UvA-DARE (Digital Academic Repository)

Lost or Found? Discovering Data Needed for Research

Gregory, K.; Groth, P.; Scharnhorst, A.; Wyatt, S.

Link to publication

Lost or Found? Discovering Data Needed for Research

Kathleen Gregory[1], Paul Groth[2], Andrea Scharnhorst[1], Sally Wyatt[3]

[1]**Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Sciences,**
[2]**Informatics Institute, University of Amsterdam,**
[3]**Faculty of Arts and Social Sciences, Maastricht University**

ABSTRACT

Finding data is a necessary precursor to being able to reuse data, although relatively little large-scale empirical evidence exists about how researchers discover, make sense of, and (re)use data for research. This study presents evidence from the largest known survey investigating how researchers discover and use data that they do not create themselves. We examine the data needs and discovery strategies of respondents, propose a typology for data reuse, and probe the role of social interactions and literature search in data discovery. We consider how data communities can be conceptualized according to data uses and propose practical applications of our findings for designers of data discovery systems and repositories. Specifically, we consider how to design for a diversity of practices, how *communities of use* can serve as an entry point for design and the role of metadata in supporting both sensemaking and social interactions.

Keywords: data discovery, data search, data reuse, research practices, open data policy, research communities

Media summary

Reusing data created by others holds great promise in research. Little is known, however, about how researchers actually discover and use data that they do not create themselves. Which data do they seek? For which purposes? Where do they look? How do they make sense of data once they find them? This article presents evidence from the largest known global survey examining these questions. 1677 respondents in 105 countries, representing a variety of disciplinary domains, professional roles and stages in their academic careers, completed the survey. The results represent respondents' data needs, the sources and strategies they use to locate data, and the criteria they employ when evaluating data for reuse. We discuss the practical applications these findings have for designers of data repositories and search systems.

1. Introduction

Stakeholders from funders to researchers are increasingly concerned with the sharing and reuse of research data (e.g., Digital Curation Centre, n.d.; Tenopir et al., 2015). Policymakers draft guidelines, systems designers create repositories and tools, and librarians develop training materials to encourage opening and sharing data, often without empirical evidence about community-specific practices (Noorman et al., 2018). It is assumed that data can and will be reused if they are shared (Borgman, 2015a). Another assumption predicating reuse is that data will actually be discovered by researchers, although relatively little empirical work exists to support this assumption.

In this article, we present the results of the largest known survey examining how researchers discover and (re)use research data that they do not create themselves, so-called *secondary data* (Allen, 2017). We consider commonalities in practices but also examine differences, looking at how data needs and search practices vary not only by disciplinary domain but also by types of data uses. Past work has explored data search practices via in-depth interviews (Borgman, Scharnhorst, & Golshan, 2019; Gregory et al., 2019a; Koesten et al., 2017). This study employs a broader approach, using a globally distributed multidisciplinary survey, with nearly 1,700 respondents, to explore these practices at a larger scale.

Our study produced a rich data set including both qualitative and quantitative data. Here, we present the discovery phase of exploring the quantitative data, relying on descriptive statistics and tests for pairwise correlations, and we draw deeply on the qualitative responses. We use these analyses to paint a detailed picture of data discovery and propose a typology for data (re)use that we use to explore the data needs and practices of participants. We also probe the role of social interactions in searching for data and explore how data search is related to other practices, such as searching for academic literature. We consider how communities of data seekers can be conceptualized and discuss our findings in light of recent efforts to increase the discoverability of research data in a theoretical discussion. We conclude with practical considerations for the application of our findings by data discovery systems designers and repository managers and suggest further directions for research.

2. Background

Although information seeking is an extensive research field, work investigating *data-seeking practices* is nascent. Practices of data seeking, which we refer to here also as *data search* or *data discovery* practices, are commonly examined through user studies of particular data platforms and repositories (e.g., Borgman, Scharnhorst, & Golshan, 2019; Murillo, 2015; Wu et al., 2019). Zimmerman (2003, 2007, 2008) investigates data search practices directly, looking at the needs, discovery strategies, and criteria for evaluating data for reuse for a small group of environmental scientists. Recent work characterizes data search and evaluation practices across disciplinary domains (Gregory et al., 2019a, 2019b) and by data professionals within and outside of academia (Koesten et al., 2017), relying primarily on in-depth interviews with data seekers or log analyses (Kacprzak et al., 2017).

Surveys investigating data practices tend to use quantitative methods and focus on data-sharing behaviors across disciplines (e.g., Kim & Zhang, 2015; Tenopir et al., 2011; Tenopir et al., 2015), within specific domains (Schmidt et al., 2016; Tenopir et al., 2018), or in different geographic locations (Ünal et al., 2019). Publishers and data repositories conduct annual surveys tracing data-sharing and management practices over time (e.g., Berghmans et al., 2017; Digital Science et al., 2018). Information about data search strategies, criteria important for reuse, and the role of social communications is

found within surveys designed to develop data metrics (Kratz & Strasser, 2015) and to determine factors affecting data reuse (Kim & Yoon, 2017; Yoon, 2017).

Interest in designing tools specifically for data search is growing (Chapman et al., 2020), evidenced by the development of search engines for research data (Noy et al., 2019; Scerri et al., 2017). Despite this trend, the limited amount of user interaction data restricts how these search tools are developed (Noy et al., 2019). There are also a growing number of policies regulating open data and data sharing (European Commission, 2019), which are seen as precursors to creating the ecosystem necessary for data discovery (Borgman, 2015b). These policies often do not accurately reflect the way that opening data and data sharing are enacted within communities (Noorman et al., 2018).

The sustainability and adoption of both search systems and data policies rely on understanding and building on extant practices (Schatzki et al., 2001). Our work aims to provide *evidence of practices of data seeking* and to inform the design of community-centric solutions and policies. To do this we take a broad approach, looking for commonalities that can be used for design, while also highlighting differences. We present a detailed examination of the behaviors of our respondents as they engage in discovering, evaluating, and reusing data; this work also provides a starting point for future analyses.

3. Methodology

3.1. Survey Design

We drew heavily on the findings of our earlier interviews with data seekers (Gregory et al., 2019a) and our analytical literature review (Gregory et al., 2019b) to design a survey addressing our principal research questions (see Table 1). Our research questions were informed by user-centered models of interactive information retrieval (e.g., Belkin, 1993, 1996; Ingwersen, 1992, 1996), particularly the synthesized model of an information journey proposed by Blandford and Attfield, 2010, and Adams and Blandford, 2005, which generally posits an actor/user with an (at times unrecognized) information need who engages in an iterative process of discovery, evaluation, and use.

Our survey employed a branching design consisting of a maximum of 28 individual items; nine of these items were constructed to allow for multiple responses. In addition to write-in responses for expanding on 'other' answers, the survey included three open-response questions. Respondents working as librarians or in research/data support also answered a slightly modified version of the survey.[1] Although we include some data from this group of 'research support professionals' in the results presented here (most notably in Figure 10), the primary focus of this article is on researchers.

Research questions	Items

RQ 1: Who are the people seeking data?	Q1, L3, **L4**, **L5 D1**, D2, D3, D4, D5, D6, D7
RQ 2: What data are needed for research and how are those data used?	Q2, **Q3**, **Q4,** Q5
RQ 3: How do people discover data needed for research?	*Q5a*, Q6, Q7, **Q7a,** Q7b, Q8, **Q9**, Q10, *Q10a*, Q11, **Q11a**
RQ 4: How do people evaluate and make sense of data needed for research?	**Q9,** Q12, Q13, Q14, Q15

Table 1. Survey questions addressed by each research question. "L" questions were asked only to librarians/research support professionals; "D" questions were assigned to demographic section. Multiple response questions indicated in bold; open response questions in italics.

Four multiple-response questions and their associated variables are of particular importance in our analysis; we include an overview of these variables to aid in navigating our results (Table 2).

Item	D1		Q3	Q4	Q9
Variable	**Disciplinary domains**		**Types of data needed**	**Types of data uses**	**Social connections**
Response Options	Agriculture	Engineering and Technology	Observational/ empirical	Basis for a new study	Discover - Conversations with personal networks (e.g. colleagues, peers
	Arts and Humanities	Environmental Sciences	Experimental	Calibrate instruments or models	Discover - Contacting the data creator
	Astronomy	Health professions	Derived / compiled	Benchmarking	Discover - Developing new academic collaborations with data creators

	Biochemistry, Genetics, and Molecular Biology	Immunology and Microbiology	Simulated	Verify my own data	Discover - Attending conferences
	Biological Sciences	Materials Science	Other	Model, algorithm or system inputs	Discover - Disciplinary mailing lists or discussion forums
	Business, Management and Accounting	Mathematics		Generate new ideas	Access - Conversations with personal networks (e.g. colleagues, peers
	Chemical Engineering	Medicine		Teaching / training	Access - Contacting the data creator
	Chemistry	Multidisciplinary		Prepare for a new project or proposal	Access - Developing new academic collaborations with data creators
	Computer Sciences / IT	Neuroscience		Experiment with new methods and techniques (e.g. to develop data science skills)	Access - Attending conferences
	Decision Sciences	Nursing		Identify trends or make predictions	Access - Disciplinary mailing lists or discussion forums

	Dentistry	Pharmacology, Toxicology and Pharmaceutics		Compare multiple datasets to find commonalities or differences	Sensemaking - Conversations with personal networks (e.g. colleagues, peers
	Earth and Planetary Sciences	Physics		Create summaries, visualizations, or analysis tools	Sensemaking - Contacting the data creator
	Economics, Econometrics and Finance	Psychology		Integrate with other data to create a new dataset	Sensemaking - Developing new academic collaborations with data creators
	Energy	Social Science		Other	Sensemaking - Attending conferences
		Veterinary			Sensemaking - Disciplinary mailing lists or discussion forums
		Information science			
		Other			

Table 2. Multiple response question numbers, variables, and response options of particular interest in our analysis.

The survey was scripted and administered with the [Confirmit](#) software. We piloted the survey instrument in two phases. We scheduled appointments with four researchers, recruited via convenience sampling, and observed them as they completed the online survey. During these observations, we encouraged participants to think out loud and ask questions. We used these

Harvard Data Science Review

Lost or Found? Discovering Data Needed for Research

comments to modify our questions before the next pilot phase. We then recruited an initial sample of 10,000 participants (using the recruitment methodology detailed herein). Once 100 participants had begun the survey, we measured the overall completion rate (41%), taking note of the points in the survey where people stopped completing questions. We used this information to further streamline the question order and to clarify the wording of some questions before recruiting our sample. A demographic analysis of the noncomplete responses was not possible, as demographic information was collected at the end of the survey questionnaire.

3.2. Sampling and Recruitment

Our population of interest consisted of individuals involved in research, in multiple disciplinary domains, who seek and use data they do not create themselves. This is a challenging population to target specifically, as information about who seeks and reuses data, particularly in a diversity of disciplines, is difficult to isolate (i.e., Park et al., 2018; Yoon, 2014). We therefore recruited our sample from a very broad population of researchers, hoping to capture a diverse sample of data seekers.

We sent recruitment emails to a random sample of an additional 150,000 authors who are indexed in Elsevier's Scopus database and who have published in the past three years. The recruitment sample was constructed to mirror the distribution of published authors by country within Scopus. Recruitment emails were sent in two batches, one of 100,000 and the other of 50,000, two weeks after the first batch. One reminder email was sent to encourage participation. A member of the Elsevier Research and Academic Relations team created the sample and sent the recruitment letter, as access to the authors' email addresses was restricted. Potential participants were informed that the purpose of the survey was to investigate data discovery practices. We therefore assume that participants who completed the survey are in fact data seekers.

We received 1,637 complete responses during a 4-week survey period in September–October 2018. We recruited an additional 40 participants by posting to discussion lists in the library and research data management communities, for a total sample of 1,677 complete responses, yielding a response rate of 1.1%. This response rate is calculated using the total number of recruitment emails distributed. It is likely that not all 150,000 individuals receiving recruitment emails matched our desired population of individuals who search for and reuse data.

Of the recruited participants, 2,306 individuals clicked on the link to the survey, but did not complete it. Forty-two percent of people who engaged with the survey completed the questionnaire, similar to the completion rate in our pilot phase. Of the individuals who did not complete the survey, 50% viewed the introduction page with the informed consent statement, but did not click through to the survey itself. This could be because participants were not interested in the scope of the survey, had negative feelings about the funders or institutions involved, or they did not agree with the information

presented in the consent form. Seventy-five percent of individuals who did not complete the survey stopped responding by the fifth question, which was still within the first of the four sections of the survey.

3.3. Analysis

We used the statistical program R to perform our analysis, in particular the Multiple Response Categorical Variable (MRCV) package (Koziol & Builder, 2014a, 2014b), to analyze questions with multiple possible responses. For these questions, we tested for multiple marginal independence (MMI) or simultaneous pairwise marginal independence (SPMI) between variables using the Bonferroni correction method. This method calculates a Bonferroni-adjusted p value (Dunn, 1961) for each possible 2 x 2 contingency table that can be constructed from a question's variables. These individually adjusted p values are then used to create an overall adjusted p value that indicates if MMI or SPMI exist; specific associations between variables can then be identified by comparing the p values from the individual 2 x 2 tables to the set significance level ($\square$ = 0.05) (Bilder & Loughin, 2015). This method was used to detect the correlations presented later in Figure 6, Figure 8, Figure 15, and Table 4. This test can produce overly conservative results, particularly when analyzing questions with many variables. Nonetheless, this approach is preferred to traditional tests for independence, as it takes into consideration the fact that a single individual can contribute to multiple counts within a contingency table. We coded and analyzed open-response questions in NVivo using a combined deductive and general inductive approach to thematic analysis (Thomas, 2006) and used R, Gephi, and Tableau to create plots and visualizations.

3.4. Reporting

Significant associations between variables are often reported by listing p values in tabular format for each combination of possible responses. Due to the complexity of our questionnaire, in particular the number of multiple response variables (Table 2), we present significant associations using visualizations. These visualizations indicate if an association between variables exists; they do not indicate the value of the p values themselves. We do this in order to increase the understandability of our results and to make them usable for a wider audience. Tables with the adjusted p values are included in the Supplementary Material for this article.

3.5. Limitations

The sampling methods, the survey design, and our analytical methods have both strengths and limitations (see Box 1). Our data and analysis are descriptive, not predictive, and only represent the practices of our respondents—a group of data-aware people already active in data sharing and reuse and confident in their ability to respond to an English-language survey. We also have limited knowledge about individuals who did not respond to the survey. The analysis presented here is not

generalizable to broader populations, but rather depicts the behaviors and attitudes of the approximately 1,700 respondents to our survey.

Box 1. Strengths and Limitations of Sampling, Questionnaire Design and Analysis Methods

Strengths and Limitations of This Study
Sample
·   Responses are from globally distributed, multidisciplinary individuals involved in research
·   Number of responses (n = 1,677)
·   Possible sampling bias due to coverage of Scopus
·   Response rate of 1.1%. Limited information about nonresponders.
Survey questionnaire
·   Multiple-response and open-response questions elicited rich qualitative and quantitative data
·   Survey questions based on past empirical work
·   Responses represent self-reported behaviors and attitudes
·   Responses to multiple-choice questions shaped by the options provided. Open response option included for each multiple-choice question.
Analysis
·   Analysis represents discovery phase for a rich, hard-to-obtain data set. Exploratory quantitative results combined with qualitative analysis.
·   Large sample size especially important in qualitative analysis.
·   Provides directions for future areas of more detailed and statistically powerful quantitative analysis
·   Exploratory tests for statistical significance are based on pairwise correlations.
·   Statistically significant results may be impacted by possible bias in the sample.

> ·   Complete-case analysis. Partial responses were not included.

The majority of our sample consists of researchers who have published in a journal indexed in the Scopus database. Certain disciplinary domains are better represented within Scopus; most notably, the arts and humanities are not as well covered (Mongeon & Paul-Hus, 2016; Vera-Baceta et al., 2019). Scopus has an extensive and well-delineated review process for journal inclusion; as of January 2020, 30.4% of titles in Scopus are from the health sciences, 15.4% from the life sciences, 28% from the physical sciences, and 26.2% from the social sciences (Elsevier, 2020). While the policies and content of Scopus could produce a potential bias in our sample, drawing from this population also helped us to target our desired population of researchers across domains.

