Are we building the data discovery infrastructure researchers want? Comparing perspectives of support specialists and researchers

Guangyuan Sun
National Institute of Education, Nanyang Technological University, Singapore. https://orcid.org/0000-0001-7352-2158

Tanja Friedrich
German Aerospace Center, Scientific Information, Cologne, Germany. https://orcid.org/0000-0003-1557-3728

Kathleen Gregory
University of Vienna, Vienna, Austria; School of Information Studies and Scholarly Communications Lab, University of Ottawa, Ottawa, Canada. https://orcid.org/0000-0001-5475-8632

Brigitte Mathiak
GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany. https://orcid.org/0000-0003-1793-9615

Abstract
Purpose Data discovery practices currently tend to be studied from the perspective of researchers or the perspective of support specialists. This separation is problematic, as it becomes easy for support specialists to build infrastructures and services based on perceptions of researchers’ practices, rather than the practices themselves. This paper brings together and analyze both perspectives to support the building of effective infrastructures and services for data discovery.

Methods This is a meta-synthesis of work the authors have conducted over the last six years investigating the data discovery practices of researchers and support specialists, like data librarians. We bring together data collected from in-depth interview studies with 6 support specialists in the field of social science in Germany, with 21 social scientists in Singapore, an interview with 10 researchers and 3 support specialists from multiple disciplines, a global survey with 1630 researchers and 47 support specialists from multiple disciplines, an observational study with 12 researchers from the field of social science and a use case analysis of 25 support specialists from multiple disciplines.

Results We found that while there are many similarities in what researchers and support specialists want and think about data discovery, there are some differences we have identified, most notably the interconnection of data discovery with web search, literature search and social networks.

Conclusion We conclude by proposing recommendations for how different types of support work can address these points of difference to better support researchers’ data discovery practices.

Keywords
Research Data Management, Data Discovery, Research Data Services, Data Reuse, Meta-synthesis

Author contributions:
All authors contributed to the study conception and design. All authors contributed the methods and findings of their respective studies. The first draft and consecutive reviews were done collaboratively by all authors. All authors contributed equally to the discussion and conclusion. Literature review was primarily done by Guangyuan Sun. Guangyuan Sun also developed the typology of support work.
1. Introduction

Research data management support is an increasingly visible component in the host of support services provided by universities and research institutions. While data librarians, archivists and educational specialists provide support with managing and sharing data [1–4], they are also interested in supporting practices afforded by well-managed data, such as data discovery.

Data discovery practices currently tend to be studied from the perspective of researchers or the perspective of support specialists. While work from both perspectives aims to inform the design of tools, infrastructures and services, these perspectives often exist separately from each other in the literature.

Studies from the perspective of researchers and data seekers situate data discovery and reuse in relation to disciplinary norms or types of data [5–8]; the use of particular repositories or data portals [9, 10]; or the influence of trust, reputation and ease of access on the process [11–13].

Work conducted from the perspective of data support specialists focuses on typical workflows [3, 14, 15], tasks [16, 17] and the details of infrastructural projects (e.g., the FREYA project1, or the GeRDI project2) which are designed to support data discovery and reuse.

With few exceptions (i.e., [1]), these two perspectives exist separately from each other in the literature. This separation is problematic, as it creates an environment where it becomes easy for support specialists to build infrastructures and services based on perceptions of researchers’ practices, rather than the practices themselves. We argue that we need to bring together and analyze both the practices and perspectives of researchers and those of data support specialists in order to build effective, sustainable infrastructures and services to support data discovery.

This paper brings together and re-analyzes data from past work, rooted in both perspectives, in a systematic meta-synthesis. We draw on six of our own studies, conducted over a six-year period, which examine both the data discovery practices of researchers and the work of support specialists to examine the following questions. How do data support specialists work to support researchers’ practices of data discovery? Do the efforts of support specialists support the data needs and discovery practices of researchers, or is there a mismatch between what researchers “need” and the efforts of support specialists? What is done well and what can be done better?

We begin by reviewing the literature to develop a typology of the different types of work done by support specialists. We then re-analyze and synthesize the data from our past studies in light of the above research questions, according to both the perspective of researchers and support specialists. These studies include data from in-depth interview studies [6, 18, 19]; a global survey [20, 21]; an observational study [22] and a use case analysis conducted by information professionals [23, 24]. We conclude by relating five emergent themes from our analysis to our typology of support work, highlighting both existing best practices and identifying areas for improvement.

2. Literature Analysis

2.1 Prior studies from the perspective of researchers

While the focus of this paper is on support specialists, studies from the researcher’s perspective are much more numerous. Many data discovery studies are motivated by studying researchers’ data reuse behaviors [7, 18, 25, 26]. They correlate concepts, such as attitude, motivation and disciplinary norms in data reuse with discovery practices, e.g., through surveys.

---

1 https://www.project-freya.eu/en
2 https://www.gerdi-project.eu/
Qualitative studies, such as [27] investigate workflows of data search in a web-based environment and introduce patterns of web-based data search: discover, explore and analyze. They characterize data discovery as “serendipitous discovery of datasets...by means of descriptive metadata”, where the input is “unstructured free text” and the output is “list of (relevant) datasets”. [28] categorize users’ data-centric tasks into two types: process-oriented (e.g., using data for machine learning processes) or goal-oriented (e.g., using data to answer a question), and identify different user information needs for the different types. [29] study user requirements for users of STEM data repositories and collected nine user requirements. [30] studied the information seeking behavior of 22 environmental scientists with regards to data reuse. They found many scientists performed known-item searches in trusted sources, and that half began their search with a web search on specific keywords or names of the data repository.

Search query and log analysis are a third way to investigate data search behavior e.g., of structured data stored on the web [28], in open government data portals [31] or in research data repositories specifically [32–34]. They yield interesting results, such that data search queries tend to be short and users generally land on data repositories from Google, which suggests that users often start searching using general search engines. However, most of the studies are specific to the source of the log analysis and the results are limited to the repository structure and the type of data stored.

2.2 Prior studies from the perspective of research data support specialists

There are many terms used to describe research data support staff such as data curator, data librarian, data manager, data steward, data specialist and research information scientist [35]. There is a lack of common terminology and a variability of both position titles and actual duties. Research data support work, as we choose to call it, can be both social and technical, involving technical knowledge on research data management (RDM) systems and social knowledge for RDM community building [36].

We identify three primary types of support work from the literature: people-oriented roles (e.g., reference librarians) providing consultations (Section 2.2.1); metadata-related roles (e.g., metadata specialists) providing data documentation services which are non-IT but focuses on enhancing the completeness of research data description and data discoverability (Section 2.2.2); and infrastructure-oriented roles (e.g., research software engineers) providing IT-focused technical tasks (Section 2.2.3). These types of work are not mutually exclusive. A support specialist may play multiple roles, depending on his/her job requirements. Our proposed categories are nonetheless useful as a way of structuring our discussion about RDM support work.

2.2.1 People-oriented support work

People-oriented research support work varies between organizations. It depends on technical infrastructure, expertise of staff, and the organization’s value, culture and structure [37]. At the libraries of top universities, such support work includes the topics of research data management, open access, scholarly publishing, research impact measurement, developing research guides, and providing research consultations and research tools recommendations [38]. Some of this work can be seen as an extension of traditional library services to research data [3]. Tammaro et al. [35] further classify skills of data curators into four categories: data management (e.g., to advise on file naming conventions), data description and documentation (e.g., to provide consultation on metadata creation following standardized schemas), data deposit/publishing (e.g., to help with data anonymization) and archiving and preservation.

Cox et al. [39] found that, across countries, libraries have provided leadership in RDM. But there are also alternative organizational models. While research support staff are often recruited from existing library staff, non-library professionals play an important role as well to serve specific communities [35]. In the Netherlands, so-called data stewards are university employees, sometimes library staff, who generally hold a PhD degree in a research domain and are appointed to a school or department to engage researchers with in-depth discussions on RDM [40]. Data champions are volunteer domain
researchers that are already embedded within research units, who are equipped with necessary research data knowledge and advocate good RDM [41].

In [39], they distinguish between advisory services and technical services, and find the former, which we consider people-oriented work, is overall more wide-spread than the latter. The main types of advisory services include providing advice on data management plans, RDM training, maintaining a website of local RDM advice and useful resources, and advice on copyright. The study finds that such services were extensively developed across libraries. Advisory services also include providing advice regarding the “awareness/retrieval of reusable data sources”; this type of service was evaluated as being “patchy”—it was not among the most common advisory services in the investigated libraries. Similarly, a more recent study also finds that library support staff tend to place more value on consultation services, as some believe such services are more valuable to researchers than infrastructure-oriented or metadata-related support [42].

