monitor using either high-resolution digital cameras or using WebRTC to record the browser content. Both the videos and the interviews were later transcribed. We thanked the participants by compensating their effort with 30 EUR. The sessions started presenting the search task and the invitation to the participant to start the data search. The duration of recorded sessions including the semi-structured interview and the questionnaire ranged from 45 to 116 minutes, with a median of 69 minutes.

This procedure was pretested with two subjects to clarify the instructions of the simulated work task, the operation of the technical set-up (voice recorder, screen capture), and the interviewer guideline for the semi-structured interview.

3.2 Sampling

We decided to populate our sample with equal numbers of early career and late-career researchers, ranging from Master's students to full professors. The prerequisites for taking part in the study were a background in social sciences and being a native German speaker. We recruited the participants through posters at the local university and by sending invitation emails to a list of people who, in an earlier study, consented to be contacted again for research purposes. We further contacted scholars teaching data analysis directly to ask them to participate and to motivate their students as well.

3.3 Analysis methods

This study provides insights into participants' behaviour and perception throughout the search task and the semi-structured interview. Both aspects are analysed separately.

3.3.1 Interaction analysis

To get an overview of the participants' activities, we decided to use a simple coding scheme according to the type of website they were interacting with or specific actions they perform with an information type (see Fig. 2). We distinguish between Google, Google Scholar, library catalogue or literature portal, data portal[7], a dedicated project website to a survey program (e.g. the website of the International Social Survey Programme), reading a document, reading additional material (like codebooks, questionnaires, reports), and downloading data. We base the interaction analysis on the screencasts recorded during the simulated work task. Unfortunately, in two cases, the quality of the recording was too bad to code reliably, so we coded the interactions only from 10 out of 12 sessions.

3.3.2 Affinity diagrams

We used the affinity diagrams method [37] to cluster participants' feedback during their search for datasets. We transcribed the voice recordings of all sessions to text, summarised the sentences into a short description each, and marked the descriptions on sticky notes. Observations and feedback from the interviews were treated identically. Each note represents an aggregation of a participant's search behaviour either from the think-aloud or interview protocol.

Notes were clustered in 10 joint sessions with all authors of approximately 27 hours in total. All authors consecutively posted the notes on a canvas, placing similar ones together and different ones apart. For example, a sticky note with "Google Scholar - search relevant papers" would be placed in proximity to "Library catalogue - enter research topic" because they both are about literature review.

The authors watched each other, while someone placed a note on the canvas. If an author's proposal was controversial, the note assignment was discussed. If consensus could not be achieved, the note was put aside and reviewed at the end of clustering round for an appropriate fit. Thus, topical clusters emerged. After reviewing all notes, we agreed on suitable labels for each cluster. Subsequently, for each note, the assignment to a cluster was checked and adjusted or confirmed.

Based on this clustering, we decided on a three-layer hierarchy that allows for appropriate levels of both abstraction and detail. In the results section, we present the clusters we found during the analysis. The pertaining notes were summarised to provide an overview of the cluster.

[7] In the context of the study, we define data portals as websites that specifically collect and distribute multiple datasets and provide metadata and documentation for those datasets.

Table 1 Career stages, disciplines, and research topic of the participants

Part.	Career Stage	Discipline	Research Topic
P1	Professor	Political Sciences	International relations and development policy
P2	Professor	Sociology	Intergenerational transmission of health
P3	Postdoc	Sociology	Consumption social groups qualitative research
P4	PhD Candidate	Sociology	Social structure research, data analysis, media user analysis, family sociology.
P5	Postdoc	Sociology	Social change life course research, sociology of religion
P6	Master Student	Political Sciences	China: Stability, Legitimacy, Social Credit System
P7	PhD Candidate	Sociology	Sociology of Education
P8	Postdoc	Political Sciences	Civil society civic engagement European integration
P9	Professor	Sociology	Sociology of the Family
P10	Professor	Media Management	Social change in the digital age, media reception, influence of media on society
P11	Master Student	Economic Sociology	Housing industry, insurance industry
P12	Master Student	Political Sciences	International relations

Table 2 List of sources for datasets

Source	# participants
Articles	12
Professors, lecturers, mentors	9
Web search (Google, Bing, Yahoo)	8
Books or Monographs	8
Colleagues or friends	7
Researcher generated dataset in prior research	6
Conference presentation	6
Support of local library	4
Search engine for research data	4
Social media (Facebook, ResearchGate, LinkedIn)	3
Online catalogue of data archives (Figshare, Zenodo or research data centre)	1
TV, Radio, Newspaper	0
Other	0

4 Participants

Our sample consists of twelve participants. Four are female, and eight are male. Eleven are affiliated to a university and one to a public research facility. Four of the participants are professors, three are postdocs, two are PhD students, and three are Master's students. Their primary research areas within the social sciences are sociology (6), political science (4), economic sociology (1), and media management (1). The participants reported their usage of research data in the last three years. They used research data for writing publications (10), for education (8), for comparison with other data (6) for developing a new research question (7), and for using research data in their thesis or dissertation (2). A summary of the participants' career stage, scientific discipline, and research topic is displayed in Table 1.

Using a pre-defined list, we asked the participants about their sources of datasets. They could select as many items as they wanted. Table 2 shows the selected sources and their frequencies for the question "How did you discover the datasets you used in the past?".

On a 5-point Likert scale ranging from "very easy" (1) to "very difficult" (5), participants rated how complex it is to search both for literature and for research data. Figure 1 shows that literature search is rated as rather easy to perform, while searching for datasets is perceived as being more difficult.

5 Results of the interaction analysis

In this section, we present the results of the interaction analysis. Due to technical problems with the video recording in two interviews, the following analysis is based on ten out of twelve participants. Figure 2 shows the interaction path of each participant. We summarised interactions with websites

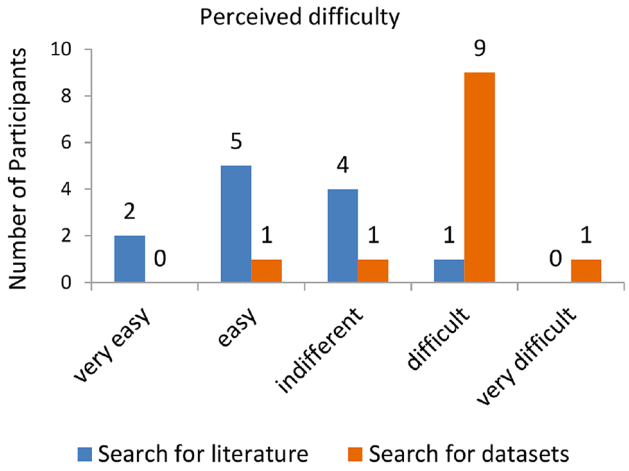

Fig. 1 Perceived complexity of dataset search. Answers given on a 5-point Likert scale to the question "If you compare the search for literature and research data, how complex are the two tasks?"

into eight categories. Please note that a single step encodes only the type of interaction, not its length.

Google was the entry point for eight participants, while two other participants used Google Scholar to begin the search. However, after the almost unanimous entry, participants went into different directions. Three navigated directly to a study or a project page, and three into a data portal. In addition to the two participants who already started with Google Scholar, two more participants began a literature review via a library. During the search, many participants returned to Google. Google served different functions: either to look up resources only known by name rather than URL (e.g. data portals, literature portals, or the homepages of studies) or to issue a topic-based search. On Google, participants found and clicked, among others, on news reports, PowerPoint slides, and government reports.

Literature review was relatively common among participants, considering that the task asked them to search for datasets, not for literature. Seven participants consulted a full-text article during the session, and one more scanned the result list of a literature search to check whether his research question had been answered before. Reading literature served multiple purposes. Relevant publications contain the name of potentially relevant datasets, sometimes describe the quality and the relevance of those datasets, and may include download links. In fact, P12 chose to concentrate entirely on literature review, never visited a dedicated data portal, and did not download any dataset.

However, literature is not the only source of information. Eight participants used a data portal website to find data. Again, one participant relied exclusively on this source (P11). Other participants chose a more balanced approach, utilising a variety of sources to conduct their dataset search.

We observed considerable efforts of the participants to assess the quality and the relevance of the datasets. In six cases, participants consulted additional material like variable reports and questionnaires to understand and interpret the data. The data documentation was read thoroughly, even in our time-limited study. This finding is in sharp contrast to similar studies on information quality on the web [36] or even medical information seeking by physicians [19].