Self-reported responses could be impacted by respondents' desire to give socially acceptable answers. Respondents may also have interpreted the survey questions in different ways; responses could be influenced by English-language fluency, individual interpretations of the wording of questions, the multiple-choice options and Likert scales provided, and the ordering of the questions themselves. We attempted to counter these issues by implementing a two-stage pilot phase and by providing open-response options for multiple-choice questions.

Our findings are based only on complete survey responses. While this facilitated our analysis, it is possible that this introduced additional biases to our sample as well as reduced potential statistical power (as described in Pigott, 2001). The test for independence that we apply to identify correlations among questions with multiple response variables has been suggested to err on the conservative side (Bilder, 2015), making it possible that some correlations were not detected with this method. The Bonferroni adjustment is an approximation; known issues include the identification of false negatives, particularly when comparing large numbers of variables, and determining the number of comparisons to use in the calculation (McDonald, 2014). For these reasons, we present our quantitative results as an initial exploration of the data, to guide deeper future statistical analyses.

3.6. Ethics and Data Availability

We received ethical approval from Maastricht University for the study. Participants had the opportunity to review the informed consent form prior to beginning the survey and indicated their consent by proceeding to the first page of questions.

The data from this survey are available in the DANS-EASY data repository under a CC-BY-4.0 license (Gregory, 2020).

4. Results and Analysis

We present our results according to the research questions presented in Table 1. Each subheading provides an answer to the proceeding research question, which is then further discussed and supported by the survey data. The research questions are therefore answered in the course of the results section, before the primary takeaways are summarized in the Conclusion. We first examine characteristics of the data seekers responding to our survey, and then proceed to look at their data needs, search and discovery strategies, and evaluation and sensemaking practices.

4.1. RQ 1: Who Are the People Seeking Data?

Respondents Are Globally Distributed and Have Research Experience. Respondents employed in 105 countries completed the survey. The United States, Italy, Brazil, and the United Kingdom were among the most represented countries (Figure 1a). This does not directly correspond with the sampling distribution (Figure 1b), where the largest difference between recruited participants and respondents occurred in China. This lower response rate could be due to language differences, perceived power differences (Harzing, 2006), or a lack of tradition in responding to research requests from researchers from other countries (Wang & Saunders, 2012). It could also indicate that data seeking is not a common practice.

The majority of survey respondents were researchers (82%) and worked in universities (69%); 40% of respondents had been professionally active for 6–15 years (Table 3). With the exception of participants recruited specifically from the library and research data management communities (or 'research support professionals'), our recruitment methodology ensures that all respondents are published authors, making it likely that they have been involved in conducting research in either their past or current roles. Nearly half of those working in research support also need secondary data for their own research, in addition to supporting researchers or students.
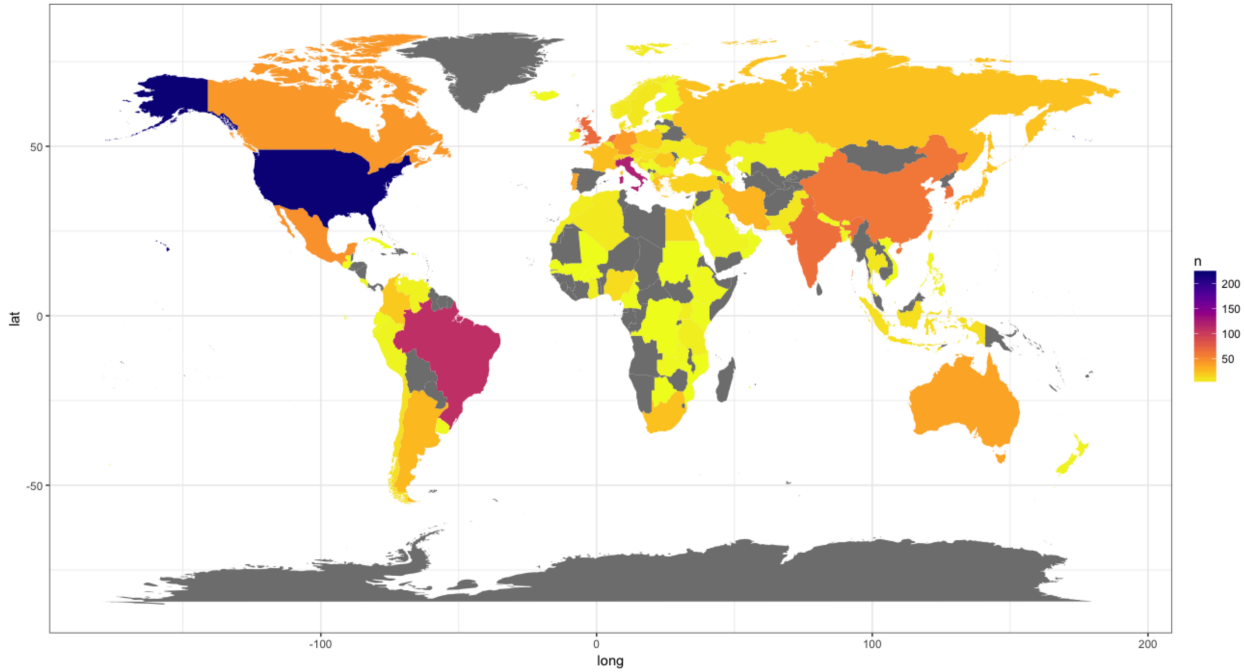
Figure 1a. Geographic distribution of respondents - Number of respondents by country of employment (n = 1,677)

Percent of recruited sample		Percent of respondents	
United States	19%	United States	13%
China	15%	Italy	7%
United Kingdom	5%	Brazil	7%
Germany	5%	United Kingdom	4%
Japan	4%	India	4%
France	4%	South Korea	4%
India	4%	Netherlands	4%
Italy	3%	China	4%
Canada	3%	Mexico	3%
Spain	3%	Canada	3%
Australia	3%	Germany	3%

South Korea	3%	Australia	2%
Brazil	2%	Portugal	2%
Netherlands	2%	Iran	2%
Russian Federation	1%	Argentina	2%
Taiwan	1%	France	2%
Iran	1%	Greece	1%
Swizerland	1%	Japan	1%
Turkey	1%	Russian Federation	1%
Poland	1%	South Africa	1%
Other	19%	Romania	1%

Figure 1b. Geographic distribution of respondents - Percent of recruited participants by country compared to percent of respondents by country ($n = 1,677$).

Table 3. Role, place of employment and years of professional experience of respondents (n=1677)

Role	Number	Percent
Researcher	1372	82
Student	73	4
Manager	54	3
Research support	47	3
Other	131	8
Organization		
University	1150	69

Research institution	287	17
Government agency	74	4
Corporate	68	4
Independant archive	2	0
Other	96	6
Professional experience		
0-5	206	12
6-15	677	40
16-30	502	30
31+	292	17

Respondents Support Data Sharing and Reuse. While 80% of all respondents reported sharing their own data in the past, participants with longer careers have done so slightly more often. Eighty-nine percent of respondents who have worked for 31+ years reported having shared their data, compared to 77% percent of respondents working for less than 5 years. Personal attitudes toward data sharing and reuse differ from the perceived attitudes of peers, disciplinary communities, and institutions (Figure 2). The majority of survey respondents are proponents of sharing their research data; they believe that the other groups identified in Figure 2 are less supportive of data sharing. This also points to a possible bias in our data, suggesting that people who have not shared their own data or who do not support data sharing were less likely to respond to our survey. Alternately, respondents could have felt compelled to provide a socially desirable answer, feeling a positive response was more acceptable. Respondents indicated that they believe data sharing is more strongly supported by their direct coworkers than by their disciplinary communities or institutions; they are most uncertain about the attitudes of their institutions. A similar pattern exists for attitudes toward data reuse, although there is more uncertainty involved.
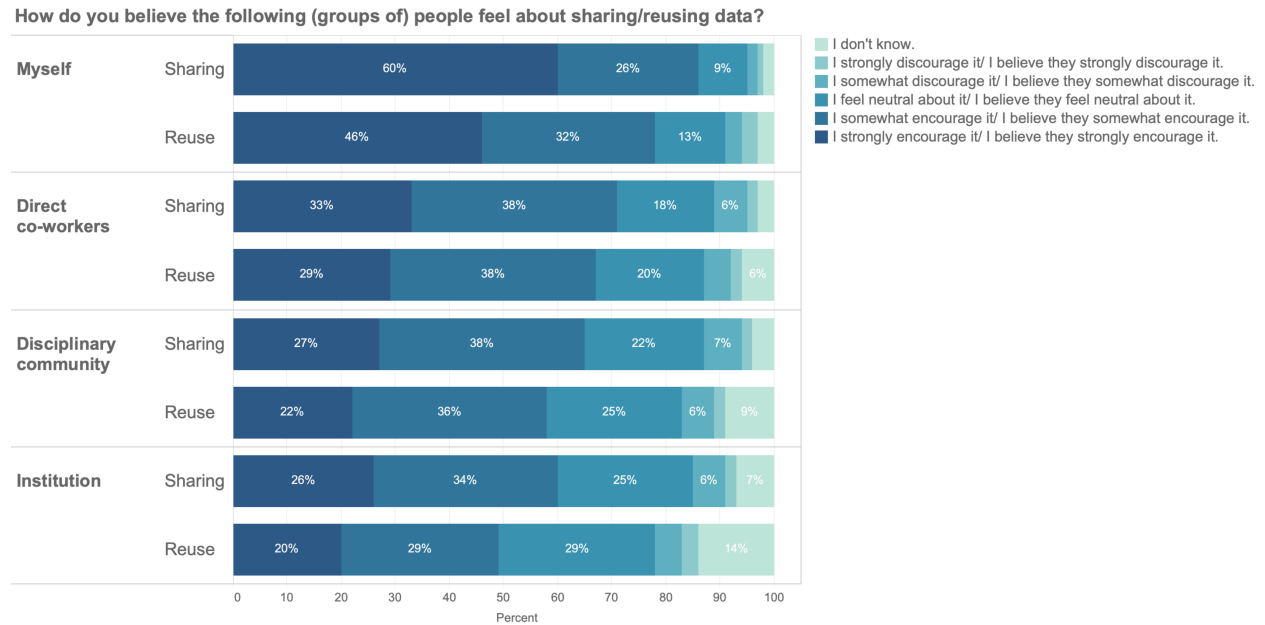
How do you believe the following (groups of) people feel about sharing/reusing data?

Figure 2. Respondents' beliefs about how they and other groups feel about data sharing and reuse (*n* = 1,677). Percent denotes percentage of respondents.

Respondents Belong to Multiple, Overlapping Domains. Respondents indicated their domain(s) of specialization from a list of 31 possibilities. This disciplinary list was originally used in a survey measuring data-sharing practices across disciplines (Berghmans et al., 2017). Engineering and technology was selected most often, followed by biological, environmental, and social sciences (Figure 3). Approximately half of the respondents selected two or more domains, with one quarter selecting more than three. This could be a factor of varying levels of granularity of the disciplinary list; it could also indicate that participants found it challenging to choose a single domain that captures the complexity of their expertise.

Figure 3 depicts the disciplinary overlaps among respondents, showing which domains were selected in conjunction with each other. The figure reveals expected disciplinary overlaps, for example, between information science and computer science, between medicine and health professions, and between material science and chemistry. Other overlaps are perhaps less expected, for example, between social science and engineering and technology or between arts and humanities and computer science, which could be indicative of the use of digital humanities methodologies among our participants. Seventy-one percent of respondents who selected engineering and technology chose at least one other discipline, most frequently computer science, environmental science, material science, and energy, as is visible in Figure 3.
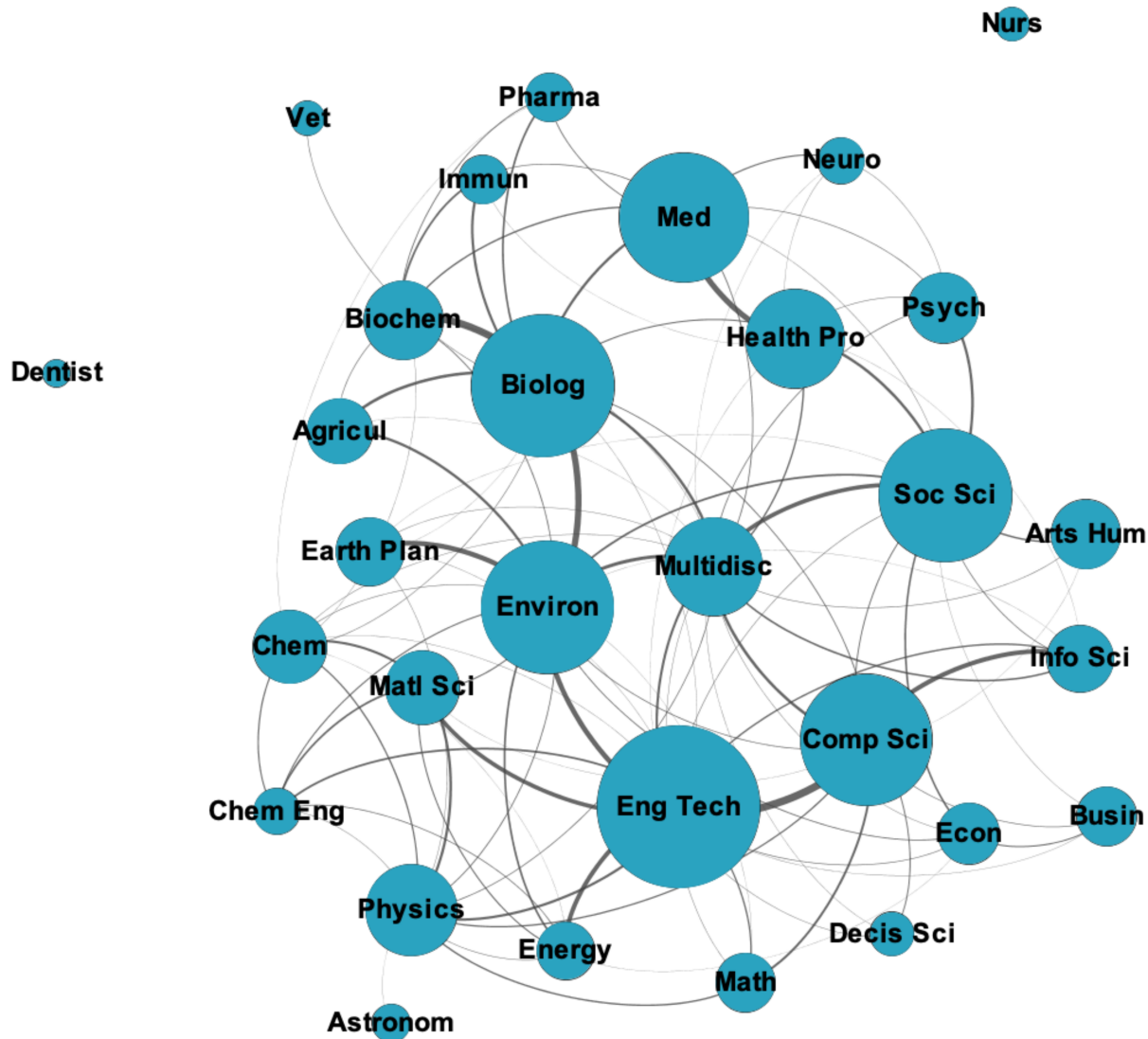
Figure 3. Disciplinary domain responses (*n* = 3,431). Node size represents the number of responses from a discipline. Width of the edges represents the number of times two disciplines were selected in conjunction with each other. Only edges with a weight greater than 20 are shown; some edges pass behind other nodes. Graph created in Gephi; graph shape set using ForceAtlas algorithm, repulsion strength = 9,000; other parameters set to default.

We also identified individuals who selected only one discipline. The greatest number of single-discipline responses were in the domains of medicine, social science, engineering and technology, and computer science (Figure 4). In future sections, we augment our analysis of respondents across all disciplinary domains with an occasional analysis of a subset of the individuals selecting only one discipline.

