Tracing data: A survey investigating disciplinary differences in data citation

Kathleen Gregory[1], Anton Ninkov[2], Chantal Ripp[3], Emma Roblin[4], Isabella Peters[5] & Stefanie Haustein[6]

[1]kathleen.gregory@uottawa.ca
University of Ottawa, School of Information Studies, Scholarly Communications Lab, 55 Laurier Avenue East, K1N 6N5, Ottawa, Ontario (Canada)
University of Vienna, Faculty of Computer Science, Research Group Visualization and Data Analysis, Währinger Straße 29, 1090, Vienna (Austria)
ORCID: 0000-0001-5475-8632

[2] anton.boudreau.ninkov@umontreal.ca
Université de Montréal, École de bibliothéconomie et des sciences de l'information, Pavillon 3200 rue Jean-Brillant, H3T 1N8, Montréal, Québec (Canada)
ORCID: 0000-0002-8276-7656

[3] chantal.ripp@uottawa.ca
University of Ottawa, Scholarly Communications Lab, 55 Laurier Avenue East, K1N 6N5,Ottawa, Ontario (Canada)
ORCID: 0000-0003-3544-8158

[4] eroblin@uottawa.ca
University of Ottawa, School of Information Studies, Scholarly Communications Lab, 55 Laurier Avenue East, K1N 6N5, Ottawa, Ontario (Canada)
ORCID: 0000-0002-9179-0620

[5]i.peters@zbw.eu
ZBW Leibniz Information Center for Economics & Kiel University, Duesternbrooker Weg 120, 24015 Kiel (Germany)
ORCID: 0000-0001-5840-0806

[6] stefanie.haustein@uottawa.ca
University of Ottawa, School of Information Studies, Scholarly Communications Lab, 55 Laurier Avenue East, K1N 6N5, Ottawa, Ontario (Canada)
Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal, CP 8888, Succ. Centre-Ville, H3C 3P8, Montréal, Québec (Canada)
ORCID: 0000-0003-0157-1430

Abstract

Data citations, or citations in reference lists to data, are increasingly seen as an important means to trace data reuse and incentivise data sharing. Although disciplinary differences in data citation practices have been well documented via scientometric approaches, we do not yet know how representative these practices are within disciplines. Nor do we yet have insight into researchers' motivations for citing - or not citing - data in their academic work. Here, we present the results of the largest known survey (n=2,492) to explicitly investigate data citation practices, preferences, and motivations, using a representative sample of academic authors by discipline, as represented in the Web of Science (WoS). We present findings about researchers' current practices and motivations for reusing and citing data and also examine their preferences for how they would like their own data to be cited. We conclude by discussing disciplinary patterns in two broad clusters, focusing on patterns in the social sciences and humanities, and consider the implications of our results for tracing and rewarding data sharing and reuse.

Keywords: data citation, research data, data reuse, data sharing, open science, survey

1. Introduction

Data sharing and reuse are pillars of open science. Sharing data can enable transparency in research, while reusing data created by other people offers the potential to validate existing findings and improve scientific efficiency (Baker, 2016; National Institutes of Health, 2023; Pasquetto et al., 2017). Although (open) *data objects,* such as databases, data collections and datasets, are reused (Federer, 2019), such reuse is often invisible or is not easy to trace (Lane et al., 2020; van de Sandt et al., 2019).

Data citations, i.e. citations in reference lists to data, are considered to be key to tracing data reuse and incentivising data sharing (Lowenberg et al., 2019). Despite numerous advocacy efforts to encourage and standardize data citation (Data Citation Synthesis Group, 2014; Make Data Count, n.d.), such citations are rare in the academic literature (Ninkov et al., 2021; Peters et al., 2016). If data reuse is acknowledged in publications, data are usually mentioned in a footnote or within the full text of publications (Park et al., 2018; van de Sandt, 2021).