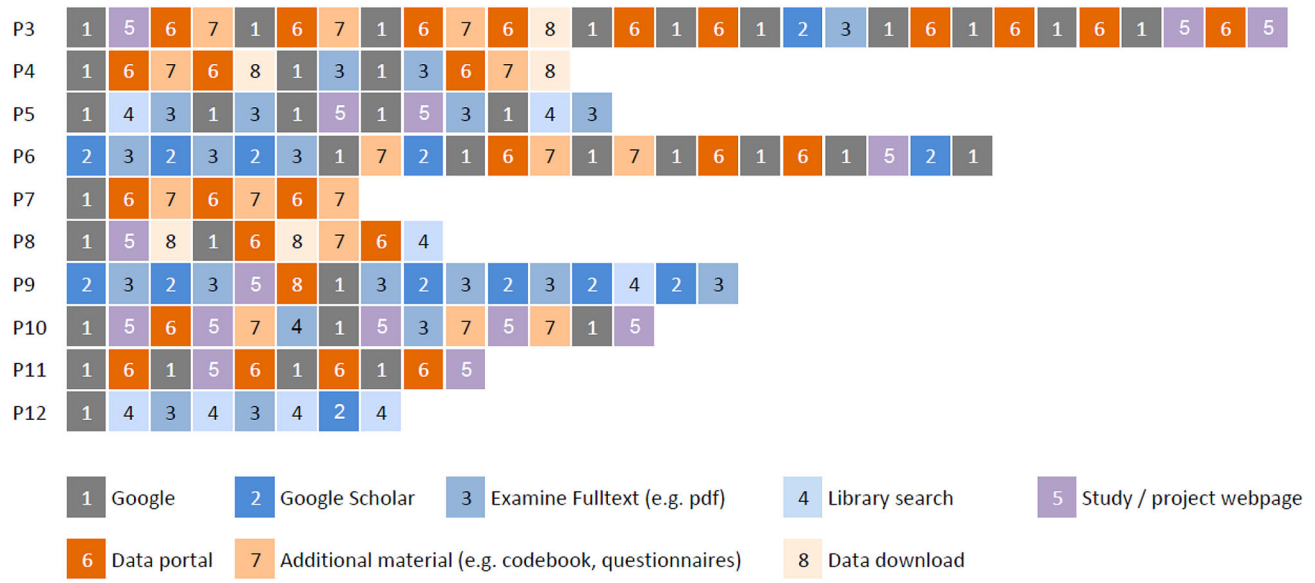

Fig. 2 Visualisation of interaction sequences of ten participants (P3-P12). Each box represents one interaction step independent of the time it took. The boxes are colour coded: grey for Google, blue colours for steps related to literature search, orange for steps related to data portals, and purple for study/project web page visits

Although Figure 2 does not show the time spent on each step, we note that reviewing the literature and scanning the data documentation were both time-consuming activities. Contrarily, data download and Google searches were often short activities. We asked the participants to think aloud during the search, which distorts the time dimensions in our study. A study using log files would provide further insight into the true lengths of interactions.

6 Results from the affinity diagrams

In this section, we present the results of the observations made during the search tasks and in the semi-structured interviews. If not stated otherwise, the order of the results does not indicate a rank of importance or frequency of mentions. We identified four main aspects of dataset search: (1) the **context** of the search including prior knowledge and external factors, (2) the **resources** used by the searcher during searches such as portals and search engines, (3) applied **working methods**, and (4) the **assessment** of intermediary and final search results. The fields and their constituents are visualised in Fig. 3.

6.1 Context

This section describes the context during dataset search, differentiating prior knowledge of the participants, their topic, strategies, and social context (network and colleagues).

6.1.1 Prior knowledge

We identified two aspects related to previous knowledge: prior knowledge about search resources on the one hand and a lack of prior knowledge on the other hand.

Known Resources Participants made use of knowledge from previous searches or their educational background to help them in their search. At the start of the search, participants used known studies and portals. Later during the search, known studies and portals served as a fall-back strategy. Familiar resources were the OECD database[8], HISTAT[9], institutionalised study programs, own previous works, historic or older datasets, and datasets from previous searches. Altogether, the set of resources known before the search makes up the "resource knowledge portfolio" of a searcher. The participants' portfolios developed through personal contact of relevant researchers who publish in portals such as HISTAT (see Subsection 6.1.4 Colleagues and Network), acquaintances at the university, dealing with datasets previously at work or during education, or presentations of datasets at conferences. One participant also reported having heard about a dataset on the radio. Participants currently enrolled in a Master's program mostly based their searches on study series and portals they encountered previously during their education.

Lack of Knowledge Participants showed a varying level of knowledge about data portals, data search engines (e.g. Google Dataset Search was unknown to all participants), and datasets of popular surveys (e.g. the ALLBUS[10] or EVS[11]). Missing prior experiences with raw datasets, in general, became apparent during the study: some participants reported avoiding to work with raw data altogether. They base their analyses purely on pre-processed data such as aggregations or diagrams and stated to avoid complex statistical analyses. A lack of knowledge can be a consequence of the impossibility to know or to retain all information. A portal is, for example, known by name, but its URL is unknown, the portal on which a well-known study series can be found is unknown, or the name of a formerly visited website cannot be recalled.

Summary In general, participants referred to known data if possible, making the researcher's own "portfolio" an essential driver for the individual dataset search. A lack of knowledge was identified for dedicated dataset portals (Google Dataset Search). Consequently, this lack of knowledge influenced data search behaviour.

6.1.2 Topic

In the user study, participants searched datasets with regards to a specific, self-chosen topic. The following paragraphs describe the role of the topic and the research questions throughout a dataset search.

Trigger There were a variety of triggers for choosing a particular topic or research question. One participant was looking to verify a hypothesis based on qualitative results. Another person was looking into the reasons why a hypothesis assumed earlier was wrong. Others were more strategic, e.g. one participant was looking for variables of a popular dataset that were overlooked so far in the publications to find promising publishing opportunities. However, most participants were simply expanding on topics they had worked on before.

Topic Granularity Participants varied in how they framed their topics. While some reported that it takes weeks to determine a precise research question, others started with a narrow topic, including temporal and spatial extent or conceptualisation. Nevertheless, we observed participants issuing unspecific searches to find interesting data or conducting searches for several sub-questions within a general

[8] OECD https://data.oecd.org/.

[9] GESIS Historical Time Series https://histat.gesis.org/histat/.

[10] German General Social Survey (ALLBUS) https://www.gesis.org/en/allbus/allbus-home/.

[11] European Values Study https://europeanvaluesstudy.eu/.

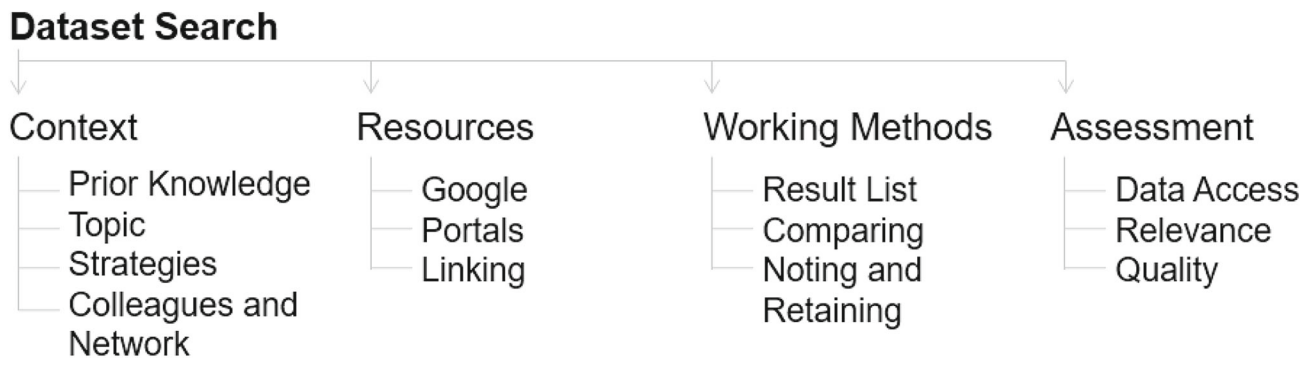

Fig. 3 Structure of user study results with four main areas and 13 sub-areas

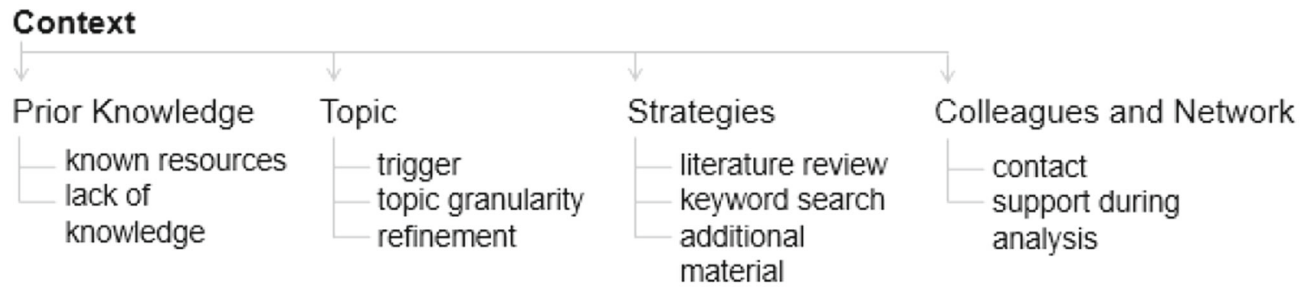

Fig. 4 Structure of "Context" Cluster

topic. Such exploratory searches were also used to determine the degree of innovativeness of a research question. The search topics in our study covered subject areas of politics (Brexit referendum), religion (secularisation in Eastern Europe), sociology (life satisfaction in China compared to Germany), criminology (perceived vs. actual murder rates), public policy (public expenditure for housing since the 80s), or sports (effects of travelling on the identities of soccer fans).