In this disciplinary subset, we included domains with more than 40 respondents, with the exception of the "other" category, as well as other domains whose data practices have been well-documented in the literature, specifically astronomy, environmental sciences, and earth and planetary sciences (see Gregory et al., 2019b, for a review of the literature exploring these disciplines). Disciplines that are included in this disciplinary subset are marked with a double asterisk in Figure 4.
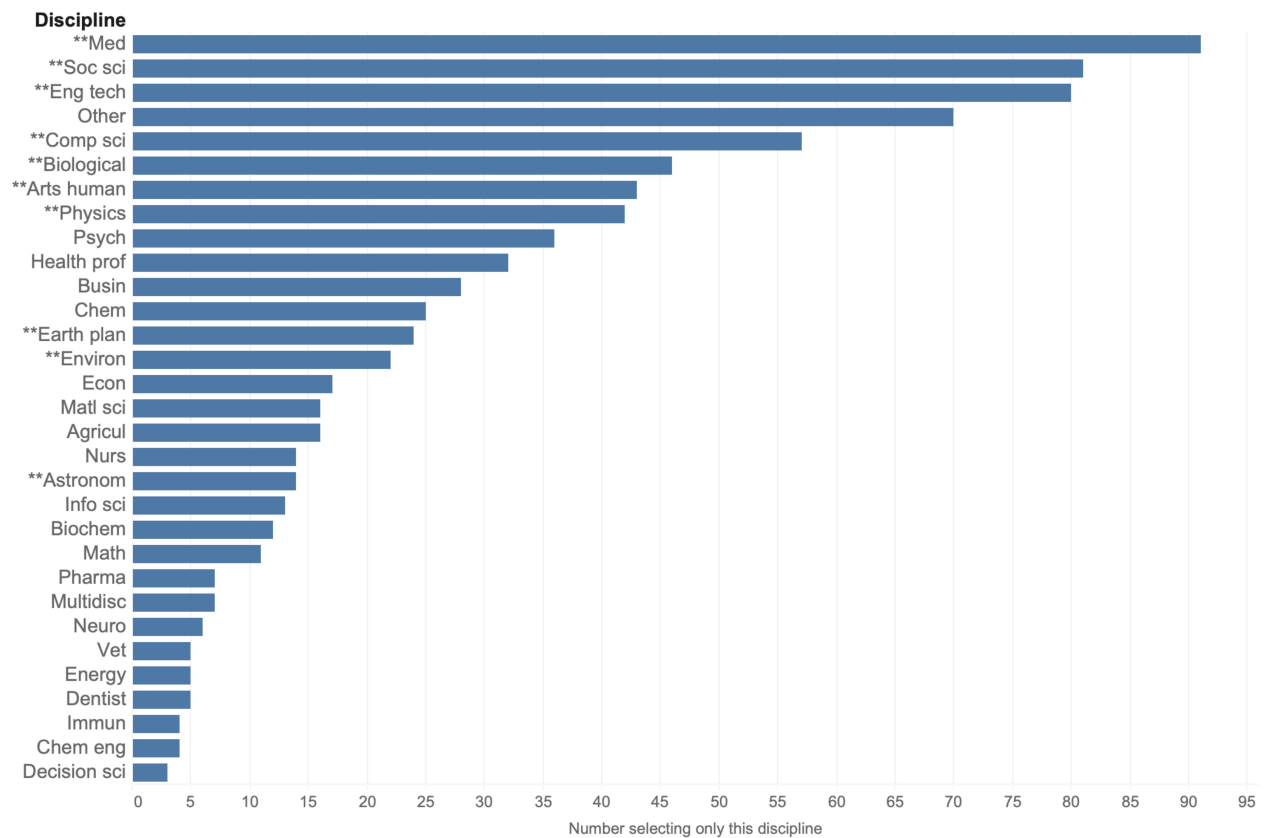

Figure 4. Number of respondents who selected only one discipline (*n* = 836). Disciplines included in the subset for further analysis are marked with **.

4.2. RQ 2: What Data Are Needed for Research and How Are Those Data Used?

Data Needs Are Diverse and Difficult to Pigeonhole. To provide a high-level view of their data needs, respondents selected the type(s) of data that they need in their work from a list derived from the categories of research data proposed by the U.S. National Science Board (2005) and used in Berghmans et al., 2017, and Gregory et al., 2019b. While observational/empirical data were selected most

frequently, 50% of participants also indicated that they need more than one type of data. Figure 5 represents the number of respondents selecting individual and multiple data types.

Please select the options that describe the secondary data that you (might) need.

❑ **Observational or empirical** (e.g. sensor data, survey data, interview transcripts, sample data, neuroimages, ethnographic data, diaries)

❑ **Experimental** (e.g. gene sequences, chromatograms, toroid magnetic field data)

❑ **Simulation** (e.g. climate models, economic models)

❑ **Derived or compiled** (e.g. text and data mining, compiled database, 3D models)
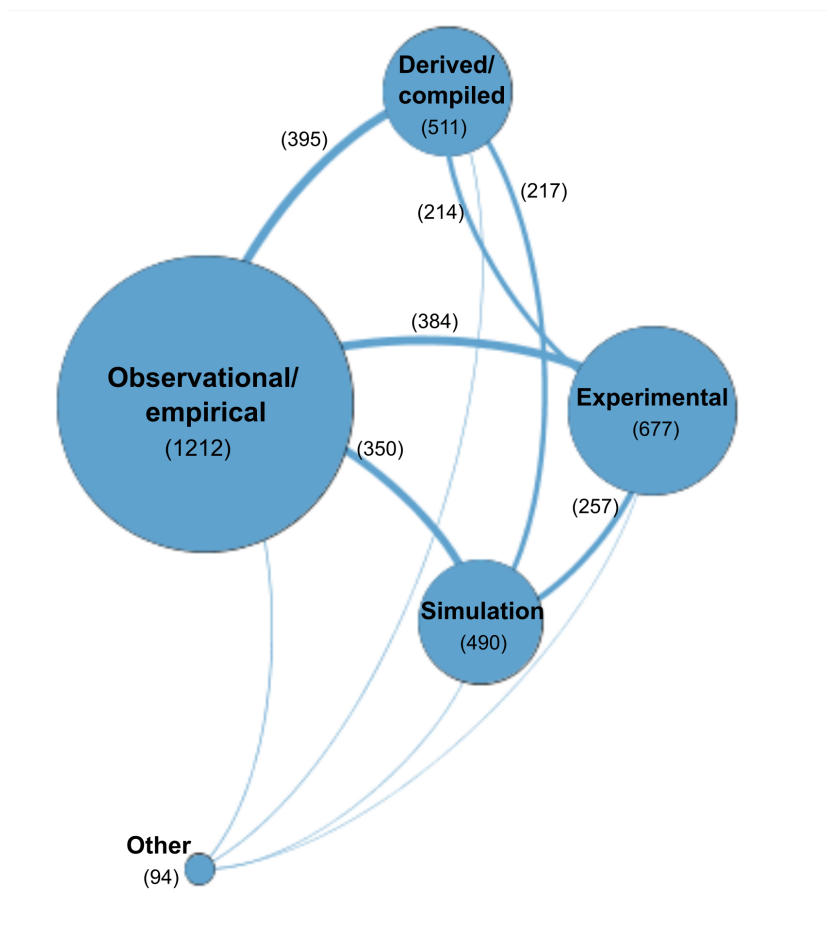
❑ **Other,** Please specify _____

Figure 5. Question from survey with descriptions of data types. Node size in visualization represents number of responses for each data type (n = 2,984). Edges represent number of times both data types were selected in conjunction with each other.

We asked participants to further expand on their data needs in an open-response question, which 78% of respondents completed. We compared these responses to the types of data that participants selected, paying special attention to those who chose multiple data categories. Respondents selecting multiple data types appear to do so for different reasons. Some require a variety of topic-specific data to conduct their research; others use a variety of data, regardless of topic, as long as it matches format and structure requirements. The following participant indicated needing observational, experimental, and derived data.

> [I need] large and small data sets that students can use for Data Science skill development. For example, transactional data from business, medical data such as interactions between patients and doctors, descriptive medical histories. Data need to be complex enough to be interesting and able to be parsed. (Respondent ID 613).

Data are difficult to categorize (Borgman, 2015b; Leonelli, 2019); in part because people may define data categories differently. While the majority of respondents stating they use census data selected observational data, others did not, choosing instead derived/compiled data. While the majority of individuals using literature corpora selected derived/compiled data, a minority selected "other," apparently not knowing which category best fit their data.

Associations (detected using the test for independence described in the Methodology section) between disciplinary domain and the type of data needed are shown in Figure 6. The highest number of disciplinary associations were detected for experimental data; arts and humanities was associated with the greatest number of data types, although surprisingly not with observational data. There are also domains where no associations were detected. This could be due to the composition of our sample; it could also be taken as evidence for the diversity of data types that participants within disciplines need.
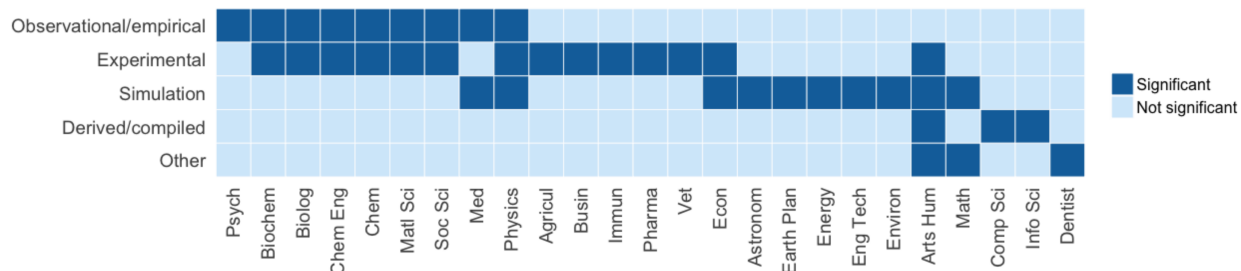
Harvard Data Science Review

Lost or Found? Discovering Data Needed for Research

Figure 6. Associations between disciplinary domain and needed data. Associations detected using adjusted Bonferroni test for simultaneous pairwise marginal independence (n = 1,677; significance level: $p < 0.05$).

Slightly more than half of respondents reported needing data outside of their disciplines. This pattern holds across domains, with the exception of environmental science, where more individuals need data external to their domain (65%), and medicine, where the trend was reversed, with less than half needing this type of data (43%).

Data Uses Are Diverse But Limited to a Core Set. Although using data in ways that support research is well-documented (e.g., Wallis et al., 2013), using data to drive new research or in teaching is not as well documented (Gregory et al., 2019b). The majority of respondents (71%) selected using data as the basis for a new study, which is in line with other very recent research (Pasquetto et al., 2019); half of respondents selected needing data to use in teaching (Figure 7). Other uses include experimenting with new methodologies and techniques such as developing data science skills or completing particular data-related tasks, such as trend identification or creating data summaries (as suggested by Koesten et al., 2017). Less than 2% of respondents indicated needing data for other purposes, suggesting that the list of uses in our survey is fairly complete.
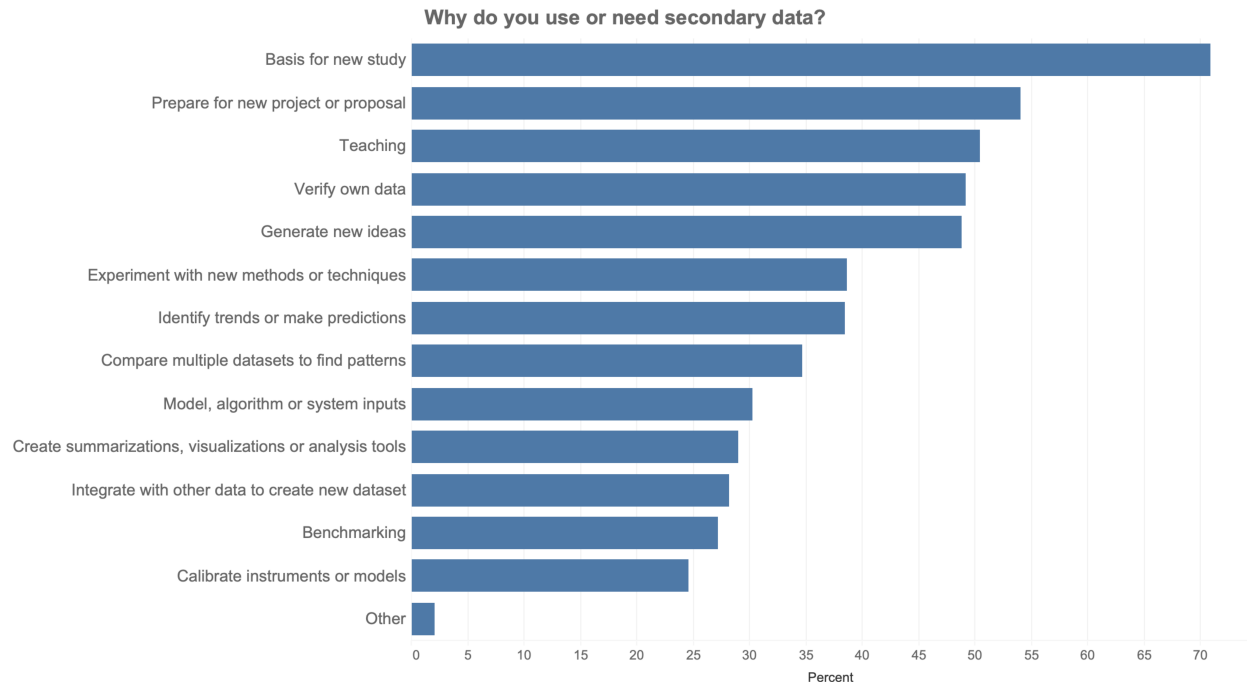
Figure 7. Reasons why respondents need secondary data. Multiple responses possible. Percent denotes percentage of respondents (n = 1,677).

Data Uses Span Research Phases. Table 4 presents significant associations between the data uses shown in Figure 7 and the types of needed data shown in Figure 5. We also tested to see if significant associations exist between the different types of data uses (Table 4). We recognize that associations between uses are difficult to interpret; if an individual selected teaching, for example, it may or may not be related to their selection of another data use. To examine these associations through a different perspective, we organize them along a conceptual typology, which we propose here, based on different phases of research.

To create this typology, we drew on the associations we identified, existing literature, and extant models of research lifecycles. In Table 4, we agree with van de Sandt et al. (2019) that reuse is one type of data use, but similar to Fear (2013), we view reuse as referring to using data as the basis for a new study. Our typology has similarities with models of the research lifecycle process (e.g., Jisc, 2013), but it also differs from these models. Typical models of research lifecycles portray research that uses data created by the researcher, rather than secondary data. They also tend to reduce the complexity of research processes (Borgman, 2019), ignore the interwoven nature of tasks involved in research (Cox & Wan Ting Tam, 2018), and depict data cycles as independent workflows (e.g., UK Data Service, 2019). With this typology, we attempt to nuance the uses of secondary data throughout phases of research, recognizing how data uses are associated with multiple data types and uses and highlighting that these uses are associated with multiple work phases.

Depending on research practices, the data uses in our typology could occur in different phases. Uses that could particularly fall into two bordering categories are marked in grey in Table 4. Integrating different data sets could occur when conducting research, for example, or verification of data could also be considered to be part of data analysis. Although not marked in grey, data analysis and sensemaking tasks are likely to occur throughout all phases of research. This is reflected in the results, where analysis activities are associated with every other data use, with the exception of instrument or model calibration.