People-oriented support work is dynamic in its requirements. Research support services need to include guiding researchers in describing their datasets with adequate metadata [43], see also section 2.2.2. Personalizing comprehensive and executable data management workflows for individual research groups is a very recent trend which includes tasks like drafting file organization schemes and creating data management roles within the team [44]. [3] identify demands in new services related to text and data mining and artificial intelligence. [45] observe that the support services for doctoral students and scientists provided by libraries follow a steady trend, with many subtle changes, particularly on the technological basis. Including data visualization services in top university libraries is investigated and discussed in [46].

In terms of geographic landscape, people-oriented research data services by academic libraries are more advanced in North America and Europe [3]. RDM services are spreading to other parts of the world where, in general, people-oriented work is still in development and the focus is rather on making policy, raising awareness and conceiving RDM IT infrastructure [47, 48].

2.2.2 Metadata-related Support Work

Rich metadata is relevant for the discoverability of both large and small datasets. Information professionals, including research data support specialists, have formed and are active members of metadata-related working groups to develop and promote standardized documentations to enhance data discovery and harvesting on the web, especially when data can be connected to other materials and resources that are also published online.

Such groups include the Controlled Vocabularies Working Group3 and the XKOS - Extended Knowledge Organization System4 under the Data Documentation Initiatives (DDI)5 or specialized groups within the Research Data Alliance, DataCite or the GO FAIR Implementation Network.

Within these groups, various knowledge representation techniques are developed and are later employed by repositories. Regarding metadata, both generic and domain-specific standards have been developed. DataCite6 is a generic metadata standard promoting research data citation and has been widely adopted by data archives for imputing metadata about datasets, as well as by infrastructures harvesting metadata from those archives. The DDI metadata standard is designed for the complete and accurate documentation of quantitative social science datasets to facilitate data discovery and metadata interoperability across different repositories [49]. It is now used by research data centers.

---

3 https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/39911435/Controlled+Vocabularies+Working+Group
4 https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/826408976/XKOS+-+Extended+Knowledge+Organization+System
5 https://ddialliance.org/
6 https://schema.datacite.org/
worldwide (e.g., Inter-university Consortium for Political and Social Research\textsuperscript{7}, Finnish Social Science Data Archives\textsuperscript{8}) as the standardized protocol for creating metadata of deposited research data.

*Data Catalog Vocabulary (DCAT)* is a W3C standard, and an RDF vocabulary originally designed for facilitating federated search of governmental data catalogs published on the web, and has now been extended to describe research data \cite{50,51}.

The more recent techniques of Linked Data and knowledge graph are also under exploration, though not yet been implemented in data repositories on a large-scale. Research data can be linked to research papers for better discoverability. Burton et al. \cite{52} develop the Scholix (Scholarly Link eXchange) Framework to model links between research articles and supporting data. Khan et al. \cite{53} conducted a case study to quantitatively test the usage of Scholix-based API (i.e. Scholexplorer) in identifying linkages between the academic literature output of researchers in University of Bath and datasets published in external data repositories.

Research data can also be linked to other types of resources. Aryani et al. \cite{54} create a Research Graph database of identity and some basic descriptive information of research datasets, of researchers (e.g., ORCID), of publications (e.g., title, DOI) and of research grants (e.g., title, participant list) to support linking research datasets stored in different data repositories. Wang et al. \cite{55} explored the conversion of records in the same Research Graph to Linked Data format to enhance their accessibility to third-party web services over the Internet, thus indirectly increasing research data discoverability. They use schema.org for the conversion, which is a widely adopted standard for marking up structured data in web pages and promoted by major search engines such as Google \cite{56}. In fact, the library world has started using Schema.org to make metadata of books and papers in library catalogs more discoverable on the web as early as 2012 \cite{57}.

Despite the numerous activities listed above, many libraries do not provide technical services for "creating/transforming metadata for data", which would require knowledge and skills in standardized metadata schemata \cite{39}.

### 2.2.3 Infrastructure-oriented support work

Research data management inevitably involves building and maintaining digital infrastructure that caters to user needs across the research data lifecycle. Such work, which is more technical in nature, is typically performed by support staff who are IT-oriented.

There are a few studies on the characteristics of research data infrastructure and related support work \cite{17,58}, collecting user requirements for these infrastructures \cite{59} and developing RDM related tools or software \cite{60,61}. Typically, these studies make recommendations on what infrastructure-oriented type of support work should entail.

Other infrastructure-oriented work, which also has overlaps with more metadata-oriented tasks, includes: the development of federated research data repositories to support cross-disciplinary discovery of research data \cite{62}; the design and implementation of discovery tools, such as DataCite Commons, that allow simple searches by works, individuals and organizations, while providing users with a detailed overview of the relationships between the entities in the research setting \cite{63}; and the creation of data catalogs to support the discovery of sensitive research data that requires special attentions paid to legal, ethical, or authorial constraints \cite{64,65}.

Khan et al. \cite{17} found that the major concern across all repositories was long-term maintenance of the repository infrastructure and lack of engagement from the users. Discovery of data was rated as a future need instead of top priorities. A similar finding by Smit \cite{66} is that technical service providers’

\textsuperscript{7} https://www.icpsr.umich.edu/web/pages/index.html
\textsuperscript{8} https://www.fsd.tuni.fi/en/
main goal focused on the maintenance of research data infrastructure—to assure its long-term availability to support data sharing and storage, rather than concentrating on making the data available.

Infrastructure-oriented support work may need to devote more effort in catering to the special needs of research data, compared with other types of resources, regarding improving discoverability. Bugaje and Chowdhury [58] evaluated different types of research data repositories—including disciplinary, institutional, location-based (e.g., ANDS Research Data Australia), content-specific (i.e. housing data of a certain file type), publisher-service (i.e. provided by journal publishers), and general-purpose (e.g., Figshare, Zenodo), and point out that existing data discovery infrastructures lack design attending to the special requirements of research data, but rather “makeshift adaptations” of the classic information retrieval systems.

Another stream of work that needs more attention may be to design different data infrastructures based on peculiarities of different data formats. Hedeland [67] focuses on infrastructure requirements for hosting audio-visual linguistic research data, and identifies a gap between what is needed in the process of digitization (i.e. domain-specific research methods) and what is provided (i.e. approaches of generic research data management). He posits that responsibilities of technical support staff should include closing this gap.

[66] uncovered that the software underlying research data infrastructure is often created and maintained by support staff who have not received formal training in software engineering, or created at organizations of which their primary tasks do not include software development to begin with. Like [66], the authors of this paper are well aware of the phenomenon that research data software is often not developed by software engineers, but domain users. An example is the well-known Dataverse software, which was developed by a Social Scientist to manage Social Science data. While the advantage of this approach is that the developers are very close to the domain and have insights into what users might want, they may be less familiar with surrounding technical issues, such as how to improve discovery.

2.3 Summary

Most RDM service providers tend to focus on people-oriented research support work like RDM advisory services and training for users. Advice on discovering reusable data on the Web are provided only by some libraries. The metadata-focused support work is, in most cases, an intellectual collaboration of support staff with other information professionals to develop standardized documentation aiming to improve the discoverability of data on the Web by both humans and machines. The infrastructure-oriented support work currently focuses on ensuring the robust and sustainability of the infrastructure, and sometimes lacks engagement with and incorporation of the need of end users of the service (i.e., researchers).

There is also a gap in research investigating the actual support work provided and the services needed by users. In order to address this issue, we need to understand perspectives from both researchers and support staff. The rest of this paper compares the differences in perspectives by doing a meta-synthesis of work we have conducted in the past six years to draw insights that may inform the future development of support work for data discovery.

3. Method

In this section, we briefly discuss the methodologies used in each individual study in our meta-synthesis (see Table 1). We then discuss our process for identifying common themes across the studies.
Table 1. Methodologies and study descriptions for data used in meta-analysis

| Study number | Methodology                     | Participants                                      | Disciplinary focus | Year of data collection | Published work |
|--------------|---------------------------------|---------------------------------------------------|--------------------|-------------------------|-----------------|
| Study 1      | Interview                       | Researchers (n=21)                                | Social sciences    | 2017                    | [19]            |
| Study 2      | Interview                       | Researchers (n=19); Support professionals (n=3)   | Multiple disciplines| 2017                    | [18]            |
| Study 3      | Interview                       | Support professionals (n=6)                       | Social sciences    | 2016                    | [6]             |
| Study 4      | Survey                          | Researchers (n=1630); Support professionals (n=47) | Multiple disciplines| 2018                    | [20, 21]        |
| Study 5      | Observational study             | Researchers (n=12)                                | Social sciences    | 2019                    | [22]            |
| Study 6      | Use case analysis by support professionals | Use cases (n=100 use cases collected; n=25 support professionals) | Multiple disciplines| 2020                    | [24]            |

3.1. Methodologies for individual studies

3.1.1. Study 1: Interviews with researchers

Study and analysis description

In Sun [19], twenty one face-to-face semi-structured interviews were conducted between August and December 2017. Participants were recruited through convenience sampling to explore the usefulness and challenges of reusing quantitative social science data. Participants discussed their experiences with data reuse and performed observed searches in the Inter-university Consortium for Political and Social Research (ICPSR) data repository to find data meeting their needs. The interview protocol focused on identifying the type of information participants used to search, browse, and assess the reusability of research data. The interviews were audio-recorded and transcribed. Transcripts were analyzed using the open coding and axial coding techniques introduced by Corbin and Strauss [68].