Specification and Refinement We observed that participants operationalised their research questions during the search session. They derived fine-grained variables from coarse-grained concepts while looking at datasets and their descriptions. They developed an understanding of how a concept was operationalised in other studies, leading to iterative adjustments of the research questions. This is further elaborated in Subsects. 6.3.2 Comparing and 6.4.2 Relevance.

Summary Data can be a driver for the research question, i.e. the data search is performed to formulate a hypothesis. Usually, however, participants initiated the data search after formulating a research question. In any case, the research question was reshaped during the data-seeking process, usually from a generic to a more fine-grained formulation.

6.1.3 Strategies for finding datasets

This section describes the search strategies employed during dataset search. The most distinctive search strategies observed in this user study were the usage of literature to obtain datasets or information on datasets, the usage of topical keywords to query search engines, and utilising additional material published alongside a dataset.

Keyword Search All participants used keyword searches in Google, data portals, and library catalogues as a search strategy.

Different keywords were used in subsequent searches.

Queries were reformulated, expanded, and replaced by synonyms (see Subsection 6.2.1 Google). Even when manually scanning texts, participants reacted on keywords in the text rather than on whole phrases or broader concepts.

Examining Literature Literature search is a crucial part of dataset search. Participants scanned the literature for mentions of datasets, including grey literature. Participants evaluated diagrams in articles as hints to underlying datasets and tables with aggregated numbers as pointers to datasets. Once a relevant article has been identified, footnote chasing and cited reference lookup were used to continue the search.

Participants commented that searching literature is a prerequisite for searching for datasets. A common understanding among participants was that the publication on a specific topic points to datasets on that topic. We observed that participants reacted positively to dataset providers that curate specific bibliographies. One participant compared finding datasets via literature with gambling, while another embraced finding serendipitous inspiration through non-relevant content.

Examining Additional Material Some aspects of a participant's search strategy depended on additional material. The questionnaire file and the codebook published alongside the dataset or publication deliver additional information on the data. Participants described the metadata and documentation as useful for their search, e.g. as provided in the ESS[12]. Participants also used the sample demography or the study results summary delivered in additional materials to gain an overview of the dataset's potential relevance.
Summary Dataset search via literature search seems to be a common strategy. Participants scanned the literature for mentions of datasets. The search through additional material of a publication points in the same direction: Proxies were used during search rather than the datasets themselves. Additional material was also perceived as essential for the relevance and quality assessment (see Subsection 6.4 Assessment), which supports the relevance of interlinkage of dataset-related resources (see Subsect. 6.2.3 Linking). With respect to the results of the interaction analysis (see 5), we have no evidence for distinctive or elaborate data search strategies.

6.1.4 Colleagues and network

Participants stated to use their network in one form or another for data search during our study. Participants asked colleagues for data when an initial search was unsuccessful, when colleagues were considered experts for a specific topic, or when they mastered the language used in a dataset-related publication. Rather than as a replacement, participants contacted their professional network in addition to their dataset search.
Making Contact for Data Acquisition Maybe because of the time pressure of the study, we observed contact via phone only once. Another participant mentioned he would ask a particular colleague in this situation, but she was unavailable. Even though we observed only two such occurrences, most of the participants mentioned personal contact as a means for finding datasets. They mentioned email, face-to-face communication in the faculty or in meetings, or direct interaction at conferences as their usual communication channels. Participants reported using literature or the platform ResearchGate[13] to contact the authors of datasets directly.

In our study, researchers stated that they ask for data from other research groups on occasion, e.g. if a Google search fails or if a dataset seems publicly unavailable. They also reported asking faculty colleagues for data when working on a related topic. Senior colleagues were also mentioned to provide help in dataset search: Participants reported that their senior colleagues at work are promising sources for relevant datasets. As such, senior colleagues play a role in supporting researchers by using their networks to acquire relevant datasets.
Support during Analysis Another moment in the research process, where colleagues seem to play a significant role, is during the analysis. Particularly when using mixed-method approaches, qualitative researchers reported to ask their quantitative colleagues for help for the analysis or to delegate the analysis to student assistants. To clarify questions on how statistics were calculated in the supplemental material of a dataset or in publications, participants indicated they would contact the authors of a dataset or other researchers who previously worked with the dataset.
Summary It seems to be more usual to contact people inside the current professional network of a researcher than to contact researchers outside the network. Personal contact is vital during data search, data acquisition, and analysis. Thus, colleagues support the development of a researcher's portfolio of known resources (see Subsect. 6.1.1 Known Resources).

6.2 Resources

This section describes the resources participants consulted during their search process, subdivided into Google, portals, and different types of links between resources.

6.2.1 Google

Google took a popular role in the search for datasets. Not only did participants use it as a search engine to find datasets and literature. It also served as the starting point to search for specialised portals and as a fall-back strategy if specialised search engines and portals did not deliver satisfactory results. Moreover, participants used Google to find open-access versions of literature as an alternative to services with pay-walls or sign-up restrictions. Some expressed assumptions about Google that drive their work with the search engine: Querying Google in English was expected to give more search results, including not only scientific resources but also other content such as newspaper articles or commercial datasets. Furthermore, participants expected Google to be better at finding literature than finding datasets. One participant expressed reservations about Google. He believes that search results depend on personalised profiles, leading to different result lists for different researchers. Particularly in the context of teaching, he prefers if all students and researchers are able to start with the same neutral and objective results.
Querying Queries issued during the user study were aimed to find either direct links to data related to the research question, or links to specialised portals. Some users tried a full-text search of the research question at hand. When targeting datasets, participants often added "data" or "dataset" as search terms to the query. All participants reformulated

[12] European Social Survey https://www.europeansocialsurvey.org/.

[13] see researchgate.net http://www.researchgate.net.

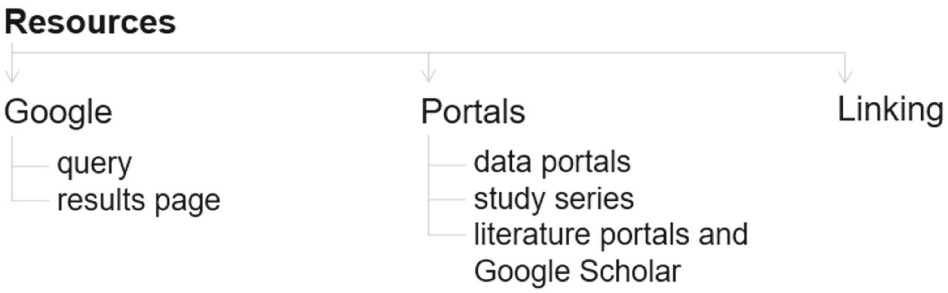

Fig. 5 Structure of "Resources" Cluster

their query one to five times throughout the search session to get more precise results or broaden the results list.

Search Results Page After issuing the query, participants evaluated Google's result page. Often, participants clicked the first result in the results list, whereas other participants scanned the results list meticulously—sometimes even multiple times. When asked how they evaluated the result list, participants indicated their judgment based mostly on the title. They scanned the title for mentions of the topic to evaluate how well it fits their research questions regarding contents, the type of data behind the result (dataset, literature, website), or the type of scientific work (theoretical or empirical).

Summary Overall, although participants expressed doubts about Google as a neutral instance and the quality of results, Google took a popular role in the present user study. Despite this preference for the Google product family, none of the participants used Google's dedicated dataset search engine, Google Dataset Search. In the post-task interviews, it became clear that Google Dataset Search[14] was not known among the participants.