Table 4. Associations Between Types of Data Use, Needed Data Type and Other Data Uses

Harvard Data Science Review

Lost or Found? Discovering Data Needed for Research

Research/work phase	Types of data uses	Associated with these types of needed data	Associated with these types of data uses
Reuse	Basis for new study (new study)	observational/empirical	new project, new ideas, integration, comparison, trends, teaching
Project creation and preparation	Prepare for new project or proposal (new project)		new study, new ideas, integration, new methods, verification, trends, comparison, summaries/vis/tools, teaching
	Generate new ideas (new ideas)		new study, new project, integration, new methods, verification, trends, comparison, summaries/vis/tools, teaching
Conducting research	Integrate with other data to create new dataset (integration)	observational/empirical, derived/compiled	new study, new project, new ideas, new methods, inputs, trends, comparison, summaries/vis/tools
	Experiment with new methods or techniques - e.g. data science skills (new methods)	experimental, simulation, derived/compiled	new project, new ideas, integration, inputs, calibration, benchmarking, verification, trends, comparison, summaries/vis/tools
	Model, algorithm or system inputs (inputs)	simulation, derived/compiled	integration, new methods, calibration, benchmarking, trends, comparison summaries/vis/tools
	Calibrate instruments or models (calibration)	experimental, simulation, derived/compiled	new methods, benchmarking, inputs, verification
	Benchmarking (benchmarking)	simulation, derived/compiled	new methods, inputs, calibration, verification, trends, summaries/vis/tools
	Verify own data (verification)	experimental, simulation, derived/compiled	new project, new ideas, new methods, calibration, benchmarking, comparison, teaching
Data analysis and sensemaking	Identify trends and make predictions (trends)	observational/empirical, simulation, derived/compiled	new study, new project, new ideas, integration, new methods, inputs, benchmarking, comparison, summaries/vis/tools
	Compare multiple datasets to find commonalities or differences (comparison)	experimental, derived/compiled	new study, new project, new ideas, integration, new methods, inputs, verification, trends, summaries/vis/tools
	Create summaries, visualizations or analysis tools (summaries/vis/tools)	observational/empirical, simulation derived/compiled	new project, new ideas, integration, new methods, inputs, benchmarking, trends, comparison, teaching
Teaching	Teaching/training (teaching)	observational/empirical	new study, new ideas, new project, verification, summaries/vis/tools

Note. Grey areas represent uses that could fall within multiple bordering research phases. Colors correspond to research/work phases. Associations detected using adjusted Bonferroni test for simultaneous pairwise marginal independence (significance level: $p < 0.05$).

Using data as the basis for a new study is associated in our results with needing observational data. Among our respondents, observational data are not significantly associated with the tasks in the 'conducting research' phase, but they are associated with activities in both data analysis and with teaching. Experimental data are related to uses involved in conducting research, as well as to comparison. Derived data are associated with all activities in both the research process and data analysis/sensemaking phases; simulation data are primarily associated with conducting research.

Using data as the basis for a new study is associated with uses such as project creation and preparation, data analysis and sensemaking, or teaching, but is not associated with any uses in the conducting research phase. Teaching is associated with only a few other uses, most of which fall within the project creation phase. Data integration is associated with activities across phases of research; calibration, however, is exclusively associated with other research process tasks.

Data Uses Are Common, But Their Enactments Are Complex. Disciplinary domains also shape how data are used. Figure 8 indicates significant associations between disciplines and uses in our sample. Most of these associations are for data uses that fall within the conducting research phase of the above typology, particularly using data as inputs and for calibration, and domains that typically make use of modeling or computational research methods. Due to the large number of variables compared in Figure 8 and the limitations of our test for significance, these associations in particular represent initial results.

To complement the associations presented in Figure 8, we also looked at the subset of our data for individuals selecting only one discipline (indicated in Figure 4). Within these disciplinary groups, we also found that respondents use data for a variety of purposes, rather than for just one type of use. While researchers in multiple domains may use data for the same purposes, uses will be enacted in different ways and have different meanings in various disciplines and contexts (Borgman, 2015b, Leonelli, 2015). While 39% of computer scientists in the disciplinary subset and 35% of those in the arts and humanities selected using data for verification, for example, the actual practice and the meaning of verifying data will be different in each of these disciplines.
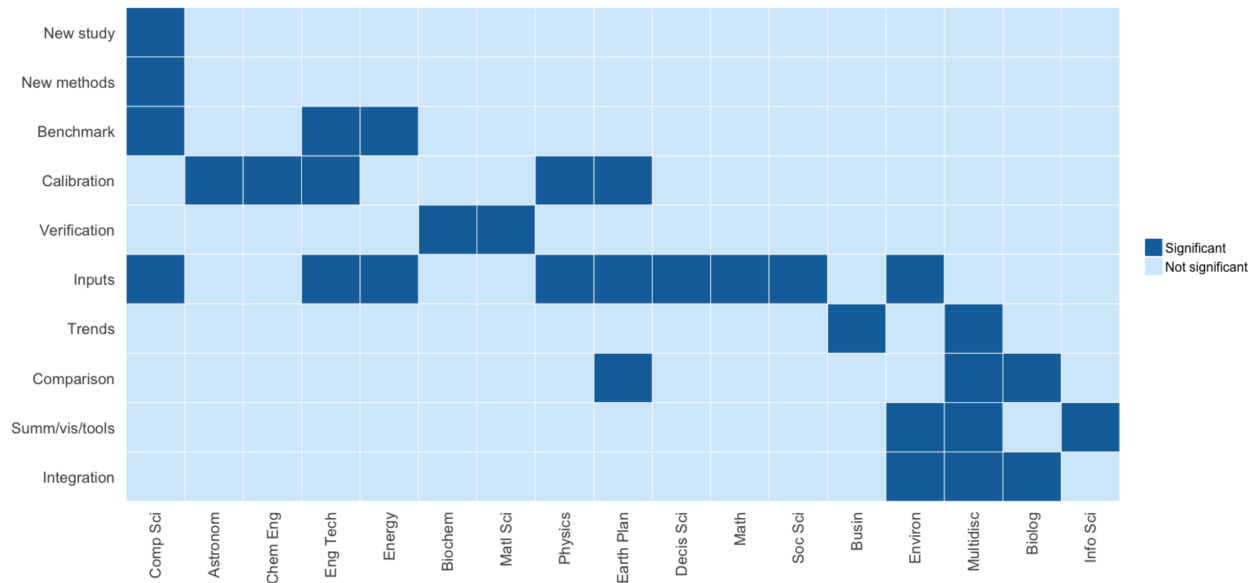
Figure 8. Associations between disciplinary domain and data use. Associations detected using adjusted Bonferroni test for simultaneous pairwise marginal independence (n = 1,677; significance level: p < 0.05).

4.3. RQ 3: How Are People Discovering Data?

Researchers believe that discovering data is a sometimes challenging (73%) or even difficult (19%) task. The greatest challenge researchers face is that data are not accessible, that is data are not available for download or analysis (68% of researcher respondents), followed by the fact that data are distributed across multiple locations (49%). One third of these respondents identified inadequate search tools, a lack of skill in searching for data, or the fact that their needed data are not digital as being additional challenges.

Via Academic Literature. Thirty percent of respondents reported no difference in how they find literature and how they find data. Figure 9 presents the disciplines selected by these respondents. Some of these respondents chose multiple disciplines; the analysis in Figure 9 is therefore not limited to respondents in the disciplinary subset. Disciplines with data repositories that are closely linked with systems for searching the academic literature of that field, such as in the biomedical sciences and physics (i.e., the resources from the National Library of Medicine and HEP-INSPIRE database, respectively), rank more highly. Respondents in disciplines where data repositories and academic literature databases are traditionally less integrated with each other, such as in the social sciences (e.g., the ICPSR database) or the environmental sciences (e.g., PANGAEA), tend to have more distinct discovery practices.
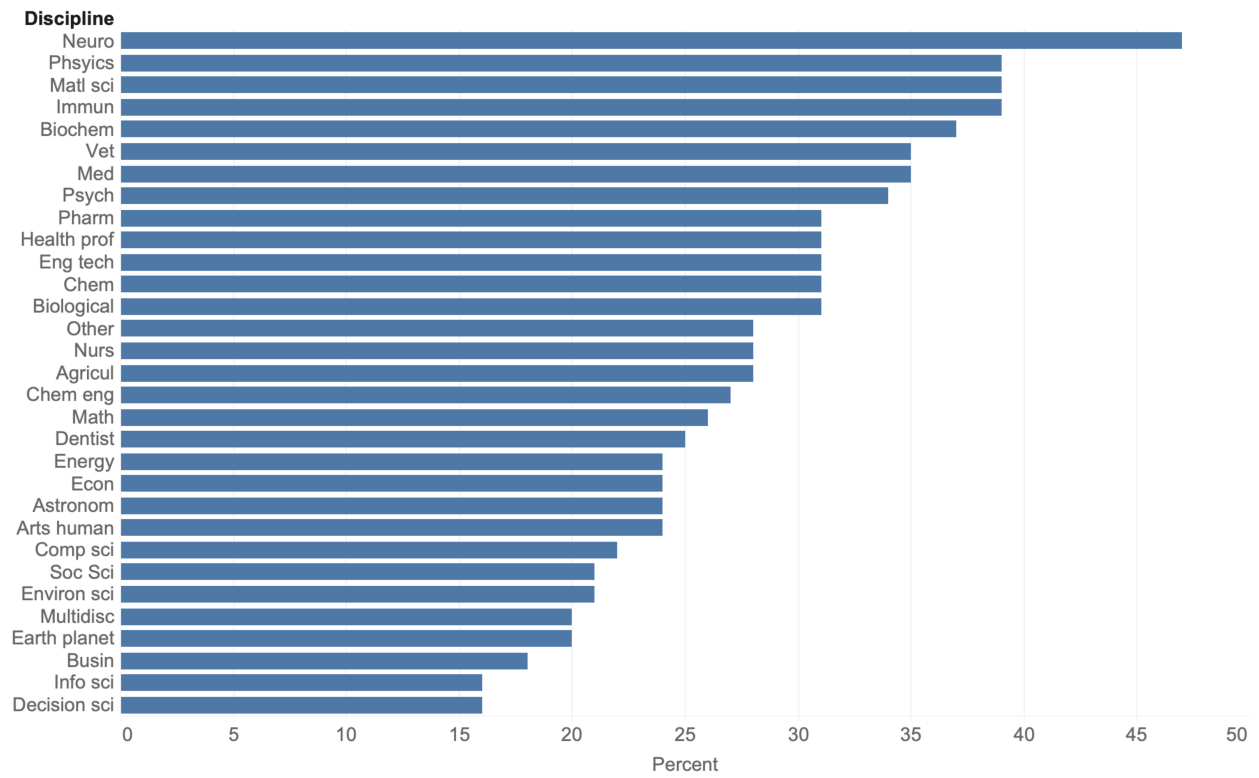
Figure 9. Percentage of responses (_n_ = 959) by discipline for individuals reporting no difference in literature and data search strategies. Multiple responses possible.

Fifty-two percent of respondents stated that their processes for finding literature and data are sometimes the same and are sometimes different; whereas they were always different for 18%. Respondents saying that the two processes are sometimes or always different (_n_ = 1,178) were asked to explain the differences in an open response question.

One of the key differences participants identified between literature search and data discovery are the sources that are used.

> I check other channels for data than for literature, e.g., if a project produces data, I check the project's site directly for their data and hope for links to repositories. (Respondent ID 4001)

> Academic literature could be found through different portals...To receive data, one often needs to know where to find it. For example, the name on the database and then contact the administrator

> for the database if you can't extract the data directly from the database. (Respondent ID 4008)

Yet the academic literature itself is a key source for discovering data for researchers, as are general search engines (e.g., Google) and disciplinary data repositories (Figure 8). Research support professionals rely less on the literature, more frequently turning to a variety of sources in their search for data (Figure 10). The importance of literature, search engines, and domain repositories across disciplines supports the findings from earlier work (Kratz & Strasser, 2015).
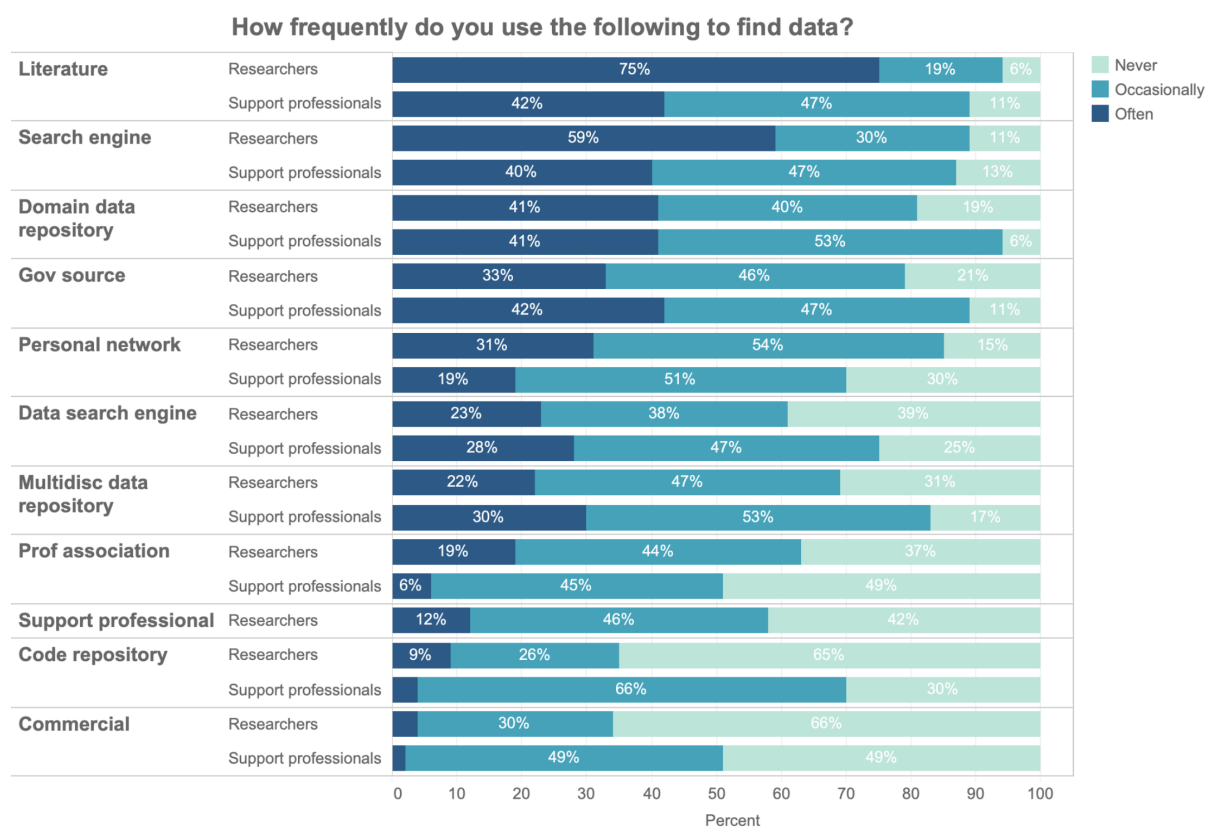

Figure 10. Sources used to find data by researchers (including students, managers, and others, *n* = 1,630) and research support professionals (*n* = 47). Percent denotes percentage of respondents for each category. Listed in order of decreasing importance for researchers.

The distribution of sources presented in Figure 10 generally holds across disciplines for researchers; literature, followed by search engines or domain repositories, are reported as used most often in nearly all domains. There are some disciplinary differences, identified by looking at the subset of respondents selecting only one discipline. In the arts and humanities, for example, turning to research support professionals was selected more often than in other disciplines; in computer science, 76% of

respondents selected that they occasionally or often consult code repositories. (A breakdown of the use of sources according to domain for our disciplinary subset is included in the supplementary materials).

Respondents use literature as a source of data—plucking data from reported tables and graphs[2]—but they also use the literature to track down the original data, making use of behaviors common in literature searching, such as citation chasing (Figure 11; the distribution presented in Figure 11 remains the same when looking at the percentage of overall responses to this question.) It is common for respondents to first find the literature, and then use the literature as a gateway to locating the data. This strategy is often planned, but it also happens serendipitously while reading or searching for literature. Roughly two thirds of participants also often or occasionally find data serendipitously outside of the literature (e.g., via email or conversations with colleagues) or in the course of sharing or managing data.