Sample description

Twenty-one faculty members and research staff at two universities in Singapore (Nanyang Technological University and National University of Singapore) were recruited via email invitation. All participants were social scientists using quantitative research methods, had interest in or experience with reusing quantitative research datasets, and self-reported working in different domains (sociology (n=11), communication studies (n=6), international studies (n=2) and business (n=2)).

3.1.2. Study 2: Interviews with researchers and support specialists

Study and analysis description

As further described in Gregory et al. [18], twenty-two one-hour, semi-structured interviews were conducted between October and December 2017, either virtually or in person. Participants (n=19) were recruited via email from a pool of 186 individuals who had visited the DataSearch portal (now Mendeley Data9) and had indicated willingness to provide feedback. An additional three participants were recruited via purposive sampling. The interview protocol focused on eliciting information about participants’ data needs, sources and strategies used for finding data, and their criteria for evaluating data which they discover. Audio-recordings were made of all interviews. Detailed summaries, including quotations with timepoints, were made following a protocol for every interview. Summaries were thematically coded and analyzed using QDA Miner Lite.

9 https://data.mendeley.com/
Sample description

Nineteen participants identified as being researchers; three participants worked as support specialists. Respondents were employed in 12 countries; the most frequently represented were the United States (n=6) and the Netherlands (n=3). Participants self-identified their broad disciplinary area and computer science (n=3) and information science (n=3) were the most common. Participants also had varying levels of professional experience: early career (0–5 years, n=5), mid-career (6–15 years, n=10), experienced (16 + years, n=6) and retired (n=1).

3.1.3. Study 3: Interviews with support specialists

Study and analysis description

As described in Friedrich [6], six support specialists from the German GESIS data archive for the social sciences were interviewed regarding their interactions and experiences with users of the data archive in June and July 2016. Following constructivist grounded theory methodology, the participants were sampled by initial and theoretical sampling, and the interviews were designed as unstructured conversations exploring eight topics: the data users’ educational/professional background; the users' research experience and data literacy; information sources and channels; the roles of intermediaries and information technology; goals, needs and purposes of users; requirements when looking for data; the role of documentation; barriers and problems when looking for data. The shortest interview lasted about 52 minutes, the longest 1 hour and 36 minutes. Audio recordings and transcripts were made of all interviews. The transcripts were analyzed according to constructivist grounded theory methodology, using the software atlas.ti.

Sample description

All six interviewees were employees of the same large social science data archive. Two participants were help desk staff who received requests via the general helpdesk line. The other four were specialized experts for data from one or more complex large-scale social science surveys. The participants had been working as data support staff for different periods of time. The most junior staff member had been working in their position for 3 years, the most senior staff for 25 years.

3.1.4. Study 4: Survey with researchers and support specialists

Study and analysis description

As described in Gregory et al. [21], a global survey was conducted to investigate practices of data discovery and reuse in September-October 2018. Respondents were recruited by emails sent to a random sample of 150,000 authors who published an article indexed in the Scopus literature database between 2015-2018. An additional 40 participants were recruited by posting to discussion lists in the library and research data management communities for a total sample of 1677 complete responses. The survey consisted of a maximum of 28 individual items, with 9 items allowing for multiple responses. Responses were recorded anonymously and were analyzed using R [69], the data from the survey are openly available at Gregory [20]. This survey was descriptive in nature; responses only describe the practices of people completing the survey.

Sample description

A total of 1630 researchers and 47 research support specialists completed the survey. Both the majority of researchers (68.4%) and support specialists (74.47%) reported working at a university or college, followed by those working at a research institution (17.18%, 14.89%). The majority of both researchers (40.18%) and support specialists (46.81%) have been active professionally for 6-15 years. The second largest group of researchers had 16-30 years of experience (30.25%), while the second largest group of support specialists had 0-5 years of experience (27.66%). The most selected
disciplinary domains of researchers were engineering and technology, biological sciences, medicine and social sciences. Among support specialists, the most selected domains were information science, environmental science, multidisciplinary research and social science.

3.1.5. Study 5: Observations of researchers

Study and analysis description

An observational study described in Krämer et al. [22] was conducted with 12 social scientists recruited through convenience sampling to learn about their data-seeking behaviors. It consisted of an initial survey, an observational period in which screen recordings were collected while participants performed a simulated data search task, and a follow-up guided interview. The audio material was transcribed. Screen captures were transcribed to show interactions with the computer in an abstract way, i.e., using web search, university library web page, etc. The interactions were also analyzed separately for patterns. The merged transcriptions were analyzed using affinity diagrams.

Sample description

At the time of the study, eleven participants were affiliated to a university and one to a public research facility. Four of the participants were professors, three were postdocs, two were PhD students, and three were master’s students. Their primary research areas were sociology (n=6), political science (n=4), economic sociology (n=1), and media management (n=1).

3.1.6. Study 6: Use case analysis and prioritization by support specialists

Study and analysis description

As described in [24], the GOFair Data Discovery Implementation Network10 collected over 100 use cases for data discovery by polling researchers and support specialists and by consolidating existing lists, i.e., [23]. These use cases were then clustered around common themes and requirements. A survey tool administered during the GOFAIR meeting (February 3rd, 2021) was used to gain consensus about the identified themes. Meeting attendees were asked to answer a demographic survey (position, domain and geographical location) and to rank the two most and least ‘relevant’ use case clusters.

Sample description

25 out of approximately 50 meeting attendees completed the survey. Most respondents were from mainland Europe (n=17), with three additional participants from the UK (n=3). Of those mainland participants, most were located in Germany (n=8), Scandinavian countries (n=4) and the Netherlands (n=3). Three attendees were from North America (2 US, 1 Canada), and one from South America (Brazil). The majority of respondents identified themselves as infrastructure providers (n=17), with eight respondents indicating that they produce data themselves. Most respondents were involved in the life sciences (n=8), general science (n=7), or natural sciences (n=5). Fewer worked in social science (n=3) and computer science.

3.2 Methodology for meta-synthesis

As authors of the above studies, we are all deeply familiar with the methodologies and resulting data. In order to synthesize the data from across all studies, we conducted four phases of analysis. First, we engaged in an exploratory phase, where we each took our individual expertise as a starting point. We each re-examined the data from our own studies above, with a focus on supporting professionals’ data discovery and reuse practices and attitudes and our research aims. We then summarized our findings.

---

10 https://www.go-fair.org/implementation-networks/overview/discovery
textually and graphically and shared them with each other before collectively identifying themes in our individual analysis. Finally, we compared these individual themes with each other and identified five cross-cutting themes which we use to present our findings below. In our conclusion we discuss these findings using the typology of support specialists that we have introduced in the literature review.

4. Findings

Our findings are arranged in five thematic clusters: challenges and problems in data seeking; data needs and purposes; the role of web search in data discovery; the role of academic literature; and the role of social connections. Within each cluster, we present findings from both the researcher perspective and the support specialist perspective and provide a short synthesis identifying commonalities and points of difference.

4.1 Challenges and problems in data seeking

4.1.1. Researcher perspective

Seventy-three percent of researchers responding to the global survey (Study 4) indicated that finding data is sometimes a challenging task, while 19% of respondents believed it to be difficult. Survey respondents identified specific challenges, including the inability to download or access data (68% of respondents) and the distributed nature of data across different locations (49%). Other identified challenges included a lack of adequate data search systems, a lack of data search literacy, or, as found in the semi-structured interviews reported in (Study 2), a lack of digital and discoverable data.

At the same time as reporting these challenges, interviewees in Study 2 felt that they were ultimately satisfied with their discovery strategies, although they recognized that they need to modify their expectations or employ multiple strategies (including consulting their social networks and personal connections) to locate the data which they use. Researchers also reported seeking help from colleagues to design efficient queries; others educate their own colleagues about sources and strategies to use when searching for data online. These more ‘social’ strategies, which are discussed further in Section 4.5, also have challenges. Study 2, e.g., found that relying on social networks alone can lead to the creation of ‘filter bubbles’ where researchers may only find data available from within their own circle of peers.