6.2.2 Portals

Besides Google, specialised portals play a major role in searching for datasets in the social sciences. The websites of study series provide detailed reports on their data, reflecting the high degree of institutionalisation of the social sciences. Likewise, libraries and literature portals are frequent starting points for a dataset search session.

Data Portals Participants used or stated to use one of the data portals for unknown-item searches. Portals mentioned for statistical data were: destatis [15], BKA[16], Deutschland in Zahlen[17] for statistical datasets related to Germany, and OECD[18] and Eurostat[19] for international datasets. For national survey data, researchers accessed GESIS DBK [20] or Polizei Hamburg. The ICPSR[21] and the national archives of several countries are used for the search for national survey datasets. Data portals were often known by name and either accessed via typing the name into Google or even entering the URL manually.

Study Series Study series are long-term studies, generally conducted by dedicated projects or institutions that produce datasets regularly and with high-quality documentation. We observed people searching for specific study programs by name. When asked, participants knew at least one of the study programs EVS[22], ESS[23], ISSP[24], ALLBUS[25], Eurobarometer[26], or PKS[27]. We found that participants tended to choose datasets from the well-known study series first. They knew that the websites are usually maintained by a study series consortium or an institution that provides an overview of the study series and a search function.

Literature Portals and Google Scholar Participants used libraries and literature portals. The local libraries with their online catalogue still played a role for the participants. University libraries were reported to be the default starting point

[14] The study was conducted in the summer of 2019. Google Dataset Search launched just a few months earlier.

[15] Statistisches Bundesamt, https://www.destatis.de/.

[16] Bundeskriminalamt, https://www.bka.de/.

[17] Germany in Numbers, https://www.deutschlandinzahlen.de/.

[18] Organisation for Economic Co-operation and Development, https://www.oecd.org/.

[19] Statistical Office of the European Union, https://ec.europa.eu/eurostat/de/home.

[20] gesis Datenbestandskatalog, no longer available

[21] Inter-university Consortium for Political and Social Research, https://www.icpsr.umich.edu/.

[22] European Values Study, https://europeanvaluesstudy.eu/

[23] European Social Survey, https://www.europeansocialsurvey.org/.

[24] International Social Survey Programme, https://www.gesis.org/issp/home/.

[25] German General Social Survey, https://www.gesis.org/allbus/allbus.

[26] https://www.europarl.europa.eu/at-your-service/de/be-heard/eurobarometer.

[27] Police Crime Statistics, https://www.bka.de/DE/AktuelleInformationen/StatistikenLagebilder/PolizeilicheKriminalstatistik/pks_node.html.

for searches. Also, participants used EBSCO [28], e-journals, and ResearchGate[29] during the study.

Besides Google's generic web search engine, participants used Google's scientific search engine, Google Scholar. While some participants used it as a search engine for literature to find background information and publications citing datasets, others added the keywords "data" or "dataset" to their query to search for datasets directly on Google Scholar.
Summary Participants consistently showed knowledge of popular portals and study series in their field. Popular studies or study series were used as a starting point for the dataset search—either directly via the URL or indirectly via a known-item search on Google (see Section 5 Results of the Interaction Analysis). Cross-disciplinary dataset portals were hardly known, and if so, they were not used during our study.

6.2.3 Linking

Participants used or would have liked to use references between different dataset-related objects, be it on study websites, study series websites, or in publications. Where available, links to similar studies were considered helpful. Participants expressed the need to navigate between different dataset-related objects, e.g. from a specific survey question or variable in a publication to the corresponding survey documentation (e.g. the codebook providing information on the question, answer options, and answer distribution), or from a raw dataset to its documentation. While the reference to a dataset in a publication was usually available in prose format in the text itself, systematic linking to the datasets or dataset metadata was not always available. Participants stated that missing linkage to the datasets used for a publication raises questions. They hypothesised that links could be held back on purpose if the dataset was of low quality and therefore not to be published, or that the link to the dataset simply was not set.

Having an overview of publications and projects that use a specific dataset was also reported to be useful, yet it was acknowledged that such a source tracking would be difficult to implement.
Summary References between dataset-related objects, be it simple hyperlinks or qualified references (entity interlinking such as related code), were perceived helpful and used, if available. A dataset can be perceived as a "one-way road" if no links to similar or related datasets exist. In general, participants considered linking as relevant for their dataset search as one path for improvement.

[28] EBSCO provides different library and research services such as EBSCOhost Research Databases, EBSCO Discovery Services, or the Publication Finder Interface, https://www.ebsco.com/e/de-de.

[29] https://www.researchgate.net/.

6.3 Working methods

We summarise the working methods participants employed, subdivided into interactions with result lists, comparing, and noting and retaining.

6.3.1 Working with result lists

Throughout the dataset search, participants made use of search engines, portals, and data catalogues that are searchable and often return a result list for a given query. In our study, we observed that participants sequentially examined the result page while opening several results at once for relevance assessment. Search lists, therefore, functioned as an "anchor" to come back to and resume search activity after viewing the details of individual search results. Participants showed varying ending conditions for their search. While some participants ended their literature search after the third search engine result page, others did not stop until all known sources have been assessed. Participants interacted or expressed the desire to interact with the results list by sorting and filtering the results.
Filtering Filtering was considered a function to find the exact dataset that is needed to answer the research question. Participants used filtering to speed up the search and relevance judgement process by eliminating irrelevant data. Some dataset providers have grouped their data by topic, which is likewise perceived as a means of filtering data. Participants with a narrow and precise research question at hand set clear restrictions for their target dataset in terms of metadata (e.g. publication date), the sample characteristics (e.g. demographics of the sample group), connected entities (e.g. author or institution), or the type of data or publication (e.g. time series vs. single survey, journal articles vs. conference proceedings). If the services did not support filtering in those dimensions, participants scanned the result list manually to filter out relevant results. Filtering was a desired functionality for both literature search and dataset search.
Sorting Some participants were also concerned with the ordering of the results lists. Although most participants used the standard sorting of the search engines, one participant especially expressed concerns about the sorting of Google Scholar, which the participant assumed to be ordered by the number of citations. Another participant assumed that search engines take click-through data into account for sorting the result lists. As described in Subsect. 6.2.1 Google, search engines were suspected of returning personalised search results.
Recommending Throughout the study, participants expressed the necessity to find more items similar to the one at hand: similar questions, similar datasets, or related literature. Similar data (single questions or whole datasets) are needed to either answer the same research question, i.e. to compare or

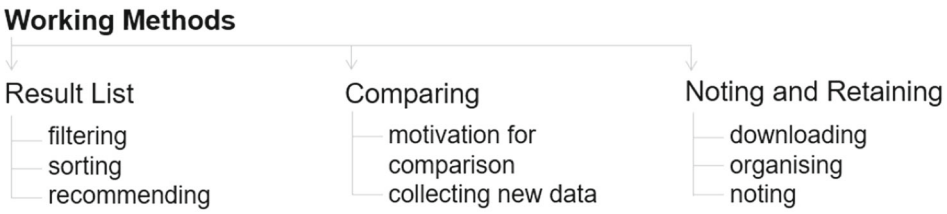

Fig. 6 Structure of "Working Methods" Cluster

support the results, or to provide context around the research question, e.g. by answering the same question but for a different age group. Participants also mentioned "suggested for you" algorithms as a source for inspiration and related literature. Similarity algorithms also showed to be helpful when formulating search queries, e.g. participants relied on Google's query completion suggestions to build their query, rather than iteratively searching for the best keyword.

Summary Overall, filtering seemed to be a more prominent tool than sorting when working with result lists. Participants used those tools to gain the most relevant results, both in literature search and in dataset search. However, in dataset search, different filters play a prominent role, e.g. a detailed description of the sample. A precondition for filtering and sorting is the availability of structured information about the sample, the content, and other metadata.

6.3.2 Comparing

Researchers showed the need to compare interim results within a data search session. This assessment is time-consuming and includes evaluating the methodology, as well as comparing the temporal scope, geographical scope, and content.

Motivation for Comparison Searching for datasets led to a large set of search results in most cases. If multiple datasets were deemed interesting, at least two preliminary datasets were selected for further inspection (see Subsection 6.4.2 Relevance). If a dataset increased the significance or the validity of the research, it was used as an additional candidate. Participants also compared datasets to detect bias in data, e.g. in commercial studies or governmental studies. Furthermore, comparisons were performed to evaluate the quality of a dataset in detail. Existing harmonisation efforts, e.g. in the international OECD database, were appreciated in general. However, datasets from national studies often provide more detail, which, in turn, requires more effort when comparing different datasets from the same country. Comparing detailed datasets was consistently perceived as a difficult task.