> Finding data is different because it often occurs as a result of finding academic literature. (Respondent ID 738)

> Literature is more direct; data is more like "bonus" finds. One finds interesting data in other contexts of work in a publication, one can contact the author to ask for the data. (Respondent ID 3179)
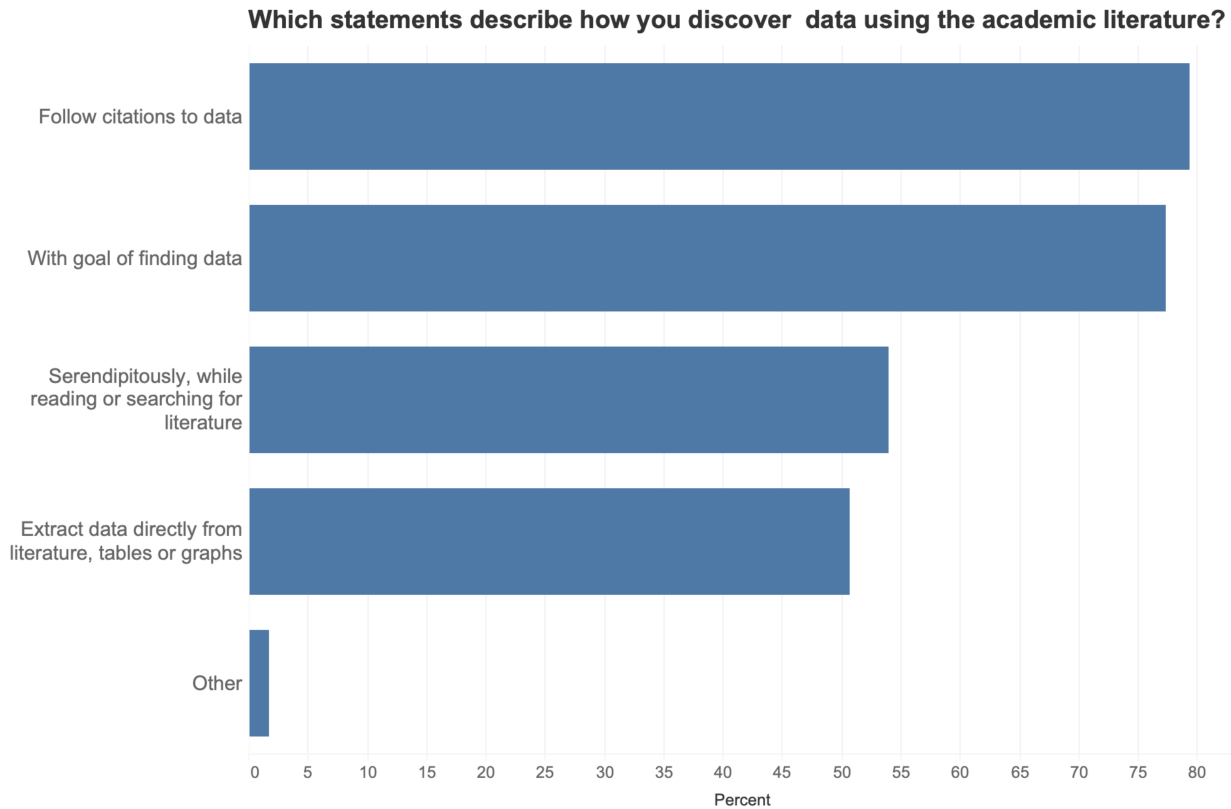
Which statements describe how you discover data using the academic literature?

Figure 11. Strategies for using the academic literature to discover data. Question asked to respondents who indicated using literature as a source. Percent denotes percentage of respondents; multiple answers were possible (n = 1,573).

Via Social Connections. Using social connections and personal outreach to discover and access data is another important difference identified between literature search and data search. This is reflected in Figure 10, where only 15% of researchers never make use of personal networks in data discovery.

> Unlike academic literature where you get the data by accessing the journal, finding data often requires contacting the institution that created the data. (Respondent ID 2357)

> I use personal networks and public access datasites to discover data, then I usually have to submit a proposal and get it accepted in order to get access to the data. I have not had the experience of just downloading data directly without going through a permission process. (Respondent ID 1416)

Attending conferences and having discussions within personal networks are the most frequent ways of mobilizing social connections to discover data (Figure 12). While personal networks remain important in actually gaining access to data, contacting data authors directly is the most often reported method for accessing data. Forming new collaborations with data creators also appears to be more important in accessing data than in first discovering them. These patterns hold, regardless of the types of data that respondents need or their intended use for the data, although there are some disciplinary differences in the percentage of respondents discovering or accessing data via conference attendance or forming new collaborations. The need to use personal connections in accessing data also reflects the finding that access remains the largest hurdle for participants.
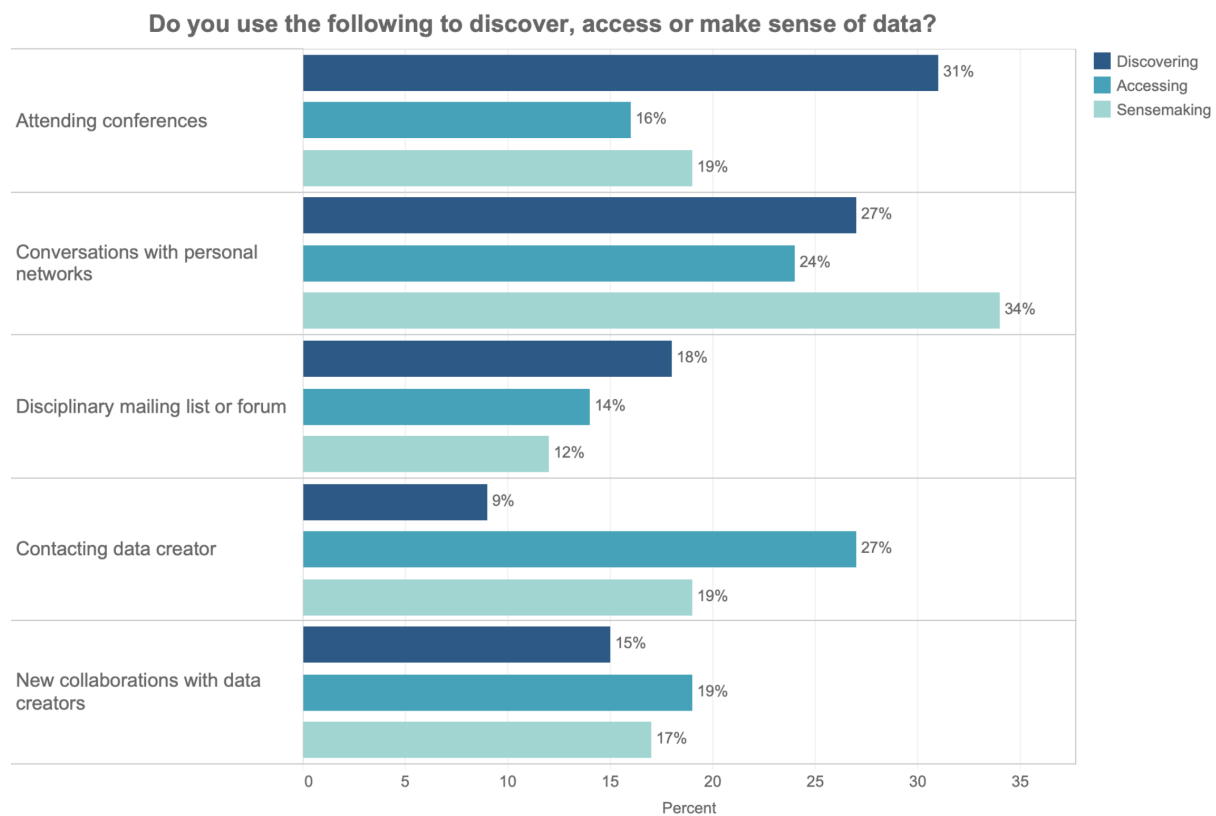
Do you use the following to discover, access or make sense of data?

Figure 12. How respondents make use of social connections in discovering data (n = 3,311), accessing data (n = 3,589), and making sense of data (n = 3,031). Percent denotes percentage of responses for each option; multiple responses possible.

Social connections are a two-way street. Not only do respondents seek data from their networks but experts also receive data without solicitation or in exchange for their knowledge. Participants receive data from individuals both within and outside of their domains of expertise.

Harvard Data Science Review

Lost or Found? Discovering Data Needed for Research

> Specialist organizations often seek my help and their members send me data. I write to schools to invite them to participate in the experimental work I do and I analyze their data for them and send them reports and suggestions to overcome the difficulties that I observe. (Respondent ID 1242)

> I am an expert in statistics. I didn't have to find the data; the researchers that owned the data find me. I have published a few original articles not being part of my field of expertise as a co-author—as the researcher responsible for the statistical analysis. (Respondent ID 3253)

Via 'Mediated' Search. For some respondents, actually locating and accessing data is a mediated process, mediated not through the work of information professionals (although this sometimes happens—see Figure 10), but rather through the literature and through personal connections. Numerous respondents first discover or encounter data via an 'intermediary' source—an article, a conversation with a colleague; they then turn to another source—a data repository, Google—to search specifically for the known data.

> I generally do not search for data blindly but I would normally know that it already exists through some previous interaction (reading scientific publication, personal communication). (Respondent ID 679)

General search engines (e.g., Google) can also serve in intermediary roles, as respondents use them not only in order to find data themselves but also to locate data repositories. These two practices—using Google for known-item searches and to locate repositories rather than data—could contribute to the fact that 38% of respondents who use general search engines found their searches to be either successful or very successful. However, the majority of respondents using general search engines reported mixed success (55%), with 7% being rarely or never successful, perhaps reflecting the higher failure rate in general in academic search compared to general web searching (Li et al., 2017).

Via Specific Searches Plus Casting a Wide Net. Much of data searching is very specific. Participants rely on particular, known data repositories and sources. Respondents have specific requirements and search parameters, and seek data for specific purposes and goals, as is evidenced in our typology of

data (re)use (Table 4). This is in contrast to literature searching, where participants report using cross-disciplinary sources, such as the Web of Science or Scopus, and where the goal is often to cast a wide net to discover ideas for use in theory or concept development.

> When I search for data I am pretty focused on finding only data sets that I need for a specific purpose. When I search for literature I read papers that are only peripherally related to the subject but they help me formulate new ideas. (Respondent ID 2128)

> I tend to search for data by specifying parameters e.g., geographical and date coverage, or by looking for data created by a specific organization. My literature searches are more general and don't have so many search filters applied. (Respondent ID 3688)

In contrast, searching for data is haphazard and less systematic than literature search for many respondents, requiring researchers to cast a wide net to discover distributed data.

> It [data search] is a little more haphazard, as I am not as comfortable with finding data. Some of this stems from my not knowing "the" sources, but some of it is also because the finding tools are not yet available. Many times, it is a "try and see" approach. (Respondent ID 3803)

> With scientific literature I know for sure where to look for it, before I start the search. In other words, sources are known to me and do not change for years. With data it is always not so. I may find data in unexpected places. (Respondent ID 3152)

By Building New Practices. This state of development causes search practices to be in flux, as individuals figure out how best to find and access the data that they need. In contrast to their practices for searching literature, their data search practices are still in formation.

> Finding academic literature is part of everyday practice. The processes for finding literature are well-established and institutionally-supported. If I need to find data I have to establish my own

> process to locate where they are held, get permission from owners, agree access rights etc.
> (Respondent ID 696)

4.4. RQ 4: How Are People Evaluating and Making Sense of Data for (Re)use?

By Using Varied Evaluation Criteria and Sensemaking Strategies. Respondents require a variety of information about the data and make use of different sensemaking strategies (Figure 13). Eighty-nine percent of respondents reported that information about data collection conditions and methodology was important or extremely important in their decisions; information about data processing/handling as well as topical relevance were also ranked highly (Figure 13a). The ease (or difficulty) of accessing data is also very important to 73% of participants. While respondents take the reputation of the data creator into consideration, with 62% of respondents indicating this is important or extremely important, the reputation of the source of the data (e.g., the repository or journal) appears to be slightly more important, as 71% of respondents identified the reputation of the source as important or extremely important. This is further evidenced later in Figure 16a; 61% of respondents selected the data creator's reputation as important/extremely important in establishing trust in secondary data, as compared to 81% who identified the reputation of the source as key to developing trust.

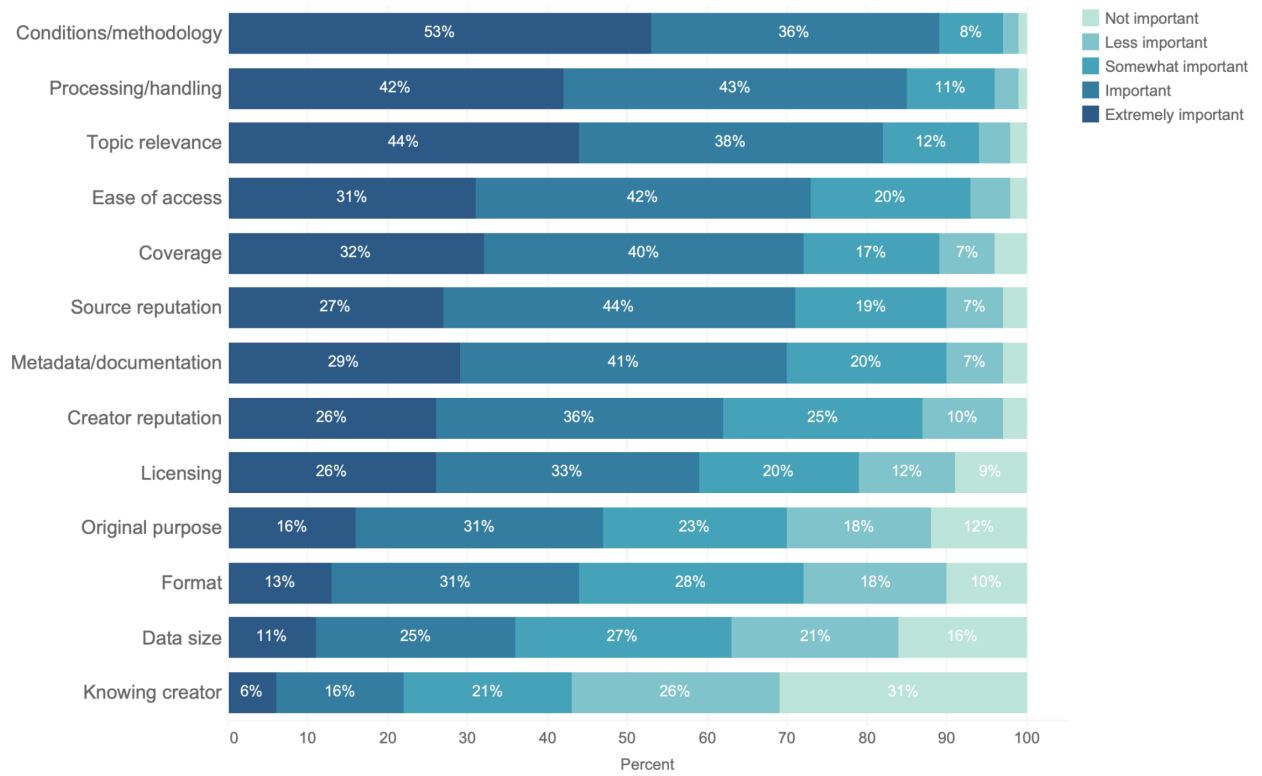**How important is the following information when deciding whether to use secondary data?**