In the pre-observation questionnaire used in Study 5, it was found that finding data is perceived to be difficult in comparison to literature search, which is usually perceived to be of low or medium difficulty. This is similar to results in [70]. One of the challenges identified in the same questionnaire is low data search literacy. During the interviews in Study 5 more challenges emerged: there were not enough links between datasets and relevant other information sources, i.e., other datasets, literature and documentation. Limited accessibility was also mentioned as a challenge to a successful dataset search.

4.1.2. Support specialist perspective

In Study 3, the interviewed support specialists reported that users requested help for both exploratory search and known-item searches, particularly in locating suitable data sources. Users also requested help when they had problems accessing data. Access-related problems included challenges related to data location and retrieval, cost of access, and restricted access.

The interviewees also discussed problems users face when they have found relevant data but cannot make use of them. These problems fall into three categories: problems with erroneous data, when users find potential errors in the data; problems with documentation, when users need assistance in finding or understanding documentation; and problems with data literacy skills, when users lack the necessary skills to actually use data they have found.
In Study 6, research support specialists were asked to prioritize clusters of data discovery use cases to identify the most urgent areas for future work. The cluster of use cases related to metadata/metadata quality was ranked as being most important, followed by the clusters data citation and convenience, the last of which refers to improving the user experience of data search. Participants were also asked which clusters they felt were least important in data discovery. Here, clusters related to lesser known approaches, such as search within data, linking with persons and machine discoverability were seen as having low importance. What is somewhat contradictory is that while metadata quality was rated as highly important, the documentation cluster was often rated as least important.

4.1.3 Synthesis

Both Studies 4 and 5 found that finding data is challenging for researchers; common reasons included lack of data accessibility and data seeking literacy. Study 4 found additional issues of inadequate data search tools and the fact of data being distributed across multiple locations reported by researchers. In Study 5, researchers described the issue of lack of links between datasets and relevant other information sources. To remedy these data finding issues, Study 2 showed that researchers would turn to their social networks for help, which in turn can lead to the problem of manifesting closed circles or ‘filter bubbles’.

Study 3 reveals that support specialists who provide data services to researchers are aware of the challenges regarding the location of suitable data, data accessibility and data literacy. The study also found that researchers express problems with data quality and documentation quality. Interestingly, in Study 6, support specialists self-assessed metadata/metadata quality, data citation, and data convenience as the top three most important factors supporting data seeking behaviors.

4.2 Data needs and purpose for using data

4.2.1. Researcher perspective

Both Studies 2 and 4 highlight that researchers need multiple types of data, often with very specific resolutions and parameters, across phases of work. In Study 4, for example, researchers classified the type of data which they need using a categorization system from the National Science Board [71]. While the majority of researcher respondents selected observational data, approximately half also selected needing more than one type of data. In Study 4, researchers also indicated that they use data for a variety of purposes, ranging from using data as the basis for a new study to using data for teaching. Figure 1, discussed further in Section 4.2.2, presents the distribution of purposes selected by respondents according to their professional role.

Study 1 identifies ten types of uses social scientists have for reusing data. These uses, which are uses that will be specifically reported in publications, were classified into three categories: to support the conceptualization of the study design; to complement the main study, or to serve as the primary data of the main study. The interviews in Study 2, however, find that data reuse can also be conceptual and not be reported in the final paper, even if data have informed the study design and development.

In Study 5, we observed on multiple occasions that social scientists decided to refine their original research question based on the data they could find. They either shifted their topic or research question, or more often, made other search refinements, such as narrowing a general search query into more specific terms.

4.2.2. Support specialist perspective

The data needs of support specialists are often the data needs of their clients. One of the interviewees in Study 2, e.g., searched for data in the literature about particular technologies such as intravascular ultrasounds to pass on to her clients. Finding data and associated literature is just one way in which support specialists work to meet the data needs of their clients. Support specialists also report
engaging in educational activities such as teaching people how to discover and evaluate data or how to curate and manage their own data (Study 4). Study 3 shows that support specialists get requests from users who need data for a broad array of purposes. The purposes include not only various research-related tasks, but also using data for professional qualification and in journalism.

Support specialists also need data for their own research and projects. Forty percent of support respondents in Study 4 (n=19) reported needing data both for their own research and to support others, while fewer individuals (n=2) need data solely for their own research.

There are some differences between how researchers and support specialists use - or view the use of - secondary data by researchers (Figure 1). In Study 4, respondents were asked about their purposes for seeking data. Researchers were asked “Why do you use or need secondary data?”, while support specialists were asked “Why do you or the people whom you support use secondary data?”

![Fig. 1 Purposes for using secondary data by role: researchers (n=1630) and support specialists (n=47). Percents are percent of respondents. Multiple responses possible.](image)

Approximately 50% of researcher respondents use data for teaching, while nearly 80% of support specialist respondents selected this use. In a separate question, support specialists indicated that they help others through teaching; the high frequency of using data for teaching indicated in Figure 1 mirrors this practice. Similarly, support specialists selected using data to make summarizations or visualizations at a higher percentage than researchers; this could be attributable to their own teaching practices also.

Support specialists also indicated that they integrate data (or believe that researchers integrate data) at higher percentages than did researchers. This could reflect the fact that data integration can be a key step in curatorial workflows for certain data, e.g., long-running survey data. It could also indicate a potential mismatch between how support specialists view the activities of researchers, as data integration is often discussed in both the literature and efforts around FAIR data, which is perhaps reflected in the answers of support specialists.
Support specialists in Study 3 described the data needs of users seeking survey data, as presented in support requests. This data was used to identify six categories of data needs, as presented in Table 2 below. Interviewees in Study 3 reported that requests with only few specifics or very general requests require more support than very specific requests.

Table 2. Data needs of users seeking survey data

| Data need                              | Example requests                                      |
|----------------------------------------|-------------------------------------------------------|
| General need for any kind of data      | “What data do you have?”                               |
| Need for new research ideas            | “Do you have data on any exciting topics?”              |
| Need for data on a specific topic      | “Which new or novel data do you have on this topic?”   |
| Need for known data                    | “Someone recommended this data to me; do you have it?” |
|                                        | “Do you have an updated version of this data?”         |
| Need for specific variables in known data | “I know the data I need, but I can’t find the variable I am looking for.” |
| Need for data with specific features   | “I need longitudinal data / linked data / comparative data.” |

4.2.3. Synthesis

Studies 1-5 provide evidence that researchers have multiple, evolving data needs, ranging from the more general to the more specific. These data needs are informed by an individual’s intended uses for data (Study 1, 2, 3, 4); not all of these uses may be cited in a research publication (Study 1, Study 2), some data use happens outside academic research (Study 3). Although there is some evidence that support specialists need data for their own research and work-related activities, i.e., teaching or creating summarizations (Study 4), the data needs of support specialists commonly reflect the data needs of researchers (Studies 2, 4, 5). Study 4, however, suggests that support specialists’ perceptions and researchers’ actual data needs may not always be in alignment.

4.3 The role of web search in data discovery

4.3.1. Researcher perspective

Across disciplines, general web search engines, such as Google or Bing, are important sources for researchers looking for secondary data. 89% of researchers in Study 4 report using general web search engines occasionally or often to locate data for reuse. They use web search both to conduct known-item searches, to locate data repositories, and to search for new datasets, a finding which was also supported in the interviews conducted in Study 2.

Study 2 found that the researchers’ success of using web search engines to locate needed data was mixed. This mirrors the findings from Study 4, where 54% of the responses from researchers indicate that searching the web to find data is sometimes successful and sometimes not; 39% of researcher responses indicate that web searches are either successful or very successful.

In Study 1, most social scientists directly searched for known datasets using general web search engines to identify the digital location of the needed data. In total, the 21 interviewees reported three types of online sources to obtain publicly available datasets: 1) the official website of a particular study, such as the Health Information National Trend Survey (n=13); 2) databases hosted by data collection organizations, such as the World Health Organization (n=9), and 3) data repositories such as ICPSR (n=3).

Study 5, the observation study, found that web search was used extensively for a number of purposes. Participants often performed known item searches, i.e., by searching for the names of datasets that
were found in the literature, but they also used web searches to find additional information on datasets, such as documentation, study websites, or related literature. Participants used web search as a quick and convenient way to chain together different resources, which would otherwise not be connected.