Collecting New Data If a comparison of candidates datasets does not lead to a satisfying dataset, participants considered collecting the data themselves. Collecting own data is perceived to provide full control over the operationalisation (e.g. variables, scales, items). However, it is more time-consuming, costly, and usually does not live up to popular international datasets in terms of sample size.

Summary Before two or more datasets can be compared and used to support or reject the working hypothesis, researchers need to ensure comparability. Participants reported interlinking between related datasets (see Subsection 6.2.3 Linking) to help compare datasets. While challenging to implement, the ability to verify data comparability was stated as a desirable feature of a dataset search engine.

6.3.3 Noting and retaining

During the dataset search task, participants used several techniques to note and retain interim search results and candidate datasets.

Downloading Participants downloaded material related to candidate datasets, even if its relevance was not clear. Materials that participants downloaded were publications, questionnaires, integrated datafiles[30], and files describing the dataset.

Organising Downloaded files were organised in the local file system and superficially scanned and evaluated. Some participants excerpted parts of publications or datasets, e.g. by transferring them into Excel sheets. One participant ranked relevant publications according to their perceived usefulness. Others kept potentially relevant documents in a dedicated place, e.g. a new folder or re-organised the tabs in their browser. One participant reported to appreciate the OECD[31] website that allows users to create personalised download packages with individually selected variables or study metadata.

Noting Participants took notes on paper, digital post-its, or in doc files. Notes mainly contained the research questions, keywords used during keyword search, relevant references, and contextual information from literature (e.g. definitions, threshold numbers). Further, participants highlighted text in files.

Summary Organising preliminary results is done by downloading files, taking notes, and highlighting. Participants organised documents and data often locally. Nevertheless, participants demonstrated that it is crucial to organise preliminary results by saving literature, search results, or data.

[30] e.g. merged datafiles of distinct waves of a longitudinal study.

[31] Organisation for Economic Co-operation and Development.

6.4 Assessment

In our user study, participants evaluated datasets they found during their search concerning their relevance and quality. A precondition for evaluating is the access to the data itself. The assessment itself includes judging the relevance and the quality of a dataset.

6.4.1 Data access

Access Process Accessing datasets often required several steps. Some services were only available to registered users, requiring the participants to log into an account or create an account first. They were asked to read the privacy policy related to the data, and, in some cases, to fill out a data usage contract, which usually took only a few minutes. And yet other platforms required the participant to indicate a purpose of use or the origin (country or institution) before their data could be accessed.
Access Barriers Participants perceived various barriers when trying to access data. A dataset might be mentioned in a publication, or metadata on that dataset might be public and findable, but the dataset itself might not. From their experience, participants reported cases where the dataset is not available online, not available at all, or the access process remains unclear to them. Moreover, some datasets are available but not accessible to them in particular, e.g. because a fee needs to be paid to gain access. Even if data are accessible, it might be offered in a proprietary data format, which makes it difficult, if not impossible, to use the dataset. Besides the linkage of datasets to documentation and literature, participants also expressed a desire for fast and easy downloads of the datasets. However, even popular portals such as destatis did not always provide a direct download link for the dataset.
Presentation of Datasets Several participants also commented on the presentation of datasets when it comes to quality. The data format seemed to be of particular importance: participants favoured a choice of data formats (e.g. Excel, SPSS, STATA) from which they selected their preferred format. Furthermore, diagrams or graphical summaries of the data provided a clear overview and fast comprehension of the dataset at hand.
Open Access According to our participants, too few datasets are openly accessible. Often, only supplementary material is openly available. Open access was considered to be a possible solution for the barriers mentioned above. One participant proposed that publishing data under an open-access scheme should be made mandatory for all publicly funded research.
Summary While participants expect and accept a certain kind of regularised process before accessing datasets, the authorisation procedure itself or the often unclear access rights are perceived as barriers. Barriers impede access and (re-)use of datasets. In any case, the interactions required for registration interrupt the data search process.

6.4.2 Relevance

Assessing the relevance of a dataset for the research question takes an essential role for social scientists during dataset search. Relevance assessment has three key aspects: relevance of the content, relevance criteria related to data characteristics, and documentation needed to assess relevance.
Topical Relevance Participants based their relevance assessment mostly on the topical fitness of a dataset, i.e. how well the content of the dataset fits the research question at hand. We observed both positive and negative outcomes of such comparative assessments. Upon closer inspection, datasets turned out to be unsuited to answer the research question at hand, while others showed to support the working hypotheses of participants. In the latter case, being able to make an initial analysis of the data in the data portal (e.g. variance analysis of a single variable, distribution of answers, or cross-tables) was reported helpful to assess its relevance.
Metadata Besides the topical relevance, access to detailed metadata is crucial during dataset search. Participants considered the type of publication of the primary research connected to a dataset (e.g. research paper, doctoral thesis, final thesis of Master's or Bachelor's students), temporal and spatial extent of the data (e.g. survey period, country or region of assessment), and characteristics of the sample (e.g. sample size, socio-economic characteristics, age groups, or recruitment method). Participants also deemed source information of a dataset informative, which discloses detailed information about authors, executives, curators, and publishers. Furthermore, the data collection method was reported to be an indicator of the "scientificness" of a dataset. Participants assessed whether the collection method meets their standards, both concerning quality and methodological fitness. Participants acknowledged that the collection methodology of the primary research connected to a dataset is more critical in dataset search than in literature search. Another relevance criteria reported by the participants were the frequency of citation or usage by other researchers. The number of connected research was deemed meaningful, as well as the quality of the related publications. Other reported relevance indicators were the naming of dataset files and the possibility of data aggregation for gaining an overview. Finally, the data needed to provide the level of granularity to answer the research question at hand.
Documentation For assessing the relevance of a dataset, participants were looking for resources that provide information on content and relevance criteria. Resources mentioned by participants were the methodological report of a survey, the questionnaire itself, interview guidelines in case of a personal

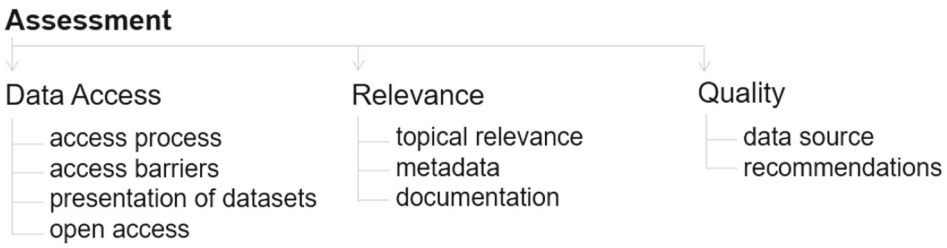

Fig. 7 Structure of "Assessment" Cluster

survey, the codebook of a dataset, and a list with definitions of terms used in the dataset and its documentation.

Summary In summary, for the participants of our study, assessing relevance represents the complex act of judging whether the data fit the research question at hand. Several different criteria played a role in relevance assessment, and various resources were used to assess the relevance criteria. The documentation of the dataset took a central role in relevance assessment.

6.4.3 Quality

Besides assessing the relevance of a dataset, participants reported evaluating its quality. When distinguishing between these two clusters, we often encountered an overlap between quality and relevance. Aspects that were important for relevance were also essential for the quality of a dataset, e.g. the extent of the metadata. Vice versa, some participants refused to work with data that would not meet their quality criteria. We found that quality is mainly judged based on the data source and its reputation, the presentation of the dataset, and the recommendations or judgement of other researchers.

Data Source The source is a central factor for perceived dataset quality. Participants indicated to use the reputation of the authors, initiators, platforms, and institutions connected to a dataset as an indicator of quality. Concerning the authors and principal investigators of a dataset, participants reported paying attention to the primary field of research, the reputation within the field of research, and other background information. The institution connected to the dataset also influenced the trust in the data. Participants expressed to be more vigilant with privately owned companies than with accredited research institutes. The opinion on the reputation of governmental datasets was ambiguous, with participants expressing reservations, and others naming governmental institutions as trustworthy (e.g. BKA[32], LKA[33], State Statistical Offices). Other sources deemed trustworthy are the World Bank, Eurostat, Leibniz institutes such as GESIS, and WZB[34]. Likewise, providers of datasets were part of the quality judgement, with commercial providers being described as less favourable than non-commercial providers. The "degree of popularity" of a dataset was mentioned as a quality attribute as well. One participant suggested that large data generation programs such as the ALLBUS or SOEP are well known and therefore trustworthy, yet quite broad. A recurrent topic was the country of origin of the dataset. Participants stated that the quality of data is dependent on the scientific standards in a country, leading to a variety of quality levels across national statistics.