Figure 13a

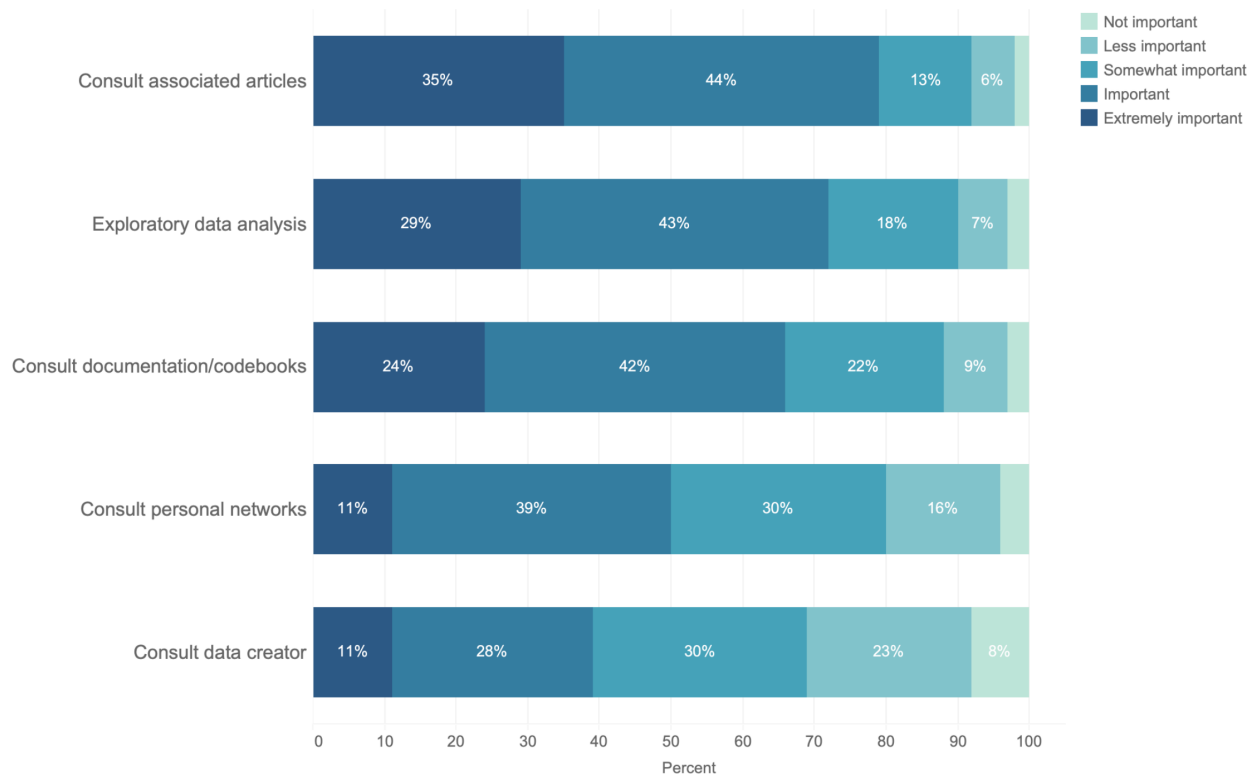**How important are the following strategies when deciding to use secondary data?**

Figure 13b

Figure 13. Evaluation criteria and sensemaking strategies A) Information used in evaluating data for reuse ($n = 1,677$). Percent denotes percentage of respondents. B) Sensemaking strategies ($n = 1,677$). Percent denotes percentage of respondents.

Other information identified in open responses includes the timeliness of data, prior usage, and the cost of obtaining data, which can determine the type of research that is pursued.

> Patent data is free to access. Data on company deals and revenue can sometimes be paid. That requires seeking research funding and typically delays the process. Using publicly free data is quicker. (Respondent ID 3220)

The academic literature plays a key role not only in discovering data, but also in understanding them. Respondents consult associated articles, as well as data documentation and codebooks (Figure 11b). Nearly three quarters of respondents report engaging in exploratory data analysis, such as performing statistical checks or graphical analysis. Participants also report triangulating data from multiple

sources as a way of understanding and determining the validity of data (e.g., Respondent IDs 3131, 2444, 1949).

By Using Social Connections in Sensemaking. Fifty percent of participants reported conversations with personal networks as being key to making sense of data. Conversations with networks are used more often in sensemaking than in either discovering or accessing data (Figure 12). Contacting data creators to make sense of data does not happen as frequently as discussions with personal networks.

Respondents also attend conferences and form new collaborations to make sense of data (see Figure 12). Some variations in the pattern in Figure 12 for sensemaking exist across disciplinary domains (e.g., see Figure 14), although engaging in conversations with personal networks is almost always chosen most frequently. These variations are likely the result of different disciplinary norms and infrastructures, which influence patterns of collaboration and communication (e.g., the role of conference attendance and publishing norms, or the existence of disciplinary mailing lists and forums, i.e., in computer science).
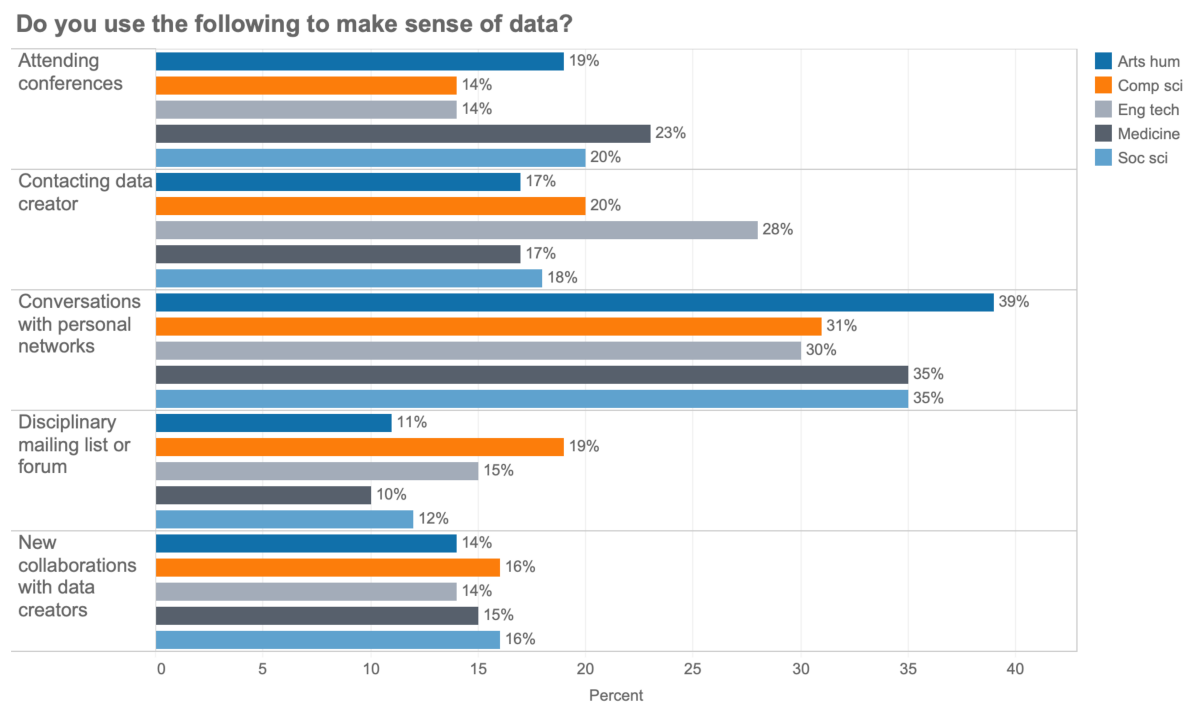
Do you use the following to make sense of data?

Figure 14. Social strategies of sensemaking in five disciplines from the disciplinary subset: arts and humanities (n = 64), computer science (n = 94), engineering and technology (n = 109), medicine (n = 158), and social science (n = 155). Percent denotes percentage of responses; multiple responses were allowed.

Harvard Data Science Review

Lost or Found? Discovering Data Needed for Research

By Using Different Contextual Information for Different Purposes. Different data uses are associated with needing different information about data. Figure 15 presents significant associations, detected using the statistical test for multiple marginal independence described in the Methodology section, between data uses and evaluation criteria. An association detected with this method indicates that responses to the survey question about evaluation criteria are correlated with responses to the question about data uses. In Figure 15, we classify the evaluation criteria presented in Figure 13a into content-related information (e.g., data collection methods and conditions, the relevance of data to a topic, the exact coverage of the data), structure-related information (format, size, the existence of detailed documentation and metadata), access-related information (ease of access, licensing), and social information (reputation of data creator and source, knowing the data creator). We then plot significant associations that exist between uses in the (re)use typology and these evaluation criteria.

This analysis allows us to begin to identify the types of information needed by respondents in different research phases. It also allows us to identify gaps. Most of the detected associations occur between content-related information and data uses in the project creation/preparation or analysis/sensemaking stages of our (re)use typology. The fewest number of significant associations were detected with calibration and benchmarking. Only one association was found for structure-related information, between teaching and data format; source reputation was also correlated with format and idea generation. The importance of information about how data are processed and handled span uses in all of the research phases; having access to detailed and complete metadata or documentation was found to be associated with experimenting with new techniques and methods, with integration, and with making comparisons.
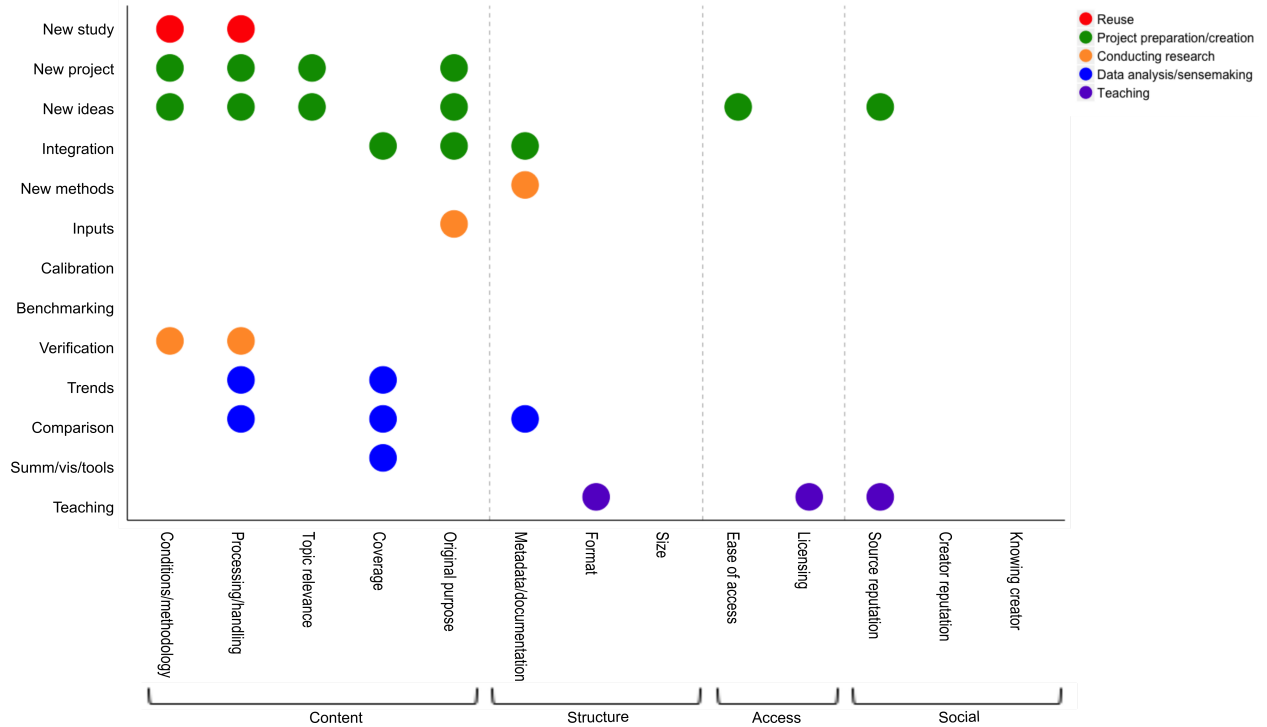
Figure 15. Significant associations between types of data use and information important in evaluating data. Associations detected using adjusted Bonferroni test for multiple marginal independence (significance level: $p < 0.05$). Colors represent phases of (re)use typology. Classifications of evaluation criteria marked with brackets.

By Establishing Trust and Data Quality. The transparency of data collection methods, followed by the reputation of the source and a minimum of errors, are critical in trust development (Figure 16a).
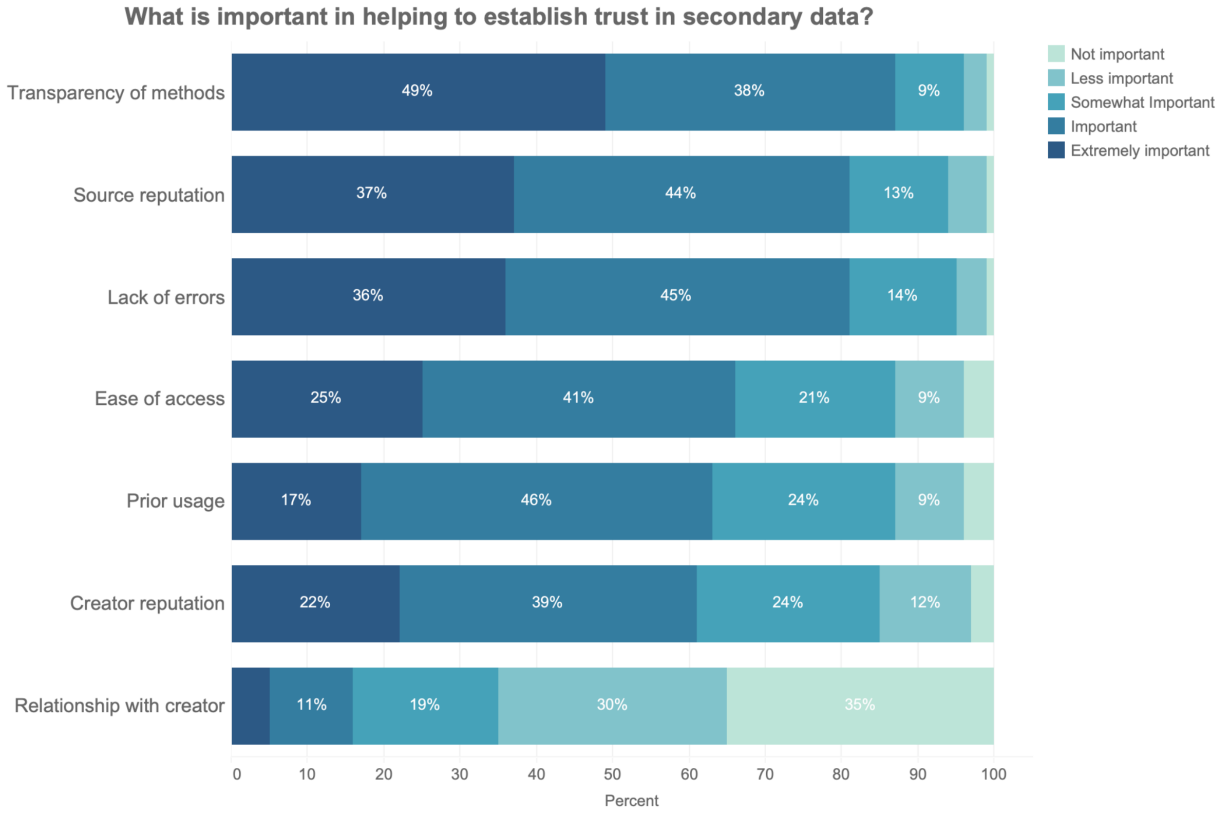
What is important in helping to establish trust in secondary data?