4.3.2. Support specialist perspective

There is an overall lack of information in our studies, and in the available literature, which documents how support specialists locate data by searching the web. Study 4 suggests that researchers may rely on general web search engines more heavily than support specialists, as more respondents who are researchers (59%) reported *often* making use of web search engines than did support specialists (40%). This difference becomes much smaller, however, when looking at the percentages for respondents who selected *never* using search engines to find data (11% of researchers and 13% of support specialists).

There was also less variety in how support specialists reported their success using web search engines to find data, with 80% of responses from support specialists indicating that they are sometimes successful/sometimes not successful in their data searches with general web search engines.

Support specialists interviewed in Study 3 indicated that researchers use web searches in particular for finding data from studies that they already know (known-item searches). They also indicated that users expect data repositories to be as easy to use as general web search engines, which they often are not.

In contrast, support specialists in Study 6 did not recognize or prioritize the role of web search in searching for data. None of the over 200 use cases in this study refer to web search and therefore had no priority assigned to them. While absence of proof is not proof of absence, web search does not seem to be the first thing that came to the minds of the many support specialists who compiled the list of use cases in the study. This is different from what we found for our next other two topics: Literature and social networks, which are both well-represented in Study 6.

4.3.3. Synthesis

As we could see in Studies 1, 2, 4, and 5, researchers use general web search engines extensively to find both data as well as data resources. One use case which we have strong support for is known item searches. Study 3 shows that support specialists are aware of this practice. They also concede that the retrieval tools at data repositories do not meet users’ expectations regarding searchability. Study 4 suggests that support specialists may rely less on web search. Study 6 likewise gives an inconclusive picture, in that web search is conspicuously missing. This leaves us with the curious conclusion that web search seems to be used often, but at the same time it may go largely unnoticed or be taken for granted by support specialists.

4.4 Data discovery source: Literature and data citation

4.4.1. Researcher perspective

Study 2, Study 4 and Study 5 document the importance of the academic literature in researchers’ practices of discovering data. Interviewees in Study 2 reported following citations to data, plucking data, including device specifications, from figures, tables, graphs and the supplementary information attached to articles. Literature search was highly prevalent in all observations of data search made in Study 5, with one person stating that literature search “enables data search”.

Seventy-five percent of researcher respondents to the survey in Study 4 reported *often* using the literature to find data, with 19% of them stating that they occasionally did so. Researchers make use of the literature in many ways, some of which are different from how support specialists do so (Figure
2). Both support specialists and researchers use the academic literature to find data by following citations. More researchers turn to the literature with the explicit goal of finding data from the literature than do support staff.

**How do you use literature to find secondary data?**

![Figure 2](image)

*Fig. 2* How respondents use the literature to discover data by role: researchers (n= 1531) and support specialists (n=42). Question only asked to respondents who report using literature as a data discovery source. Percents are percent of respondents. Multiple responses possible.

Study 5 reports that most of the participants used literature for their data search. Some used it to assist with the search, e.g., for relevance assessment, or to see how someone else did an analysis. Other participants conducted their whole dataset search through literature search, seeing literature search as being a prerequisite for searching for datasets. Similarly in Study 1, one social scientist’s normal practice involves discovering reusable variables in datasets by first searching for relevant literature.

### 4.4.2. Support specialist perspective

Study 4 suggests that support specialists make more use of a diversity of sources to locate data than do researchers, with respondents selecting that they *often* use the literature (42%), search engines, domain repositories and governmental sources at roughly the same percentage.

Support specialists tend to view data discovery and literature discovery as separate and distinct practices, whereas more researchers turn to the literature with the explicit goal of finding data from the literature than do support staff (Figure 2). In Study 4, 49% of support specialists (n=23) stated that their process for finding data and for finding literature is always different, while less than 20% of researchers responding to this question say that their processes are always different (Figure 3).

**Do you discover data differently than literature?**

![Figure 3](image)

*Fig. 3* Purposes for using secondary data by role: researchers (n=1630) and support specialists (n=47). Percents are percent of respondents. Multiple responses possible.

In Study 3, the interviewed support specialists reported that research papers are one of the main sources for learning about existing data for data archive users. One support specialist also reported observing a frequent use of DOI references to enter records in the data catalog, which suggests that users are citation chaining from papers. In particular for students or young researchers, textbooks were seen to be another important source for discovering data.
Of all the data discovery use cases collected by support specialists in Study 6, data citation and more generally, the connection between literature and datasets, had the greatest number of use cases. However, when it came to prioritization, some ranked it as high priority, while others explicitly ranked it as low priority (they only could pick two of each).

4.4.3 Synthesis

All six studies show support in various ways that academic literature is important to data discovery, a result that is consistent between different disciplines. The degree of importance, however, seems to be less pronounced among data support specialists (Study 4 and 6), although they do seem to recognize that this is common practice among researchers (see interviews in Study 3).

4.5 Data seeking via social connections

4.5.1. Researcher perspective

In Study 4, researchers report often (31%) or occasionally (54%) using personal networks to discover data. The ways in which researchers and support specialists make use of social connections to discover data has similarities (i.e., both groups report relying on personal conversations) and differences, namely in that researchers report attending conferences as a way to discover data more and support specialists report making use of disciplinary forums (Figure 4).11

**Fig. 4** How respondents mobilize social connections in data discovery. Percents are percentages of responses for each option. Researchers: discovering (n=3192), accessing (n=3497), sensemaking (n=2949) and support specialists: discovering (n=119), accessing (n=92), sensemaking (n=82).

There are also similarities and differences in how the two groups make use of social connections to access and make sense of data. Researchers selected forming new collaborations with data creators and attending conferences slightly more often than did support specialists. Support specialists selected contacting data creators as a way to access and make sense of data more frequently.

Study 1 also revealed that most of the interviewed social science researchers mainly used personal networks to discover and access data. They reported two types of reused data: publicly available data downloadable from the web and non-publicly available data obtained from research collaborators. For the former, most interviewees reported that they could not recall spending particular effort in data discovery. Rather, their knowledge of the existence of the data was accumulated incidentally from past research experiences (e.g., training, hearsay from colleagues). For the latter type, interviewees described obtaining access to the data through collaborating with data collectors (e.g., PhD advisors, colleagues, or non-research organizations that pay researchers for data analysis).

11 Responses also indicate envisioned behavior of researchers.
In Study 5, participants explained in the post-observation interview that they would get information about datasets from colleagues. During the observation, when given the option to communicate with colleagues, one participant made a phone call, and others wrote emails or said they would write an email. Supervisors played a special role for younger scientists, in that they recommended both the datasets as well as guided on data analysis. It seems that secondary analysis is often teamwork with specialized roles among team members, such as defining research questions and analyzing data.

4.5.2. Support specialist perspective

Personal networks are also important for support specialists, as indicated in Figure 4 above. Study 3 revealed that community involvement plays an important role in finding data. According to the requests that data support specialists receive, researchers request data that they have been pointed to, for example by supervisors. Others exchange datasets with peers, even data with access restriction (data dealing). The interviews also revealed that support specialists themselves make use of a larger data-related community when catering to users’ needs. These data communities include, for instance, principal investigators who have collected the data and work closely with data professionals (data managers, data curators, data archivists, data librarians and others) to prepare the data for archiving, publication or reuse. The interviews also suggest that data reusers themselves are part of these communities, for example, when they are invited to detect and report errors in datasets or make other suggestions for improvements or enhancements.

In Study 6, the use case cluster regarding social connections was one of the smallest. This cluster was not seen as relevant by the support specialists.

4.5.3 Synthesis

With exception of Study 6, all studies show that social connections are viewed as important by both researchers and support specialists. The ways of connecting may be different as Study 4 suggests.

For researchers, social connections are an important means to find data (Study 1, 2, 4 and 5) and are of value in reusing and making sense of data (Study 5). Study 4 in particular revealed that more than 80 percent of researchers at least occasionally use personal networks to find data. Study 3 shows support specialists confirm this finding and also describes the composition of these social networks, which include data producers, primary researchers, and support specialists. For support specialists, these connections are important to prepare data as well as to assist researchers in finding and reusing data.

Study 1 suggests that the relevance of social connections may depend on the accessibility of the data; finding open datasets apparently doesn’t require social interactions as much as finding access-restricted datasets. Similarly, Study 3 suggests that in personal networks, informal sharing of restricted datasets is common. It also seems that the use of social connections is particularly important when help is needed to analyze or make sense of data (studies 4 and 5).

5. Discussion

In this paper, we conducted a meta-synthesis of six of our own studies on data discovery, with the aim of bringing together the perspectives of both researchers and research data support specialists. In doing so, we identified both points of alignment and difference in the two perspectives in all five of the themes which we identified in our analysis (Table 3). We discuss these points of alignment and difference here, before relating them to possibilities of future work for support specialists in the Conclusion section, according to the typology of support specialists identified in the Literature Analysis.
While we find high alignment on the challenges and data needs between researchers and support specialists, there are three areas of lower alignment which emerge in our analysis: Web search, use of literature and social connections.