Although much research has shown that the extent to which data are reused and cited differs across disciplines (Borgman, 2015; Borgman et al., 2021; Park & Wolfram, 2017; Robinson-García et al., 2016; van de Sandt, 2021), we do not know how wide-spread these practices are within broad disciplinary groups. We also do not have a good understanding about *why* people cite, or do not cite, data in their academic work (Mayernik, 2012; Silvello, 2018). This combination of a lack of information about broader disciplinary practices and citation motivations makes it difficult to place existing data citations in context and to develop meaningful ways of measuring and understanding data sharing and reuse.

This paper presents the results of the largest known survey (n=2,492) to explicitly investigate data citation practices, preferences and motivations, using a representative sample of academic authors by discipline, as represented in the Web of Science (WoS). We explore questions about researchers' current practices and motivations for reusing and citing data and also examine their preferences for how they would like their own data to be cited.

Past work examining data citation practices has taken scientometric approaches, relying either on bibliometric databases, such as the Data Citation Index (e.g. Park & Wolfram, 2017) or on broader corpuses containing data and other research outputs, such as DataCite (e.g. Ninkov et al., 2021). While helpful in painting a broad picture of the academic landscape, these studies are limited in terms of the data which are indexed in these bibliometric sources; inconsistent and often incomplete metadata (Robinson-Garcia et al., 2017); and an overall difficulty in detecting in-text references to data (Lane et al., 2020). Scientometric studies also cannot account for researchers' preferences and motivations.

This paper addresses this gap by surveying researchers directly. While our survey has produced a rich dataset (Ninkov et al., 2023), we focus here on how data citation and reuse practices vary by discipline, adding to and contextualizing past scientometric work. We present descriptive results of our sample as a whole and identify statistically significant differences according to authors' academic disciplines. We conclude by discussing patterns in two broad disciplinary clusters, focusing particularly on patterns in the social sciences and humanities, and consider the implications of our results for tracing and rewarding data sharing and reuse.

2. Background

Data citation is shaped by factors ranging from formalized citation standards and recommendations to disciplinary norms (Borgman, 2016). While data citations can act as a sign that data have been shared, citations also signify that data have been (re)used (Silvello, 2018); we therefore see data reuse as an important precursor to data citation.

2.1 Data reuse practices

Researchers reuse data in a variety of ways (Pasquetto et al., 2019) which are dependent on both individual practices as well as research norms and infrastructures at the disciplinary level (Kim & Yoon, 2017). Existing work proposes classifications and typologies to better understand data reuse and enable discussions across disciplines. Wynholds and colleagues (2012) develop the idea of *background* data uses, which support other research practices and *foreground* data uses, which drive new research (Wallis et al., 2013; Wynholds et al., 2012). More recent work classifies data reuse according to phases of academic work, i.e. project preparation, conducting research, or teaching (Gregory et al., 2020), or according to specific goal- or process-oriented data tasks, i.e. creating data summaries or seeking answers to questions (Koesten et al., 2017). Data reuse is increasingly conceptualized as a practice which exists on a spectrum (Gregory, 2021; Pasquetto et al., 2019). Pasquetto and colleagues (2019) in particular propose a continuum of data reuse spanning from more-frequent comparative activities to less-frequent integrative uses, i.e. bringing together data for new analysis or to identify new patterns.

Although it is commonly accepted that data citations indicate some type of use, not all uses of data will be captured in a publication or in a citation (Borgman, 2016; Federer, 2019). Using data in teaching, to calibrate instruments or to verify results, e.g., may not typically be recognized or cited in an academic publication (Gregory, 2021).

2.2. Practices of citing and mentioning data

The terminology surrounding data citation practices varies in the literature. Here, we differentiate between *data citations*, which we define as referring to data objects (i.e. datasets, repositories, collections of data) in reference lists, and *data mentions*, which refer to data objects throughout other sections of a publication, including in footnotes, supplementary material, figures and acknowledgements. Building on the work of van de Sandt et al. (2019), we further define *indirect data citations* as citations to publications related to data, i.e. to papers analyzing data or to data papers. These definitions are based primarily on the location of a reference. Unlike other proposed definitions, we do not use the terms *formal* or *informal* to differentiate between types of citations, as “formality” is defined differently across communities. We also differentiate between how these methods can be used to trace signs of data reuse. Table 1 summarizes these definitions and relates them to other terms used in the existing literature.