Recommendations When asked about quality assessment, participants also reported the importance of recommendations from other researchers, either directly through suggestions for a specific dataset or indirectly via the reputation of a well-known dataset in the community.

Summary Overall, searchers made a quality judgement to decide whether they want to work with a relevant dataset. Frequently recurring criteria for quality assessment were the reputation and country of origin of institutions and individuals who worked on the dataset. Another important factor was the data format. Nevertheless, the threshold of quality for accepting a dataset seemed to be very subjective.

6.5 General feedback

We summarised general feedback of the participants into the areas *perceived risks and problems with data search*, *suggestions for improvement*, and *prognosis*.

Risks and Problems with Data Search Participants reported several perceived risks during data search: missing a relevant dataset, finding only partially suitable datasets, and finding no suitable data at all. In the case of a large number of hits, one participant sees the potential of early over-specification of the search query.

Participants also encountered problematic situations during their search. Comparing data at the international level was perceived as challenging due to diverting quality standards at different locations. Furthermore, possible systematical errors during data collection were not mentioned in the documentation. While codebooks seemed too extensive and overwhelming for a participant, others noticed outdated information, bad keyword assignments, or inconsistent definitions as obstacles in finding datasets.

[32] Bundeskriminalamt

[33] Landeskriminalamt.

[34] Wissenschaftszentrum Berlin.

Fig. 8 Structure of "General Feedback" Cluster

Both, too many and too few search results could be problematic. The latter was often encountered when searching for specific topics.

Suggestions for Improvements Participants suggested building a global, dedicated data search engine to aggregate local search engines. They also requested domain-specific search engines as they can offer better support for question and variable searches. Recommendations on improving the usability of such data search portals included harmonisation of data, topic classification to enable topical browsing, statistics for each dataset (e.g. number of downloads and views), more filter options in general, as well as the possibility to join and compare variables. Additionally, participants wished that data search portals provide better support for creating questionnaires or formulating questions.

Participants suggested shortening the click paths to datasets or related material through direct (deep) links. Direct links to items in a questionnaire, for example, would eliminate the need for frequent use of the "CTRL-F" shortcut. Search recommenders were requested as well, e.g. in the form of search term suggestions or recommender for similar datasets. One participant suggested a functional query flag "raw data" that would lead directly to raw data of relevant studies.

Variable names with low information content, such as *"attr_01"* or *"BNX"*, were seen as an obstacle to getting an overview over the data quickly. The use of standardised metadata would improve previews and thus the findability of datasets. In this context, dataset documentation was requested to be as comprehensive as possible. In the case of longitudinal studies, data of all waves should be available at a glance.

Apart from functional recommendations, an independent quality label for datasets was proposed. Furthermore, participants felt that dataset search should be part of the university curricula to increase dataset search literacy.

Prognosis When asked for their expectations regarding research data in the future, participants predicted an increase in dataset search importance. The terminology of the studies in the social sciences and methods applied are expected to evolve. In the case of longitudinal studies, this will pose additional requirements to tools that aim to support researchers in finding research data. Some participants expected more replication studies to be conducted as well as more panel studies and multi-level analyses that drive the importance of dataset reuse. Others mentioned decreasing response rates as an essential factor for dataset reuse. Additionally, the analysis of digital behavioural data and the link between digital behavioural data and "traditional" data are expected to increase.

Summary Many suggestions for improvements require increased metadata quality and availability. They are often related to particular functional aspects of dataset search portals, such as recommender systems, interlinked entities, and the possibility of comparing items or variables. Despite many reported problems, three participants expressed their satisfaction with dataset search.

7 Discussion

7.1 Key Findings

From the results, we derive seven key findings, which we summarise below:

1. **Literature search is an important part of dataset search.** Literature can provide information that is needed to make relevance and quality assessments. We observed participants splitting their time between looking at datasets, looking at the documentation, and looking at literature. The links between these different entities are rarely directly and explicitly provided. Instead, participants used the name of a dataset in a publication to search for it on Google.
2. **Tools are unknown.** Many participants wished for more support during the search process, e.g. consistent linking between literature and datasets and more explanation for variables. Such support exists, e.g. the entity linking in GESIS search or ICPSR, but the participants were unaware of those tools. There are two reasons for this lack of knowledge. First, these support tools do not show up in the observed workflows of the participants. Second, a general lack of dataset literacy can be observed, which we will discuss in detail in finding 6.
3. **Tools are creatively misused.** Complementary to finding 2, we saw participants use popular tools like literature search in remarkably creative ways for dataset search. They searched for datasets on literature search engines (e.g. Google Scholar) and manually scanned literature for mentions of the datasets used. We will give more examples below (see Subsection 7.3 Data Search Literacy).

4. **Relevance assessment is very complex.** To find all relevant information, participants have to vigilantly inspect the documentation, including methodology, sampling, and precise definition of the variables. As mentioned before, many participants relied on a literature review to assess the dataset's prominence by the number of citations. There are many aspects of a dataset to consider during relevance assessment. However, relevant attributes are often either not present in a standardised, comparable way, or not available at all.
5. **Accessibility to datasets is limited.** Access procedures such as registration or authorisation cause interruptions in the dataset search workflow. Researchers might even abandon the process if it remains unclear if and to what extent access is granted afterwards. Filling in a form and waiting for the response pose a temporary missing link to the desired datasets. Therefore, access procedures should be revised for potential automatisation, e.g. integration with single sign-on services. Visitors should be able to easily recognise the level of accessibility of a dataset which is accessible to them. For example, a colour scheme (e.g. green for direct access, yellow for restricted access, and red for no access) as an indicator for the time required to access could be displayed so that researchers can decide whether they are willing to take these steps.
6. **Dataset search suffers from missing interlinks.** The information needed to assess relevance is not all in one place. Instead, they are highly distributed across the documentation, literature, and other datasets. This is particularly tedious when looking for more than one dataset, e.g. when comparing two independent sources. In some cases, the relevant information is not available at all. Therefore, more complete and standardised dataset metadata is required. In this case, the only source of information is the professional network of the participants, particularly professors for students, and colleagues for more experienced researchers. Personal contacts are used for data search (e.g. asking for relevant datasets), data acquisition, and analysis of the datasets. In some cases, it was reported that personal contacts were used explicitly for such a task (cf. section 6.1.4).
7. **Dataset search literacy is low.** The search for literature is a very common task, but searching for datasets is not. While literature search is often taught at universities, similar courses for handling data are much rarer. We observed very different approaches to dataset search in our study. Particularly the more experienced researchers had individual "tricks" and were familiar with popular portals and study series in their field. Still, there is a lack of common ground or best practices for dataset search.

7.2 Comparison to related work

Our findings overlap with prior studies regarding dataset search, in particular findings 4 to 6. In both, Gregory et al. [16] and Pasquetto et al. [32], the authors conclude that dataset search is a complex process that transcends a simple portal-user model. Therefore, the complexity observed in our study is not a unique trait of social scientists, but a phenomenon that occurs in many disciplines. Similarly, Pasquetto et al. [32] stress the importance of professional networks and the need for collaboration between data producers and data consumers. Although Kern et al. [21] compare literature and dataset retrieval as two separate tasks, to our knowledge, the intermingling of those two tasks has not been mentioned in literature so far. Gregory et al. examined the literature on dataset retrieval concerning *Users and Needs*, *User Actions*, and *Evaluation* for different scientific disciplines [15]. The findings of our observational study confirm many practices identified by Gregory et al. in literature for the social sciences [15]. These include the use of personal networks, the well-supported identification of data from national, publicly funded datasets, and the importance of additional documentation for evaluating a dataset. Nevertheless, we found mentions of literature search in a guide to dataset search [38]. Likewise, the CESSDA guide for data discovery[35] describes an introductory fictive data discovery story similar to what our participants did. The guide itself, however, does not explicitly mention literature as a proxy for dataset search.
The use of local libraries was similarly not mentioned in prior literature, although we found it to be used quite often (see Table 2 List of sources for datasets.).

7.3 Data search literacy

Several participants wished for an international, interdisciplinary repository for research data so that they no longer have to browse numerous data portals. In fact, several such repositories exist already. DataCite[36] contains nearly all datasets registered with a DOI. Similarly, Google Dataset Search[37] is a collection of datasets that have been annotated with schema.org[38].