Table 3. Comparison between the findings of researchers and support specialists (cf. section 4) and their alignment. Numbers in parentheses refer to study numbers.

| Theme                  | Findings: Researchers' Perspective | Findings: Specialists' Perspective | Alignment                      |
|------------------------|-----------------------------------|-----------------------------------|---------------------------------|
| Challenges             | data discovery is challenging (2, 4, 5) | researchers need assistance (3) see a challenge to locate data (3) | High                            |
|                        | lack of search tools (2, 4, 5)      | see a challenge in data access (3) |                                 |
|                        | lack of data (seeking) literacy (4, 5) | data literacy of researchers too low (3) |                                 |
|                        | data accessibility (4, 5)           | erroneous data (3)                 |                                 |
|                        | modifying strategy to fit the data (2, 5) | priorities: metadata, data citation, ease of use (6) |                                 |
|                        | lack of links (5)                   | non-priorities: documentation, links to persons, machine discoverability (6) |                                 |
| Data needs             | need multiple data types (2,4)      | specialists inherit researchers needs (2,3,4) | High, with slightly different emphasis |
|                        | multiple different purposes for data (1) | less likely to conduct their own research (4) |                                 |
|                        | sometimes very specific and complex needs (2,4) | more likely to use data for teaching (4) |                                 |
|                        |                                       | specialists are perhaps more likely to do data integration/curation (4) |                                 |
| Web search             | web search is a highly used data discovery tool (1,2,4,5) | do not use web search as much (4) | Low, this seems to be underestimated by specialists |
|                        | used for a variety of purposes (5)  | do not recognize web search as part of data discovery (6) |                                 |
| Use of literature      | very important part of data discovery (1,2,4,5) | use literature differently than researchers, but also heavily (4) aware that researchers discover data by citation chaining from papers (3) literature is a priority for some specialists, but not for others (6) see literature search as being more separated from data discovery (4) | Medium, with slightly different emphasis |
|                        | “Enables data discovery” (5)       |                                 |                                 |
| Use of social connections | important part of data discovery (1,2,4,5) | rely less on social networks for data discovery than researcher (4) expect to contact data creators directly (4, 5) is not seen as a priority (6) | Medium, some differences |
|                        | used for finding and making sense of data (4,5) |                                 |                                 |

5.1 High alignment between researchers and support specialists

Data discovery is perceived as challenging by researchers and support specialists alike. Challenges in data seeking as reported by researchers are mostly about lack of data search skill, lack of data literacy, and lack of access to data. Support staff also realize that these are important challenges for researchers, and work to support these challenges through training and consultancy efforts (Study 3, [39]).

All studies, with the exception of Study 6 provide information about researchers’ data needs, characterizing data needs as evolving and being both general and specific. The interviewed support specialists from Study 3 proved to be aware of both very general as well as very specific needs and purposes for data use. Support specialists sometimes need data for their own research, albeit to a lesser extent than researchers. Studies 1-4 also highlight that both support specialists and researchers need data to use in their own teaching.
5.2 Medium alignment between researchers and support specialists

Researchers mention technical challenges, such as a lack of suitable search tools and links between research data and other relevant sources of information, in particular the academic literature. Support staff in our studies don’t see priorities in this area and are more cognizant of the challenges of poor metadata and challenges related to improving/standardizing data citation than are researchers. This reflects the fact that support specialist work includes tasks related to metadata curation and improving metadata and data standardization [4, 49, 63].

Though in our findings, data being distributed is reported by researchers but neglected by support specialists, it is a challenge recognized by support specialists and informational professionals in other studies and they develop metadata standards and build tools for federated dataset search to support data discoverability (e.g., [62]).

All of our studies show that researchers make use of personal networks, social exchanges in data discovery, to find/locate data, perhaps especially when data access is challenging, but also to evaluate and understand data. This reliance is in line with many other studies on data discovery practices and data reuse [13, 28, 72, 73]. Despite this awareness, social connections are not defined as a formal way of data discovery that deserves attention from the support specialist community (Study 6). This contrasts the fact that support work related to data discovery is itself a collaborative process that involves many personal exchanges (Study 2, [42]).

Another point of lesser alignment of perspectives is the use of academic literature. Researchers use literature as a source for finding data. Our findings indicate that support specialists are aware of the use and importance of literature in discovering data by researchers, but use literature less themselves as an avenue for finding data and perhaps see data discovery as a practice that is more distinct from literature discovery. Support specialists in Study 6 also did not prioritize data citation as a pressing topic. Nevertheless, there are increasing efforts in linking research data to papers to enhance data discoverability such as Scholix [52] and Research Graph [54, 74].

Many support specialists’ own data needs are service-oriented and are tied to the support work that they do (i.e., teaching, summarizing, or visualizing data). Study 4 suggests that support specialists are perhaps more invested in pooling or integrating datasets. This finding may reflect the curatorial nature of some support specialists’ work, i.e., integrating and maintaining longitudinal survey data, as found in prior studies (e.g., [4]).

Differences regarding the importance of technical work as well as the use of literature may be explained with the different roles both groups find themselves in. Researchers tend to be highly specialized within their domain and more likely to do complex analysis involving several datasets or complex methodology. Literature is important to them to gauge the impact of their work within their community. Research support specialists, on the other hand, tend to be more generalist, often responsible for whole fields, like social science. They often search data as part of a delegated task, without the need to understand the bigger picture of that particular research.

5.3 Low alignment between researchers and support specialists

General web search engines such as Google or Bing are an important source for researchers to discover data, as they search the web to find data repositories, conduct known-item searches and to find data directly (Study 1, 2, 4 and 5, [34]). While there are indications that support specialists are aware of this behavior (Study 3, [75]), there also appear to be differences in perspectives. Study 4 points to differences both in how support specialists use web search themselves to locate data, suggesting that support specialists do not rely on general web search engines as much as researchers do, but rather spread their efforts across sources. Study 6 suggests that support specialists may not see web search as being important enough to consider when thinking about supporting researchers’ data discovery practices, as it was not mentioned in the use case prioritization.
6. Conclusion

6.1 Limitations

Putting all these different findings together is challenging given their different methods (survey, interview, user study, use case collection) and the different demographics of the subjects. The studies in their variety of methodologies do not support representativity, but rather provide in-depth data from a variety of perspectives. As such, we do not try to make representative claims. Another limitation of this analysis is that there are subtly different understandings of the data discovery process in play. It is not always clear whether this process includes data reuse practices or data search behaviors specifically.

As mentioned in the literature review, there are also difficulties in defining what a “data support specialist” is. This term could refer to a librarian, a data archivist with a domain background, a metadata expert, or an IT specialist. We address this difficulty by creating a typology of work foci in the literature review, which includes people-oriented, metadata-oriented, and infrastructure-oriented work, a typology which we use to make the below recommendations regarding future work.

6.2 Recommendations for the different areas of support work

The alignments and differences offer a window into thinking about how the work of support specialists could better support data discovery practices of researchers. We conclude by thinking about how different types of support work identified in Section 2.2 could address these points of difference – both by identifying existing best practices and by making suggestions for improvement.

**People-oriented type of support work**, especially in university libraries, may consider fully incorporating data literacy skill training and consultation (e.g., skill of searching data and evaluating results) into the existing research support toolkits. We understand that data literacy is a complex concept and the training may range from “where to search for what data” to “advanced skills on data analysis”. We see an opportunity for support specialists to provide different levels of data literacy training for researchers at varying stages of their research process. This may require support staff working with different disciplines to develop targeted training content.

On a similar note, we suggest that libraries offer training for students to find research data in the same way as they are already doing for students to find literature, which would be an extension of the traditional library-related reference support work to the new field of open data and RDM. Many university libraries are currently providing data literacy training as an optional provision on top of the more established information literacy skill training such as “how to use citation databases” or “how to search literature”.

**Metadata-oriented support work** may need a general improvement with regard to the needs of researchers. While metadata quality is a problem, as seen in Study 3 where data seekers were not satisfied with documentation which has errors, another problem to address is to decide the right level of granularity of research contextual information included in data documentation read by researchers. The usability of data documentation itself would impact the reusability of data.

Another promising strain of metadata-oriented support work is to continue expanding the scope and scale of the existing initiatives such as Scholix [52] of building “links” between research data to other research materials (primarily literature) and research entities (e.g., researchers, research institutions). Our studies reveal that researchers need such links for discovering research data and evaluating data reusability (study 1, 2, 4, 5), whereas support specialists don’t assign link-building tasks as high priority (study 6). This is a conspicuous discrepancy that needs to be addressed.