Figure 16a

What is important in helping to establish the quality of secondary data?
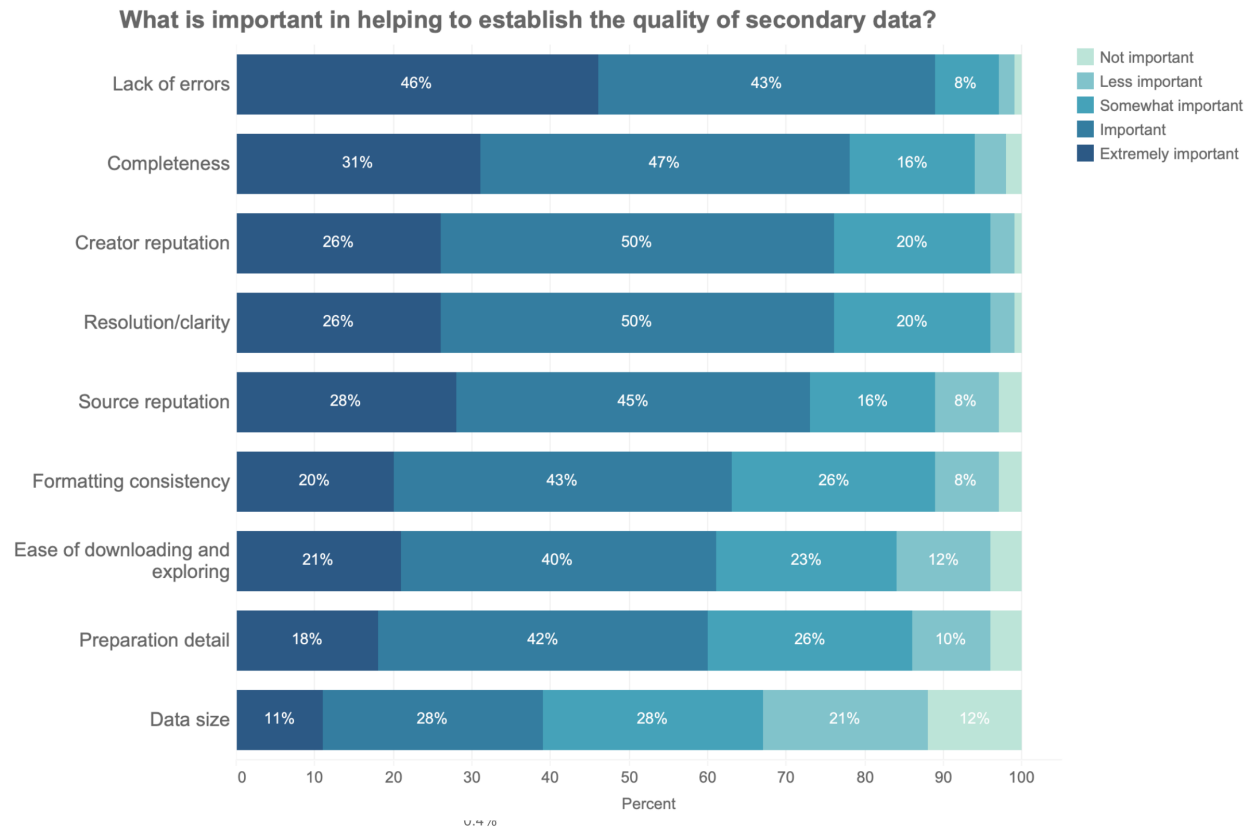

Figure 16b

Figure 16. Criteria for establishing trustworthiness and quality of data A) Importance of criteria used to establish **trust** in secondary data ($n = 1,677$). Percent denotes percentage of respondents. B) Importance of criteria used to establish **quality** of secondary data ($n = 1,677$). Percent denotes percentage of respondents.

In some disciplines, a completely error-free data set may actually raise suspicions, as it may indicate that the data have been tampered with or manipulated.

> Lack of errors would not necessarily help establish trust—errors are normal, so a perfect data set without errors might be a fabricated data set. (Respondent ID 3970)

> If you know the field you also know what to look for with respect to unreliable data. Sometimes the occasional error actually speaks to the reliability of a data set: It indicates a person was involved somewhere in data entry. (Respondent ID 1648)

Establishing data quality also depends heavily on the absence of errors and data completeness. Both developing trust and determining quality involve social considerations (Faniel & Yakel, 2017; Yoon, 2017). Although respondents across disciplines consistently ranked having a personal relationship with the data creator as being unimportant in establishing trust, they still weigh other 'social' factors—such as thinking about human involvement in data creation or the reputation of the source—in their decisions. The reputation of the data creator appears to be more important to respondents when evaluating data quality than in trust development.

5. Discussion

We identify and apply four analytical themes to further discuss our findings about practices of data seeking and reuse. We also consider each theme's relation to recent efforts to increase the discoverability of research data before concluding by suggesting future areas for both practical and conceptual work.

5.1. Communities of Use

The term 'community' is often used without considering how communities are formed or their exact composition. Community boundaries are shifting and porous, rather than fixed and stable, and individuals often belong to multiple communities simultaneously (Birnholtz & Bietz, 2003).

We see this clearly in our data. Although communities are typically thought of in terms of disciplinary domains, more than half of our respondents identified with multiple disciplines. Fifty percent also indicated needing data outside of their domains of expertise, perhaps reflecting funders' efforts to promote interdisciplinary research (Allmendinger, 2015). Data communities can also be thought of in terms of the type of data that a particular group uses (Cooper & Springer, 2019; Gregory et al., 2019b). However, we show here that respondents need multiple types of data for their work, and that these data needs can be difficult to classify in broad terms.

Communities can form around particular methodologies and ways of using and working with data (Leonelli & Ankeny, 2015), as is the case, for example, in the digital humanities and sociology or economics (Levallois et al., 2013). Our (re)use typology (Table 4) allows for conceptualizing data communities in terms of broad uses of data, for example, using data in conducting research, and in more specific terms, for example, using data for calibration. We also found initial signals that particular data uses are associated with needing certain information about data (Figure 15). Content-related metadata, such as information about collection conditions and methodologies, is important for our respondents in preparing for new projects; structure-related metadata, for example, format, is

important in teaching. We saw a similar relationship, particularly for teaching data science, in our qualitative data. This suggests another way of conceptualizing data-seeking communities—by broad research approaches. Individuals relying on data science techniques, no matter their discipline, may rely more on structure-related information when evaluating data for reuse; content-specific considerations could be more important for more traditional research approaches.

Open data policies and guidelines recognize the importance of communities, but often equate communities with disciplinary domains. The FAIR data principles (e.g., Wilkinson et al., 2016) call for the use of domain-relevant community standards as well as relevant attributes to facilitate findability and reuse. Our analysis encourages a multidimensional way of thinking about communities, recognizing that community-relevant metadata can be defined by considering other factors (e.g., data use) in conjunction with disciplinary domains.

5.2. Interwoven Practices

Data discovery is interwoven with other (re)search practices, particularly searching for academic literature. Roughly 80% of respondents stated that their practices for finding data and literature are either sometimes or always the same. The academic literature itself is the go-to source for finding data for the majority of participants. Despite the immature state of data citation practices in many disciplines (Robinson-Garcia et al., 2016), respondents use a strategy common in literature searching—following citations—to locate data from the literature.

Data citation is not equivalent to bibliographic citation (Borgman, 2016). Most data citations indicate some type of data 'usage' (Park & Wolfram, 2017), but little is known about why people cite data or the details of how data have been used in a work (Silvello, 2018). This presents an additional challenge for people seeking data to use for a particular purpose. A citation model that typifies how data have been used could potentially facilitate data discovery and evaluation practices that begin by following data citations, as well as add value to the multiplicity of data uses that we observed.

5.3. Social Connections

Discovering and accessing data are also mediated by personal networks. Respondents find out about data from their connections and then hunt down the data digitally. This process also occurs in reverse— respondents find data digitally and then access the data by personally contacting data creators. The use of social connections in discovery and sensemaking is intertwined with discipline-specific practices of communication (e.g., the role of conference attendance) and collaborations.

Participants identify using social connections as an important difference between searching and accessing data, as opposed to literature. This difference could be a result of the fact that infrastructures to support data search and access are still in development. It could also be due to the complexity of the sociotechnical issues surrounding data access. Researchers question which data to

make available for whom (Levin & Leonelli, 2017), and sensitive data containing private information about participants cannot be made openly accessible.

Accessible data, as defined by the FAIR principles, do not necessarily need to be open data (Mons et al., 2017). Access to data can be mediated by automated authorization protocols (Wilkinson et al., 2016), but automatic denials of access may not mean that data are completely closed to a human data seeker. Researchers can still contact data authors directly if access is denied to learn more about restrictions and possibly form collaborations that would enable reuse (Gregory et al., 2018). Our results also show that ease of access is a top consideration in using data, especially during early phases of research (Figure 15). As certain data become easier to seamlessly and automatically access, other data, those that are more challenging to access, will likely not be used as often, which will shape the research that is or is not pursued.

5.4. Practices in Flux

Practices and infrastructures are closely linked (Shove et al., 2015); this is especially true for practices of data discovery and reuse. We see this in the tension that we found between specific and haphazard search practices. For some respondents, data infrastructures are still in a state of development, which requires casting a wider net to locate appropriate sources. For others, finding data involves going directly to a particular, well-known data repository in the field.

Data infrastructures consist of assemblages of policies, people, technology, and data (Borgman et al., 2015, Edwards, 2010). As data are described in more standardized ways, repositories will have different methods of structuring data, be linked to other data and repositories more seamlessly, and will build new services. These services and linkages will change how data seekers interact with and discover data.

Innovations combining new technologies with existing practices will not only alter current practices, but will also bring new considerations to the forefront. Executable papers (e.g., Gil et al., 2016), where readers can interact with data directly, build on linkages between literature and data searching as well as on the importance of exploratory data analysis in sensemaking. They also blur the line between where a paper ends and data begins. As the boundaries between data and papers become less defined, the importance of archiving those data in sustainable ways (e.g., Vander Sande et al., 2018) and questions of properly citing data creators, rather than paper authors, will become more visible.

6. Conclusion

In summary, we have examined the data needs, uses, discovery strategies and sensemaking practices for the 1,677 respondents to our survey by presenting an initial quantitative analysis and by drawing on the qualitative survey data. Possible practical applications for this work are many. We conclude by

highlighting some key takeaways from our analysis and draw attention to their potential applications, in particular for designers of data discovery systems and managers of data repositories.

6.1. Diversity Is Normal, Not Abnormal

Past in-depth ethnographic work has documented the diversity of data practices in particular research groups and projects within disciplines such as astronomy, environmental engineering, and biomedicine (see Borgman, Wofford et al., 2019). Both our quantitative and qualitative results show that this multiplicity is not limited to these specific communities. Rather than being the exception, a diversity of data needs, data uses, and data sources appears to be the rule. We also see signals of other forms of diversity in our results, finding, for example, that data needs and search practices are both specific and broad, that data uses are spread across phases of research, and that practices of data discovery are intermingled with other search practices.

We have suggested elsewhere that data discovery systems implement a variety of differentiated search interfaces, including visual and graphical navigation systems, for different users (Gregory et al., 2019a), which would allow data seekers freedom to explore data in different ways. The diversity we observe here also supports the creation of flexible, interlinked designs for data discovery systems and repositories with different levels of specificity. Linking diverse data across locations via standardized approaches is gaining significant momentum (see, e.g., Wilkinson et al., 2016), and would pave the way for general data search engines and federated search efforts. Domain agnostic data search engines (i.e., Google Dataset or DataSearch) allow data seekers to search for data broadly before being directed to data repositories, where more specific searching and exploration is often possible. This puts an onus on repositories, however, as they become responsible for implementing a new generation of tools supporting more specific search and sensemaking activities within their environments (i.e., generous interfaces, Mutschke et al., 2014, and Whitelaw, 2015).

Data search engines could also feasibly help searchers looking for data outside of their domain to better identify potential repositories of interest. Our findings also suggest the need to integrate data search tools with literature databases and data management tools, which we also suggested based on our earlier interviews with data seekers (Gregory et al., 2019a). Figure 9 shows that this integration may be more relevant for certain disciplines than for others.

6.2. Communities of Use as an Entry Point to Design

While we suggested differentiated interfaces for different users in the past, the multiplicity of practices we observed makes it challenging to design tools based on 'user profiles'; there is also often a gap between the user profiles that designers imagine and the actual users themselves (see Wyatt, 2008). Both our quantitative and qualitative data suggest that searching for data is purpose-driven; researchers look for data for a specific purpose or use. This finding is supported by Koesten et al.

(2017), who use interviews with data professionals and log analyses to suggest that data search is task-dependent. We suggest here also that data communities can be conceptualized by uses of data and show in Figure 7 that these uses, although spread across research phases and not limited to a particular point in time (see also Pasquetto et al., 2019), appear to be limited to a core set. This set of data uses could provide an entry point for design, allowing data seekers to search for or filter by data used for these specific purposes.

The set of uses we identify is general and needs further research to test its comprehensiveness. Performing cluster analyses using the data from our survey may be one way to further identify and validate the *communities of use* that we propose. It is also likely that repositories may be able to identify their own specific communities of use. Identifying these will require innovative approaches by repositories and communication with both data depositors and consumers (see Section 6.4).

6.3. Metadata to Support Sensemaking and Reuse

Metadata plays a key role in facilitating data reuse (Mayernik, 2016; Mayernik & Acker, 2017; Pasquetto et al., 2019); however, the general metadata needed for discovery is often not rich enough to support the sensemaking needed for reuse (Zimmerman, 2007). In Figures 13 and 16, we identify common evaluation criteria and considerations in developing trust and quality that could inform the development of metadata, at both broader and more specific levels, to support making decisions about using data. Figure 15, in particular, suggests that certain evaluation criteria can be used to support certain types of data uses, although additional testing needs to be done to validate the associations we detected and to investigate how the factors identified in Figures 13 and 16 vary by use or disciplinary domain.

6.4. Tenacity and Value of the Social

An increasing amount of quantitative (e.g., Digital Science et al., 2019) and qualitative work (e.g., Yoon, 2017) shows that using secondary data requires communication and collaboration. Our findings demonstrate the importance of social interactions in discovery, access, and sensemaking, in particular, showing that researchers rely on conversations with personal networks to make sense of secondary data. We believe that these social interactions will continue to be important in data discovery and reuse, and should be seen as something to support when designing systems and repositories.

Our earlier suggestions, for example ranking data sets by social signals or integrating offline and online interactions around data, still hold promise (see Gregory et al., 2019a). We could also see a role for an expanded metadata schema as a way to open a conversation between data producers and multiple data consumers within the context of the data themselves at the repository level. In such a system, both data producers and data consumers could contribute information about the data to specific fields. Data producers could describe their own use of the data as well as their concerns about

what potential reusers need to know about the data before reusing them. Data users could describe how they used or plan to use the data, as well as pose questions about the data to other users or to the producer. Such an implementation would result in a layered metadata record, including general metadata needed for discovery, metadata with common elements supporting sensemaking (e.g., those identified in Figure 13), and a layer of interactive, or communication-based metadata. In the future, we envision a coevolution of metadata schemes driven both by designated communities and repository managers. Such a dynamic way to co-construct metadata schemas and indexing could enable repositories to identify new communities of use.

Our quantitative data provides the opportunity to more deeply probe and test the results we present here. Creating multilevel models to further explore the influence of data uses, data types, or disciplinary domains could be a fruitful next step, as could investigating methods to test the generalizability of our data to broader populations. This future work, in conjunction with the results in this article, could also inform deeper theoretical work. Designing useful, sustainable tools and services requires considering the interconnections between different practices, infrastructures, and communities that we have begun to investigate. Further conceptual work needs to be done to highlight these connections in a way that can be easily communicated and that can practically inform design.

Disclosure Statement

Acknowledgments

Supplementary Materials

Supplementary materials for this article can be found [here](.).

References

Adams, A., & Blandford, A. (2005). Digital libraries' support for the user's "information journey." *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries–JCDL '05*, 160. https://doi.org/10.1145/1065385.1065424

Allen, M. (2017). Secondary data. In *The SAGE Encyclopedia of Communication Research Methods*. https://doi.org/10.4135/9781483381411

Allmendinger, J. (2015). *Quests for interdisciplinarity: A challenge for the ERA and HORIZON 2020* [Policy Brief]. Research, Innovation, and Science Policy Experts (RISE) No. EUR 27370). https://ec.europa.eu/research/openvision/pdf/rise/allmendinger-interdisciplinarity.pdf

Belkin, N. J. (1993). Interaction with texts: Information retrieval as information-seeking behavior. In G. Knorz, J. Krause, & C. Womser-Hacker (Eds.), *Information retrieval '93: Von der Modellierung zur Anwendung* (pp. 55–66). Universitaetsverlag Konstanz.

Belkin, N. J. (1996). Intelligent information retrieval: Whose intelligence? In J. Krause, M. Herfurth, & J. Marx (Eds.), *ISI '96: Proceedings of the Fifth International Symposium for Information Science* (pp. 25–31). Universitaetsverlag Konstanz.

Berghmans, S., Cousijn, H., Deakin, G., Meijer, I., Mulligan, A., Plume, A., & Waltman, L. (2017). *Open data: The researcher perspective*. Elsevier; Leiden University's Centre for Science and Technology Studies (CWTS). https://www.elsevier.com/about/open-science/research-data/open-data-report

Bilder, C. R., & Loughin, T. M. (2015). *Analysis of categorical data with R*. CRC Press.