Participants also wished for dataset citation and extensive linkage of literature and datasets. Yet, none of the participants looked for data bibliographies or utilised DOI data citation. Similarly, none of the participants mentioned the

[35] https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/7.-Discover/The-process-of-data-discovery.

[36] https://datacite.org/.

[37] https://datasetsearch.research.google.com/.

[38] https://schema.org/.

makedatacount.org[39] initiative, which provides centralised standards for data citation. Uncertainties at the researchers' side regarding the completeness of a dataset search could be reduced using such a platform. When asked for the common sources of datasets (see Table 2 List of sources for datasets.), participants showed that using dedicated data search or data sharing platforms such as Zenodo or Figshare is not the standard option. More commonly, datasets are known through research articles, professors, lecturers, mentors, and web search. In some interviews, we explicitly asked for Google Dataset Search, and it was generally unknown to the participants[40].

Many use cases such as matching datasets by geography, time, and content were often not directly supported by the data portals. Thus, participants had to inspect and compare a large number of datasets manually. One of the most common use cases is the inspection of the variables, but these are often not part of the metadata. Instead, the participants had to download the documentation and search for it using the "CTRL-F" shortcut or manually scan the texts. Data documentation standards, such as DDI[41], address this problem. However, only a fraction of data portals uses the DDI standard.

In our study, contact and interaction with colleagues were reported as important. However, joint searches for data did not occur. Participants reported seeking help only to resolve data-related problems. For sharing data with colleagues, participants stated to use Dropbox[42] rather than a dedicated data-sharing platform. Although ResearchGate was considered a possible communication channel to contact other researchers, we observed no collaborative activities on that channel in our study. This is surprising, given the various stages during the data-seeking process in which other researchers are involved, and given the variety of existing collaborative data analysis platforms[43].

The most obvious case of wrong tool support was the problem of wrong data formats. While there are tools that can convert between different data formats, these are often proprietary and not well understood. Some data portals provide their data in more than one format, but many offer only a single format. This is a problem, especially when the data format is proprietary, such as the popular SPSS[44] format. Not all institutions offer access to the number of software needed to work with proprietary formats.

[39] https://makedatacount.org/.

[40] This was several months after the launch.

[41] https://ddialliance.org/.

[42] see http://www.dropbox.com.

[43] e.g. the Jupyter Project https://jupyter.org/ or Apache Zeppelin https://zeppelin.apache.org/.

[44] https://www.ibm.com/de-de/analytics/spss-statistics-software.

7.4 Strengths and weaknesses of the methodology

The strength of the observational study, we conducted, is that we were able to directly observe behaviour, instead of having to rely on the recounting via interviews. Also, unlike a log analysis, we get a full picture of all search-related activities, instead of only the portion that we happen to have access to. We attempted to exclude as many factors as practical that would change the behaviour in its "natural" form. The participants were visited in their natural environment. They were asked to search for data they might want to search for as part as their on-going research activities. They were free to decide on how to conduct this search, including writing emails and making phone calls.

Still, there are some possible distortions: we only observed a part of the full process that, under other circumstances, might have spread over several search sessions. Moreover, we may have primed the participants towards using their networks by making the possibility explicit in the initial task description.

Furthermore, the dataset search was initiated by our study rather than an intrinsic need within an actual research project. In our study, some researchers moulded the research questions to fit the data (also known as the "data first" approach). Doing so is controversial. Bhattacherjee calls it a common mistake in research design [4]. Contrarily, Golub stresses the practicality of this approach and its successes in cancer research [14]. However, most participants tried to stay true to their initial enquiry, making only small adjustments when necessary.

In our observational study, we asked the participants to stop the dataset search whenever they think they have finished the search process. There was, however, some variety in how participants interpreted the stopping condition. Although we provided a definition of "data" in the task description, some participants were satisfied with aggregated data (i.e. percentages or diagrams), while others sought raw data with individual data points. This variety in understanding "data" led to a variety of retrieved data types. Furthermore, some participants limited the scope of their data search to browsing the web in search of suitable data. At the same time, one participant also considered finding good collaborators to be part of the data search process. We believe that this diversity is a strength of our study, allowing us to find out what the participants thought dataset search should mean, instead of limiting them to our pre-conceived notions. However, it does make the interpretation more complicated, as the findings are broad and multifaceted. Although all of the participants were social scientists (though at various stages of their academic career), we could observe a variety of strategies and methodologies employed. We do believe that a similar diversity exists within other disciplines as well.

8 Conclusion

This paper presents the results of an exploratory study on information-seeking behaviour during dataset search. Our study includes observations and semi-structured interviews with 12 participants from the social sciences. We provide new insights regarding how dataset search is performed in the wild and identified challenges that social scientists encounter during dataset search. Our seven key findings directly or indirectly relate to the FAIR guiding principles [43], covering findability, accessibility, interoperability, and reusability.

Findability implies globally unique and persistent identifiers used for both datasets and literature in its first principle *F1 ((Meta)data are assigned a globally unique and persistent identifier)*. In our study, we found that while the participants did not use these identifiers all that much, it would have been helpful to facilitate links between literature and datasets (see Finding 1 and 6). For *F4 ((Meta)data are registered or indexed in a searchable resource)*, we found that care should be taken that this searchable resources themselves are easily discoverable (see Finding 2). There is no point in having a dataset registered when the registry itself is hard to discover along the typical search paths. Per Finding 7, it cannot be expected that researchers just know where to find the best information.

Accessibility is often hindered by time-consuming access procedures (Finding 5). Stronger adherence to *A1 ((Meta)data are retrievable by their identifier using a standardised and open communications protocol)* would alleviate that problem.

Interoperability is about the need to integrate data with other data and allowing for interoperation with applications and workflows. *I1 ((Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation)* is often not met, creating difficulties to the researchers in finding the information they need (Finding 4). Likewise, *I3 ((Meta)data include qualified references to other (meta)data)* is sorely lacking leading us to Finding 6.

Reusability is at stake when *R1 ((Meta)data are richly described with a plurality of accurate and relevant attributes)* is not met (Finding 4). Adaptation of the DDI standard[45] for empirical social science data, as recommended in *R1.3 ((Meta)data meet domain-relevant community standards)*, might help in that regard. Due to the complexity of relevant metadata, care should be taken that the descriptions are easy to navigate.

More advanced search infrastructures that can capture the complexity of concepts and their relations, such as variables and corresponding questions, are needed. At the European level, an effort in this direction is the CESSDA Euro Question Bank[46], which focuses on the question level and links to corresponding studies in the CESSDA Data Catalogue[47].

From a social scientist's viewpoint, strategies to mitigate the challenges in dataset search should focus on infrastructures, tools, and education: Data search should become a first-class citizen in the social scientist toolbox. It should become part of social science curricula and professional education for individuals working with research datasets. Teaching dataset search should include raising awareness for existing services, hands-on advice on how to assess the quality of datasets accurately, and help on comparing datasets.

Acknowledgements This work was partly funded by the DFG, Grant no. 388815326 (VACOS Project) and Grant No. 410845317 (ConDATA Project) at GESIS and the University of Duisburg-Essen.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Athukorala, K., Hoggan, E., Lehtiö, A., Ruotsalo, T., Jacucci, G.: Information-seeking behaviors of computer scientists: Challenges for electronic literature search tools. Proc. Am. Soc. Inf. Sci. Technol. **50**(1), 1–11 (2013). https://doi.org/10.1002/meet.14505001041
2. Bates, M.J.: Where should the person stop and the information search interface start? Inf. Process. Manage. **26**(5), 575–591 (1990). https://doi.org/10.1016/0306-4573(90)90103-9
3. Belkin, N.J., et al.: Interaction with texts: Information retrieval as information seeking behavior. Inf. Retr. **93**, 55–66 (1993)
4. Bhattacherjee, A.: Social science research: Principles, methods, and practices. (2012). URL: https://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1002&context=oa_textbooks

[45] https://ddialliance.org/.

[46] https://www.cessda.eu/About/Projects/Work-Plans/Work-Plan-2019#eqb19.

[47] https://datacatalogue.cessda.eu/.