**Infrastructure-oriented support work** may fork into the development of search tools customized for different data reuse tasks, in line with what Koesten at al. [28] suggest. To do this, an in-depth
understanding of data-centric reuse tasks would be needed. We call for more user studies into researchers’ data reuse behaviors, needs and requirements.

More user studies may also be needed to improve the findability of research data via general web search engines and to increase usability of data repositories, which may be a topic for both metadata-oriented and infrastructure-oriented support work. It’s not uncommon that researchers start with web search engines for searching, as they are simple and easy to use. There are existing efforts exploring using Linked Data technology and schema.org vocabularies to expose research data to the Internet and other web services [55]. The more relevant issue is, perhaps, to retain researchers after they land on a data repository webpage.

Extant studies often use log analysis to investigate users’ interaction with data repositories. We argue that more user studies with an eclectic mix of research methods (e.g., think aloud tasks or observational studies) are needed to inform the design of data repositories to improve both exploration of data and retention of data seekers.

Perhaps even more important is conducting studies that brings together researchers and supports specialists in co-design activities. We gather and compare practices and perspectives of both sides by re-analyzing data from our past work. Further research is needed to take this into account at the beginning of study design.

References

1. Borgman CL, Scharnhorst A, Golshan MS (2019) Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. Journal of the Association for Information Science and Technology 70:888–904. https://doi.org/10.1002/asi.24172

2. Cope J, Baker J (2017) Library Carpentry: Software Skills Training for Library Professionals. International Journal of Digital Curation 12:266–273. https://doi.org/10.2218/ijdc.v12i2.576

3. Cox AM, Kennan MA, Lyon L, Pinfield S, Shaffi L (2019) Maturing research data services and the transformation of academic libraries. Journal of Documentation 75:1432–1462. https://doi.org/10.1108/JD-12-2018-0211

4. Lafia S, Thomer A, Bleckley D, Akmon D, Hemphill L (2021) Leveraging Machine Learning to Detect Data Curation Activities. In: 2021 IEEE 17th International Conference on eScience (eScience). IEEE, Innsbruck, Austria, pp 149–158

5. Faniel I, Kansa E, Whitcher Kansa S, Barrera-Gomez J, Yakel E (2013) The challenges of digging data: a study of context in archaeological data reuse. In: Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries. Association for Computing Machinery, New York, NY, USA, pp 295–304

6. Friedrich T (2020) Looking for data. Dissertation, Humboldt-Universität zu Berlin

7. Joo S, Kim S, Kim Y (2017) An exploratory study of health scientists’ data reuse behaviors: Examining attitudinal, social, and resource factors. Aslib J Inf Manag 69:389–407. https://doi.org/10.1108/AJIM-12-2016-0201

8. Shen Y (2017) Burgeoning Data Repository Systems, Characteristics, and Development Strategies: Insights of Natural Resources and Environmental Scientists. Data and Information Management 1:115–123. https://doi.org/10.1515/dim-2017-0009
9. Murillo AP (2014) Examining data sharing and data reuse in the dataone environment. Proceedings of the American Society for Information Science and Technology 51:1–5. https://doi.org/10.1002/meet.2014.14505101155

10. Xiao F, Wang Z, He D (2020) Understanding users’ accessing behaviors to local Open Government Data via transaction log analysis. Proceedings of the Association for Information Science and Technology 57:e278. https://doi.org/10.1002/pra2.278

11. Faniel IM, Frank RD, Yakel E (2019) Context from the data reuser’s point of view. Journal of Documentation 75:1274–1297. https://doi.org/10.1108/JD-08-2018-0133

12. Kim Y, Yoon A (2017) Scientists’ data reuse behaviors: A multilevel analysis. Journal of the Association for Information Science and Technology 68:. https://doi.org/10.1002/asi.23892

13. Yoon A, Lee YY (2019) Factors of trust in data reuse. Online Inf Rev 43:1245–1262. https://doi.org/10.1108/oir-01-2019-0014

14. Gowen E, Meier JJ (2020) Research Data Management Services and Strategic Planning in Libraries Today: A Longitudinal Study. Journal of Librarianship and Scholarly Communication 8:eP2336. https://doi.org/10.7710/2162-3309.2336

15. Reichmann S, Klebel T, Hasani-Mavriqi I, Ross-Hellauer T (2021) Between administration and research: Understanding data management practices in an institutional context. Journal of the Association for Information Science and Technology 72:1415–1431. https://doi.org/10.1002/asi.24492

16. Cruz MJ, Böhmer J, Gramsbergen E, Teperek M, Dunning A, Smaele M de (2018) From Passive to Active, From Generic to Focussed: How Can an Institutional Data Archive Remain Relevant in a Rapidly Evolving Landscape? International Journal of Digital Curation 13:385–391. https://doi.org/10.2218/ijdc.v13i1.613

17. Khan N, Thelwall M, Kousha K (2021) Are data repositories fettered? A survey of current practices, challenges and future technologies. Online Information Review ahead-of-print: https://doi.org/10.1108/OIR-04-2021-0204

18. Gregory KM, Cousijn H, Groth P, Scharnhorst A, Wyatt S (2019) Understanding data search as a socio-technical practice. Journal of Information Science 46:459–475. https://doi.org/10.1177/0165551519837182

19. Sun G (2019) Knowledge representation of social science quantitative research data for data curation and reuse. Dissertation, Nanyang Technological University

20. Gregory K (2020) A dataset describing data discovery and reuse practices in research. Sci Data 7:232. https://doi.org/10.1038/s41597-020-0569-5

21. Gregory K, Groth P, Scharnhorst A, Wyatt S (2020) Lost or Found? Discovering Data Needed for Research. Harvard Data Science Review 2:. https://doi.org/10.1162/99608f92.e38165eb

22. Krämer T, Papenmeier A, Carevic Z, Kern D, Mathiak B (2021) Data-Seeking Behaviour in the Social Sciences. Int J Digit Libr 22:175–195. https://doi.org/10.1007/s00799-021-00303-0

23. de Waard A, Khalsa SJ, Psomopoulos F, Wu M (2017) RDA IG Data Discovery Paradigms IG: Use Cases data

24. Mathiak B (2022) Paper is accepted with revisions and a citation will be added to the final paper
25. Curty RG (2016) Factors Influencing Research Data Reuse in the Social Sciences: An Exploratory Study. International Journal of Digital Curation 11:96–117. https://doi.org/10.2218/ijdc.v11i1.401

26. Joo YK, Kim Y (2017) Engineering researchers’ data reuse behaviours: a structural equation modelling approach. Electron Libr 35:1141–1161. https://doi.org/10.1108/EL-08-2016-0163

27. Borst T, Limani F (2020) Patterns for searching data on the web across different research communities. LIBER Quarterly: The Journal of the Association of European Research Libraries 30:1–21. https://doi.org/10.18352/lq.10317

28. Koesten LM, Kacprzak E, Tennison JFA, Simperl E (2017) 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. Association for Computing Machinery, New York, NY, USA, pp 1277–1289

29. Wu M, Psomopoulos F, Khalsa SJ, Waard A de (2019) Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories. Data Science Journal 18:3. https://doi.org/10.5334/dsj-2019-003

30. Bishop BW, Hank C, Webster J, Howard R (2019) Scientists’ data discovery and reuse behavior: (Meta)data fitness for use and the FAIR data principles. Proceedings of the Association for Information Science and Technology 56:21–31. https://doi.org/10.1002/prab.24

31. Kacprzak E, Koesten L, Tennison J, Simperl E (2018) Characterising Dataset Search Queries. In: Companion Proceedings of the The Web Conference 2018. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp 1485–1488

32. Brickley D, Burgess M, Noy N (2019) Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In: The World Wide Web Conference. Association for Computing Machinery, New York, NY, USA, pp 1365–1375

33. Hemphill L, Pienta A, Lafia S, Akmon D, Bleckley DA (2022) How do properties of data, their curation, and their funding relate to reuse? Journal of the Association for Information Science and Technology n/a: https://doi.org/10.1002/asi.24646

34. Sharifpour R, Wu M, Zhang X (2022) Large-scale analysis of query logs to profile users for dataset search. Journal of Documentation ahead-of-print: https://doi.org/10.1108/JD-12-2021-0245

35. Tammaro AM, Matusiak KK, Sposito FA, Casarosa V (2019) Data Curator’s Roles and Responsibilities: An International Perspective. Libri 69:89–104. https://doi.org/10.1515/libri-2018-0090

36. Wilson L, Colborne A, Smit M (2017) Preparing data managers to support open ocean science: Required competencies, assessed gaps, and the role of experiential learning. IEEE Computer Society, pp 3984–3993