Birnholtz, J. P., & Bietz, M. J. (2003). Data at work: Supporting sharing in science and engineering. *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work*. https://dloi.org/10.1145/958160.958215

Blandford, A., & Attfield, S. (2010). Interacting with information. *Synthesis Lectures on Human-Centered Informatics*, *3*(1), 1–99. https://doi.org/10.2200/S00227ED1V01Y200911HCI006

Borgman, C. L. (2015a). If data sharing is the answer, what is the question? *ERCIM NEWS*, *100*, 15–16.

Borgman, C. L. (2015b). *Big data, little data, no data: Scholarship in the networked world*. MIT Press.

Borgman, C. L. (2016). Data citation as a bibliometric oxymoron. In C. R. Sugimoto (Ed.), *Theories of informetrics and scholarly communication* (pp. 93–115). Walter de Gruyter GmbH & Co KG.

Borgman, C. L. (2019). The lives and after lives of data. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.9a36bdb6

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., & Traweek, S. (2015). Knowledge infrastructures in science: Data, diversity, and digital libraries. *International Journal on Digital Libraries, 16*(3–4), 207–227. https://doi.org/10.1007/s00799-015-0157-z

Borgman, C. L., Scharnhorst, A., & Golshan, M. S. (2019). Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *Journal of the Association for Information Science and Technology, 70*(8), 888–904. https://doi.org/10.1002/asi.24172

Borgman, C. L, Wofford, M. F, Golshan, M. S, Darch, P. T, & Scroggins, M. J. (2019). *Collaborative ethnography at scale: Reflections on 20 years of data integration*. UCLA: Center for Knowledge Infrastructures. https://escholarship.org/uc/item/5bb8b1tn

Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez-Gonzalez, L.-D., Kacprzak, E., & Groth, P. (2020). Dataset search: A survey. *The VLDB Journal, 29***,** 251–272. https://doi.org/10.1007/s00778-019-00564-x

Cooper, D., & Springer, R. (2019). *Data communities: A new model for supporting STEM data sharing*. Ithaka S&R. https://doi.org/10.18665/sr.311396

Cox, A. M., & Wan Ting Tam, W. (2018). A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management, 70*(2)*,* 142–157. https://doi.org/10.1108/AJIM-11-2017-0251

Digital Curation Centre. (n.d.). Overview of funders' data policies. Retrieved June 28, 2018, from http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies.

Digital Science, Hahnel, M., Fane, B., Treadway, J., Baynes, G., Wilkinson, R., Mons, B., Schultes, E., Olavo, L., da Silva Santos, B., Arefiev, P., & Osipov, I. (2018). *The state of open data report 2018 (Version 2)*. https://doi.org/10.6084/m9.figshare.7195058.v2

Digital Science, Fane, B., Ayris, P., Hahnel, M., Hrynaszkiewicz, I., Baynes, G., & Farrell, E. (2019). *The state of open data report 2019* (Version 2). https://doi.org/10.6084/m9.figshare.9980783.v2

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. MIT Press.

Elsevier. (2020). *Scopus content coverage guide*. Elsevier BV. Retrieved March 19, 2020, from https://www.elsevier.com/__data/assets/pdf_file/0007/69451/Scopus_ContentCoverage_Guide_WEB.

[pdf](#).

European Commission. (2019). *Facts and figures for open research data*. Retrieved August 7, 2019, from https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-science-monitor/facts-and-figures-open-research-data_en

Faniel, I. M., & Yakel, E. (2017). Practices do not make perfect: Disciplinary data sharing and reuse practices and their implications for repository data curation. *Curating Research Data, Volume One: Practical Strategies for Your Digital Repository*, *1*, 103–126.

Fear, K. (2013). *Measuring and anticipating the impact of data reuse* [Doctoral dissertation]. University of Michigan. https://deepblue.lib.umich.edu/handle/2027.42/102481

Gil, Y., David, C. H., Demir, I., Essawy, B. T., Fulweiler, R. W., Goodall, J. L., Karlstrom, L., Lee, H., Mills, H. J., Oh, J.-H., Pierce, S. A., Pope, A., Tzeng, M. W., Villamizar, S. R., & Yu, X. (2016). Toward the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance. *Earth and Space Science*, *3*(10), 388–415. https://doi.org/10.1002/2015EA000136

Gregory, K. (2020). Data discovery and reuse practices in research [Data set]. DANS-EASY. https://doi.org/10.17026/dans-xsw-kkeq

Gregory, K., Khalsa, S.J., Michener, W.K., Psomopoulos, F. E., de Waard, A., & Wu, M. (2018). Eleven quick tips for finding research data. *PLOS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1006038

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019a). Understanding data search as a socio-technical practice. *Journal of Information Science*. Advance online publication. https://doi.org/10.1177/0165551519837182

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019b). Searching data: A review of observational data retrieval practices in selected disciplines. *Journal of the Association for Information Science and Technology, 70*(5), 419–432. https://doi.org/10.1002/asi.24165

Harzing, A.-W. (2006). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management*, *6*(2), 243–266. https://doi.org/10.1177/1470595806066332

Ingwersen, P. (1992). *Information retrieval interaction*. Taylor Graham.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, *52*(1), 3–50. https://doi.org/10.1108/eb026960

Jisc. (2013). *How Jisc is helping researchers: Research lifecycle diagram.* https://webarchive.nationalarchives.gov.uk/20140702225057/http://www.jisc.ac.uk/whatwedo/campaigns/res3/jischelp.aspx#simulate

Kacprzak, E., Koesten, L. M., Ibáñez, L.-D., Simperl, E., & Tennison, J. (2017). A query log analysis of dataset search. In J. Cabot, R. De Virgilio, & R. Torlone (Eds.), *Web Engineering. ICWE 2017. Lecture notes in computer science* (Vol. 10360, pp. 429–436). Springer. https://doi.org/10.1007/978-3-319-60131-1_29

Kim, Y., & Yoon, A. (2017). Scientists' data reuse behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology, 68,* 2709–2719. https://doi.org/10.1002/asi.23892

Kim, Y., & Zhang, P. (2015). Understanding data sharing behaviors of STEM researchers: The roles of attitudes, norms, and data repositories. *Library & Information Science Research, 37*(3), 189–200. https://doi.org/10.1016/j.lisr.2015.04.006

Koesten, L. M., Kacprzak, E., Tennison, J. F. A., & Simperl, E. (2017). The trials and tribulations of working with structured data: A study on Information seeking behaviour. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1277–1289. https://doi.org/10.1145/3025453.3025838

Koziol, N., & Bilder, C. R. (2014a). MRCV: Methods for analyzing multiple response categorical variables (MRCVs) (Version 0.3-3). Retrieved February 13, 2019 from https://CRAN.R-project.org/package=MRCV

Koziol, N., & Bilder, C.R. (2014b). MRCV: A package for analyzing categorical variables with multiple response options. *The R Journal, 6*(1), 144. https://doi.org/10.32614/RJ-2014-014

Kratz, J. E., & Strasser, C. (2015). Making data count. *Scientific Data, 2*(1), 150039. https://doi.org/10.1038/sdata.2015.39

Leonelli, S. (2015). What counts as scientific data? A relational framework. *Philosophy of Science, 82*(5), 810–821. https://doi.org/10.1086/684083

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

Leonelli, S., & Ankeny, R. A. (2015). Repertoires: How to transform a project into a research community. *BioScience, 65*(7), 701–708. https://doi.org/10.1093/biosci/biv061

Levallois, C., Steinmetz, S., & Wouters, P. (2013). Sloppy data floods or precise social science methodologies? Dilemmas in the transition to data-intensive research in sociology and economics. In P.

Wouters, A. Beaulieu, A. Scharnhorst, & S. Wyatt (Eds.) *Virtual knowledge: Experimenting in the humanities and the social sciences* (pp. 151–182). MIT Press.

Levin, N., & Leonelli, S. (2017). How does one "open" science? Questions of value in biological research. *Science, Technology, & Human Values*, 42(2), 280–305. https://doi.org/10.1177/0162243916672071

Li, X., Schijvenaars, B. J. A., & de Rijke, M. (2017). Investigating queries and search failures in academic search. *Information Processing & Management*, 53(3), 666–683. https://doi.org/10.1016/j.ipm.2017.01.005

Mayernik, M. S. (2016). Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, 67, 973–993. https://doi.org/10.1002/asi.23425

Mayernik, M. S., & Acker, A. (2017). Tracing the traces: The critical role of metadata within networked communications [Opinion paper]. *Journal of the Association for Information Science and Technology*, 69, 177–180. https://doi.org/10.1002/asi.23927

McDonald, J. H. (2014). Multiple Comparisons. In McDonald, J.H. (Ed.) *Handbook of biological statistics, 3rd ed.* (pp. 254-260). Sparky House Publishing. http://www.biostathandbook.com/multiplecomparisons.html

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, 106, 213–228. https://doi.org/10.1007/s11192-015-1765-5

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data Guiding Principles for the European Open Science Cloud. *Information Services & Use*, 37(1), 49–56. https://doi.org/10.3233/ISU-170824

Murillo, A. P. (2015). Examining data sharing and data reuse in the DataONE environment. *Proceedings of the ASIST Annual Meeting*, 51(1). https://doi.org/10.1002/meet.2014.14505101155

Mutschke, P., Mayr, P., & Scharnhorst, A. (2014). Editorial for the Proceedings of the Workshop Knowledge Maps and Information Retrieval (KMIR2014) at Digital Libraries 2014. In Mutschke, P., Mayr, P., & Scharnhorst, A. (Eds.) *KMIR 2014–Knowledge Maps and Information Retrieval: Proceedings of the First Workshop on Knowledge Maps and Information Retrieval* (Vol. 1311, pp. 1–6). CEUR-WS.org. http://ceur-ws.org/Vol-1311/

National Science Board. (2005). *Long-lived digital data collections: Enabling research and education in the 21st century* (No. NSB0540). National Science Foundation. https://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

Noorman, M., Wessel, B., Sveinsdottir, T., & Wyatt, S. (2018). Understanding the "open" in making research data open: Policy rhetoric and research practice. In A. Rudinow Sætnan, I. Schneider, & N. Green (Eds.), *The Politics and policies of big data: Big data Big Brother* (pp. 292–317). Routledge.

Noy, N., Burgess, M., & Brickley, D. (2019). Google dataset search: Building a search engine for datasets in an open Web ecosystem. *The World Wide Web Conference on–WWW '19*, 1365–1375. https://doi.org/10.1145/3308558.3313685

Park, H., & Wolfram, D. (2017). An examination of research data sharing and re-use: Implications for data citation practice. *Scientometrics*, *111*(1), 443–461. https://doi.org/10.1007/s11192-017-2240-2

Park, H., You, S., & Wolfram, D. (2018). Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. *Journal of the Association for Information Science and Technology*, *69*(11), 1346–1354. https://doi.org/10.1002/asi.24049

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and reuses of scientific data: The data creators' advantage. *Harvard Data Science Review, 1*(2). https://doi.org/10.1162/99608f92.fc14bf2d

Pigott, T. D. (2001). A review of methods for missing data. *Educational research and evaluation*, *7*(4), 353–383. https://doi.org/10.1076/edre.7.4.353.8937

Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2016). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*, *67*(12), 2964–2975. https://doi.org/10.1002/asi.23529

Scerri, A., Kuriakose, J., Deshmane, A. A., Stanger, M., Cotroneo, P., Moore, R., Naik, R., de Waard, A. (2017). Elsevier's approach to the bioCADDIE 2016 Dataset Retrieval Challenge. *Database*, 1–12. https://doi.org/10.1093/database/bax056

Schatzki, T., Knorr Cetina, K., & van Sevigny, E. (Eds.). (2001). *The practice turn in contemporary theory*. Routledge.

Schmidt, B., Gemeinholzer, B., & Treloar, A. (2016). Open data in global environmental research: The Belmont Forum's open data survey. *PLoS ONE*, *11*(1). https://doi.org/10.1371/journal.pone.0146695

Shove, E., Watson, M., & Spurling, N. (2015). Conceptualizing connections: Energy demand, infrastructures and social practices. *European Journal of Social Theory*, *18*(3), 274–287. https://doi.org/10.1177/1368431015579964

Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, *69*(1), 6–20. https://doi.org/10.1002/asi.23917

Tenopir, C., Christian, L., Allard, S., & Borycz, J. (2018). Research data sharing: Practices and attitudes of geophysicists. *Earth and Space Science*. https://doi.org/10.1029/2018EA000461

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, *6*(6), e21101–e21101. https://doi.org/10.1371/journal.pone.0021101

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLOS ONE*, *10*(8), e0134826. https://doi.org/10.1371/journal.pone.0134826

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, *27*(2), 237–246. https://doi.org/10.1177/1098214005283748

UK Data Service. (2019). Research data lifecycle. Retrieved August 8, 2019, from https://www.ukdataservice.ac.uk/manage-data/lifecycle.aspx

Ünal, Y., Chowdhury, G., Kurbanoğlu, S., Boustany, J., & Walton, G. (2019). Research data management and data sharing behaviour of university researchers. *Proceedings of ISIC*, *24(1)*, isic1818. http://www.informationr.net/ir/24-1/isic2018/isic1818.html

Van de Sandt, S., Dallmeier-Tiessen, S., Lavasa, A., & Petras, V. (2019). The definition of reuse. *Data Science Journal*, *18*(22), 1–19. https://doi.org/10.5334/dsj-2019-022

Van der Sande, M., Verborgh, R., Hochstenbach, P., & van de Sompel, H. (2018). Toward sustainable publishing and querying of distributed Linked Data archives. *Journal of Documentation, 74*(1), 195-222. https://doi.org/10.1108/JD-03-2017-0040

Vera-Baceta, M. A., Thelwall, M., & Kousha, K. (2019). Web of Science and Scopus language coverage. *Scientometrics*, *121*(3), 1803–1813. https://doi.org/10.1007/s11192-019-03264-z

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE*, *8*(7), e67332–e67332. https://doi.org/10.1371/journal.pone.0067332

Wang, C. L., & Saunders, M. N. (2012). Non-response in cross-cultural surveys: Reflections on telephone survey interviews with Chinese managers. In *West Meets East: Toward Methodological Exchange* (pp. 213–237). Emerald Group Publishing Limited.

Whitelaw, M. (2015). Generous interfaces for digital cultural collections. *DHQ: Digital Humanities Quarterly*, *9*(1), 1–16. http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Richard Finkers, R., Gonzalez-Beltran, A., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Wu, M., Psomopoulos, F., Khalsa, S. J., & Waard, A. de. (2019). Data discovery paradigms: User requirements and recommendations for data repositories. *Data Science Journal*, *18*(1), 3. https://doi.org/10.5334/dsj-2019-003

Wyatt, S. (2008). Technological determinism is dead; Long live technological determinism. In E. J. Hackett, O. Amsterdamska, M. Lynch, & J. Wajcman (Eds.), *The handbook of science & technology studies* (Vol. 7, pp. 165-181). MIT Press.

Yoon, A. (2014). "Making a square fit into a circle": Researchers' experiences reusing qualitative data. *Proceedings of the American Society for Information Science and Technology*, *51*(1), 1-4. https://doi.org/10.1002/meet.2014.14505101140

Yoon, A. (2017). Role of communication in data reuse. In S. Erdelez & N.K. Agarwal (Eds.), *Proceedings of the Association for Information Science and Technology* (pp. 463– 471.) Wiley. https://doi.org/10.1002/pra2.2017.14505401050

Zimmerman, A. S. (2003). *Data sharing and secondary use of scientific data: Experiences of ecologists* [Doctoral dissertation]. University of Michigan. https://deepblue.lib.umich.edu/handle/2027.42/61844

Zimmerman, A. S. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, *7*(1–2). https://doi.org/10.1007/s00799-007-0015-8

Zimmerman, A. S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science Technology and Human Values*, *33*(5), 631–652. https://doi.org/10.1177/0162243907306704

Footnotes

1. The survey instrument is available in the Supplementary Material for this article. ↩

2. This practice in particular confirms the importance of asking users what they actually do, rather than assuming an ideal notion of data search and reuse. ↩