5. Borlund, P.: The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. Information research 8(3), (2003)
6. Borlund, P., Dreier, S.: An investigation of the search behaviour associated with Ingwersen's three types of information needs. Inf. Process. Manage. **50**(4), 493–507 (2014). https://doi.org/10.1016/j.ipm.2014.03.001
7. Case, D.O.: Looking for information: A survey of research on information seeking, needs, and behavior (4th edition). *Journal of the Association for Information Science and Technology*, 68(9):2284–2286, (2017). URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23778, https://doi.org/10.1002/asi.23778
8. Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.D., Kacprzak, E., Groth, P.: Dataset search: A survey. VLDB J. **29**(1), 251–272 (2019). https://doi.org/10.1007/s00778-019-00564-x
9. Cohen, T., Roberts, K., Gururaj, A.E., Chen, X., Pournejati, S., Alter, G., Hersh, W.R., Demner-Fushman, D., Ohno-Machado, L., Xu, H.: A publicly available benchmark for biomedical dataset retrieval: the reference standard for the,?: biocaddie dataset retrieval challenge. Database **2017**, 2017 (2016). https://doi.org/10.1093/database/bax061
10. Curty, R.G.: Factors influencing research data reuse in the social sciences: an exploratory study. IJDC **11**(1), 96–117 (2016). https://doi.org/10.2218/ijdc.v11i1.401
11. Darby, R., Lambert, S., Matthews, B., Wilson, M., Gitmans, K., Dallmeier-Tiessen, S., Mele, S., Suhonen, J.: Enabling scientific data sharing and re-use. In *2012 IEEE 8th International Conference on E-Science*, pages 1–8. IEEE, (2012). https://doi.org/10.1109/escience.2012.6404476
12. Ellis, D.: A behavioural approach to information retrieval system design. *Journal of Documentation*, 45(3):171–212, (1989). URL: https://www.learntechlib.org/p/141664
13. Friedrich, T.: *Looking for data*. PhD thesis, Humboldt-Universität zu Berlin, Philosophische Fakultät, (2020). https://doi.org/10.18452/22173
14. Golub, T.: Counterpoint: data first. Nature **464**(7289), 679 (2010). https://doi.org/10.1038/464679a
15. Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., Wyatt, S.: Searching data: A review of observational data retrieval practices in selected disciplines. *Journal of the Association for Information Science and Technology*, (2019). URL: https://pure.knaw.nl/portal/files/9478410/1707.06937v3.pdf
16. Gregory, K.M., Cousijn, H., Groth, P., Scharnhorst, A., Wyatt, S.: Understanding data search as a socio-technical practice. *Journal of Information Science*, page 0165551519837182, (2019). 10.1177/0165551519837182
17. Hedrick, T.E.: Justifications for the sharing of social science data. Law and Human Behavior **12**(2), 163–171 (1988). https://doi.org/10.1007/bf01073124
18. Hemminger, B.M., Lu, D., Vaughan, K., Adams, S.J.: Information seeking behavior of academic scientists. J. Am. Soc. Inf. Sci. Technol. **58**(14), 2205–2225 (2007). https://doi.org/10.1002/asi.20686
19. Hughes, B., Joshi, I., Lemonde, H., Wareham, J.: Junior physician's use of web 2.0 for information seeking and medical education: a qualitative study. International journal of medical informatics **78**(10), 645–655 (2009). https://doi.org/10.1016/j.ijmedinf.2009.04.008
20. Kelly, D.: Methods for evaluating interactive information retrieval systems with users. Inf. Retr. **3**(1–2), 1–224 (2009)
21. Kern, D., Mathiak, B.: Are there any differences in data set retrieval compared to well-known literature retrieval? In *International Conference on Theory and Practice of Digital Libraries*, pages 197–208. Springer, (2015). https://doi.org/10.1007/978-3-319-24592-8_15
22. Kim, J.: Describing and predicting information-seeking behavior on the web. J. Am. Soc. Inf. Sci. Technol. **60**(4), 679–693 (2009). https://doi.org/10.1002/asi.21035
23. Koesten, L.M., Kacprzak, E., Tennison, J.F.A., Simperl, E.: The trials and tribulations of working with structured data: -a study on information seeking behaviour. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 1277–1289, New York, NY, USA, 2017. Association for Computing Machinery. https://doi.org/10.1145/3025453.3025838
24. Kuhlthau, C.C.: Inside the search process: information seeking from the user's perspective. J. Am. Soc. Inf. Sci **42**(5), 361–371 (1991)
25. Kules, B., Capra, R.: Creating exploratory tasks for a faceted search interface. *Proc. of HCIR 2008*, pages 18–21, (Jan. 2008). URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.596.8865&rep=rep1&type=pdf
26. Kules, B., Capra, R., Banta, M., Sierra, T.: What do exploratory searchers look at in a faceted search interface? In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 313–322, (2009). https://doi.org/10.1145/1555400.1555452
27. Marchionini, G.: Exploratory search: from finding to understanding. Commun. ACM **49**(4), 41–46 (2006). https://doi.org/10.1145/1121949.1121979
28. Marée, R., Rollus, L., Stévens, B., Hoyoux, R., Louppe, G., Vandaele, R., Begon, J.-M., Kainz, P., Geurts, P., Wehenkel, L.: Collaborative analysis of multi-gigapixel imaging data using cytomine. Bioinformatics **32**(9), 1395–1401 (2016). https://doi.org/10.1093/bioinformatics/btw013
29. Meho, L.I., Tibbo, H.R.: Modeling the information-seeking behavior of social scientists: Ellis's study revisited. J. Am. Soc. Inf. Sci. Technol. **54**(6), 570–587 (2003). https://doi.org/10.1002/asi.10244
30. Murillo, A.P.: Examining data sharing and data reuse in the dataone environment. Proc. Am. Soc. Inf. Sci. Technol. **51**(1), 1–5 (2014). https://doi.org/10.1002/meet.2014.14505101155
31. Niu, X., Hemminger, B.M., Lown, C., Adams, S., Brown, C., Level, A., McLure, M., Powers, A., Tennant, M.R., Cataldo, T.: National study of information seeking behavior of academic researchers in the united states. J. Am. Soc. Inf. Sci. Technol. **61**(5), 869–890 (2010). https://doi.org/10.1002/asi.21307
32. Pasquetto, I.V., Borgman, C.L., Wofford, M.F.: Uses and Reuses of Scientific Data: The Data Creators' Advantage. *Harvard Data Science Review*, 1(2), 11 (2019). https://hdsr.mitpress.mit.edu/pub/jduhd7og. https://doi.org/10.1162/99608f92.fc14bf2d
33. Perspectives, K.: Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. *SCARP Synthesis Study, Digital Curation Centre*, (2010). URL: https://www.dcc.ac.uk/sites/default/files/SCARPSYNTHESIS_FINAL.pdf
34. Pontis, S., Blandford, A.: Understanding influence: an exploratory study of academics processes of knowledge construction through iterative and interactive information seeking. J. Assoc. Inf. Sci. Technol. **66**(8), 1576–1593 (2015). https://doi.org/10.1002/asi.23277
35. Pontis, S., Blandford, A., Greifeneder, E., Attalla, H., Neal, D.: Keeping up to date: an academic researcher's information journey. J. Assoc. Inf. Sci. Technol. **68**(1), 22–35 (2017)
36. Rieh, S.Y.: Judgment of information quality and cognitive authority in the web. J. Am. Soc. Inf. Sci. Technol. **53**(2), 145–161 (2002). https://doi.org/10.1002/asi.10017
37. Takai, S., Ishii, K.: A use of subjective clustering to support affinity diagram results in customer needs analysis. Concurr. Eng. **18**(2), 101–109 (2010). https://doi.org/10.1177/1063293X10372792
38. Watteler, O.: *Grundlagen sozialwissenschaftlichen Arbeitens: eine anwendungsorientierte Einführung*, chapter Recherche nach sozialwissenschaftlichen Forschungsdaten, pages 127–155. UTB, (2017)

39. Wen, L., Ruthven, I., Borlund, P.: The effects on topic familiarity on online search behaviour and use of relevance criteria. In *European Conference on Information Retrieval*, pages 456–459. Springer, (2006). https://doi.org/10.1007/11735106_40
40. White, R.W., Drucker, S.M.: Investigating behavioral variability in web search. In *Proceedings of the 16th international conference on World Wide Web*, pages 21–30, (2007). https://doi.org/10.1145/1242572.1242576
41. White, R.W., Roth, R.A.: *Exploratory search: Beyond the query-response paradigm*. Number 3. Morgan & Claypool Publishers, (2009). https://doi.org/10.2200/s00174ed1v01y200901icr003
42. Wildemuth, B.M., Freund, L.: Assigning search tasks designed to elicit exploratory search behaviors. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, pages 1–10, (2012). https://doi.org/10.1145/2391224.2391228
43. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Sci. Data (2016). https://doi.org/10.1038/sdata.2016.18
44. Yoon, A.: Making a square fit into a circle: Researchers' experiences reusing qualitative data. Proc. Am. Soc. Inf. Sci. Technol. **51**(1), 1–4 (2014). https://doi.org/10.1002/meet.2014.14505101140