37. Darch PT, Sands AE, Borgman CL, Golshan MS (2020) Library cultures of data curation: Adventures in astronomy. Journal of the Association for Information Science and Technology 71:1470–1483. https://doi.org/10.1002/asi.24345

38. Si L, Zeng Y, Guo S, Zhuang X (2019) Investigation and analysis of research support services in academic libraries. The Electronic Library 37:281–301. https://doi.org/10.1108/EL-06-2018-0125
39. Cox AM, Kennan MA, Lyon L, Pinfield S (2017) Developments in research data management in academic libraries: Towards an understanding of research data service maturity. Journal of the Association for Information Science and Technology 68:2182–2200. https://doi.org/10.1002/asi.23781

40. Cruz M, Dintzner N, Dunning A, Kuil A van der, Plomp E, Teperek M, Velden YT der, Versteeg A (2019) Policy Needs to Go Hand in Hand with Practice: The Learning and Listening Approach to Data Management. Data Science Journal 18:45. https://doi.org/10.5334/dsj-2019-045

41. Savage JL, Cadwallader L (2019) Establishing, Developing, and Sustaining a Community of Data Champions. Data Science Journal 18:23. https:// doi.org/10.5334/dsj-2019-023

42. Joo S, Schmidt GM (2019) Establishing, Developing, and Sustaining a Community of Data Champions. Data Science Journal 18:23. https://doi.org/10.5334/dsj-2019-023

43. Lassi M, Johnsson M, Golub K (2016) Research data services: An exploration of requirements at two Swedish universities. IFLA Journal 42:266–277. https://doi.org/10.1177/0340035216671963

44. Borycz J (2021) Implementing Data Management Workflows in Research Groups Through Integrated Library Consultancy. Data Science Journal 20:9. https://doi.org/10.5334/dsj-2021-009

45. Das A, Banerjee S (2021) Optimising research support services through libraries: A review of practices. Library Philosophy and Practice (e-journal) 5515

46. Zakaria MS (2021) Data visualization as a research support service in academic libraries: An investigation of world-class universities. The Journal of Academic Librarianship 47:102397. https://doi.org/10.1016/j.acalib.2021.102397

47. Huang Y, Cox AM, Sbaffi L (2021) Research data management policy and practice in Chinese university libraries. Journal of the Association for Information Science and Technology 72:493–506. https://doi.org/10.1002/asi.24413

48. Marlina E, Purwandari B (2019) Strategy for Research Data Management Services in Indonesia. Procedia Computer Science 161:788–796. https://doi.org/10.1016/j.procs.2019.11.184

49. Vardigan M, Heus P, Thomas W (2008) Data Documentation Initiative: Toward a Standard for the Social Sciences. International Journal of Digital Curation 3:107–113. https://doi.org/10.2218/ijdc.v3i1.45

50. Löbe M, Ulrich H, Beger C, Bender T, Bauer C, Sax U, Ingenierf J, Winter A (2022) Improving Findability of Digital Assets in Research Data Repositories Using the W3C DCAT Vocabulary. MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation 61–65. https://doi.org/10.3233/SHTI220032

51. Tomoyose K, Arakaki ACS (2022) Vocabularies for Publishing Research Data. Cataloging & Classification Quarterly 60:69–85. https://doi.org/10.1080/01639374.2021.2011812

52. Burton A, Aryani A, Koers H, Manghi P, La Bruzzo S, Stocker M, Diepenbroek M, Schindler U, Fenner M (2017) The Scholix Framework for Interoperability in Data-Literature Information Exchange. D-Lib Magazine 23:. https://doi.org/10.1045/january2017-burton

53. Khan N, Pink CJ, Thelwall M (2020) Identifying Data Sharing and Reuse with Scholix: Potentials and Limitations. Patterns 1:100007. https://doi.org/10.1016/j.patter.2020.100007
54. Aryani A, Poblet M, Unsworth K, Wang J, Evans B, Devaraju A, Hausstein B, Klas C-P, Zapilko B, Kaplun S (2018) A Research Graph dataset for connecting research data repositories using RD-Switchboard. Sci Data 5:180099. https://doi.org/10.1038/sdata.2018.99

55. Wang J, Aryani A, Wyborn L, Evans B (2017) Providing Research Graph Data in JSON-LD Using Schema.org. In: Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp 1213–1218

56. Guha RV, Brickley D, Macbeth S (2016) Schema.org: evolution of structured data on the web. Commun ACM 59:44–51. https://doi.org/10.1145/2844544

57. Fons T, Penka J, Wallis R (2012) OCLC’s Linked Data Initiative: Using Schema.org to Make Library Data Relevant on the Web. ISQ 24:29–33. https://doi.org/10.3789/isqv24n2-3.2012.05

58. Bugaje M, Chowdhury G (2018) Data Retrieval = Text Retrieval? In: Chowdhury G, McLeod J, Gillet V, Willett P (eds) Transforming Digital Worlds. Springer International Publishing, Cham, pp 253–262

59. Bugaje M, Chowdhury G (2018) Identifying Design Requirements of a User-Centered Research Data Management System. In: Dobrevska M, Hinze A, Žumer M (eds) Maturity and Innovation in Digital Libraries. Springer International Publishing, Cham, pp 335–347

60. Liu S, Su Y, Yin H, Zhang D, He J, Huang H, Jiang X, Wang X, Gong H, Li Z, Xiu H, Wan J, Zhang X (2021) An infrastructure with user-centered presentation data model for integrated management of materials data and services. npj Comput Mater 7:1–8. https://doi.org/10.1038/s41524-021-00557-x

61. Mannheimer S, Clark J, Hagerman K, Schultz J, Espeland J (2021) Dataset Search: A lightweight, community-built tool to support research data discovery. Journal of eScience Librarianship 10:e1189. https://doi.org/10.7191/jeslib.2021.1189

62. Garnett A, Leahey A, Savard D, Towell B, Wilson L (2017) Open Metadata for Research Data Discovery in Canada. Journal of Library Metadata 17:201–217. https://doi.org/10.1080/19386389.2018.1443698

63. Arora S, Chakravarty R (2021) Making research data discoverable: an outreach activity of Datacite. Library Philosophy and Practice (e-journal) 5199

64. Read K, Athens J, Lamb I, Nicholson J, Chin S, Xu J, Rambo N, Surkis A (2015) Promoting Data Reuse and Collaboration at an Academic Medical Center. International Journal of Digital Curation 10:260–267. https://doi.org/10.2218/ijdc.v10i1.366

65. Sheridan H, Dellureficio A, Ratajeski M, Mannheimer S, Wheeler T (2021) Data Curation through Catalogs: A Repository-Independent Model for Data Discovery. Journal of eScience Librarianship 10:.. https://doi.org/10.7191/jeslib.2021.1203

66. Smit M (2019) Code Convention Adherence in Research Data Infrastructure Software: An Exploratory Study. In: 2019 IEEE International Conference on Big Data (Big Data). pp 4691–4700

67. Hedeland H (2020) Providing Digital Infrastructure for Audio-Visual Linguistic Research Data with Diverse Usage Scenarios: Lessons Learnt. Publications 8:33. https://doi.org/10.3390/publications8020033
68. Corbin J, Strauss A (2014) Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. SAGE Publications

69. R Core Team R: a language and environment for statistical computing. https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing. Accessed 13 Jul 2022

70. Kern D, Mathiak B (2015) Are There Any Differences in Data Set Retrieval Compared to Well-Known Literature Retrieval? In: Kapidakis S, Masurek C, Werla M (eds) Research and Advanced Technology for Digital Libraries. Springer International Publishing, Cham, pp 197–208

71. National Science Board (2005) Long-lived digital data collections: Enabling research and education in the 21st century (No. NSB0540). National Science Foundation.

72. Kim Y (2017) Fostering scientists’ data sharing behaviors via data repositories, journal supplements, and personal communication methods. Inf Process Manag 53:871–885. https://doi.org/10.1016/j.ipm.2017.03.003

73. Yoon A (2017) Role of communication in data reuse. Proceedings of the Association for Information Science and Technology 54:463–471. https://doi.org/10.1002/pra2.2017.14505401050

74. Färber M, Lamprecht D (2021) The data set knowledge graph: Creating a linked open data source for data sets. Quantitative Science Studies 2:1324–1355. https://doi.org/10.1162/qss_a_00161

75. Pienta AM, Akmon D, Noble J, Hoelter L, Jekielek S (2017) A Data-Driven Approach to Appraisal and Selection at a Domain Data Repository. International Journal of Digital Curation 12:362–375. https://doi.org/10.2218/ijdc.v12i2.500