Term	Definition	Location in publication	Synonymous with…	Traceability
Data citations	Referring to data objects in reference lists Data citations may vary in their adherence to guidelines, formats, completeness, and use of persistent identifiers (PIDs).	In, e.g., bibliographies, works cited, reference lists Reference lists may be in traditional publications (i.e. journal articles, book chapters) or in other entities (i.e. blog posts, teaching syllabi).	Formal data citation, as in Park et al. (2018); citations with varying degrees of completeness, i.e. those which lack PIDs, as in Moss and Lyle (2018)	Currently directly traceable by citation indexes, automatic methods, human readers
Data mentions	Referring to data objects throughout a publication	Throughout sections of the body of a publication, footnotes, supplementary material, figures, acknowledgements, etc.	Informal data citation as in Park et al. (2018); intratextual citation, as in Mayo et al. (2016)	Difficult to be traced automatically, more traceable by humans
Indirect data citations	Referring to data by including a reference to other related publications (i.e. an article analyzing the data or a data paper)	In, e.g., bibliographies, works cited, reference lists	Indirect citations, as in van de Sandt et al. (2019)	Involves intermediate step of locating publication; may or may not lead to the data directly; camouflages sign of data reuse

Table 1. Definition and explanation of data citation terms.

Many studies of data citation practices employ scientometric approaches, i.e. analyzing bibliographies, entire publications, or bibliometric databases to detect mentions of data objects and traces of data reuse. Most scientometric studies draw on similar bibliometric sources, i.e. Clarivate's Data Citation Index (DCI), a database of data records from selected repositories with related citation information (Clarivate, 2022), or DataCite, a nonprofit organization providing persistent identifiers (PIDs) and services for research data and other research outputs (DataCite, n.d.). Other studies draw on data from curated data repositories, particularly the Inter-university Consortium for Political and Social Research (ICPSR) (e.g. Banaeefar et al., 2022; Lafia et al., 2022; van de Sandt, 2021). Studies across these sources document differences regarding data supplies and a variation in methods for both citing and mentioning data.

2.2.1. Data citations

In an analysis using the DCI, Robinson-Garcia et al. (2016) find that 88% of indexed data remained uncited. The majority of data objects with citations in the DCI are from repositories in crystallography and biomedicine, perhaps reflecting more established infrastructures and data sharing norms in these fields (Robinson-García et al., 2016). Other studies confirm the uncitedness of most data indexed in the DCI and document differences in broadly-defined disciplinary fields, particularly in the natural and life sciences (e.g. Peters et al., 2016). Park and Wolfram (2017) also confirm the greater number of data citations in the biomedical and physical sciences observed by Robinson-García (2016) and further suggest that self-citation and citation of co-authors is common.

In a study of DataCite, Robinson-Garcia et al. (2017) observe a variety of skewed distributions, i.e. where 2% of data centers account for 80% of data objects and a handful or repositories contain DOIs for related scientific publications. Such skewed distributions have also been observed in ocean science data in DataCite, where data reuse can be primarily attributed to data from a few organizations or by the data creators (Dudek et al., 2019).

A more recent analysis of DataCite documents an overall lack of citation relations between data and publications; approximately 1% of datasets in the corpus of nearly 8.5 million records contain citation information (Ninkov et al., 2021). The authors also identify a dearth of disciplinary metadata in the corpus; of the datasets which have both disciplinary and citation information, the majority come from the natural sciences, specifically earth and environmental or biological sciences.

Robinson-Garcia et al. (2016) also find that different disciplines cite different types of data objects. The majority of data citations in the DCI in the social sciences and humanities are to *data studies*, defined as studies and experiments with associated data, i.e. census data (Clarivate, 2022). Nearly all citations in engineering and technology and ‘science’ are to *data sets*, e.g. single data files. Even if data objects are cited in reference lists, data citations vary in their formats, completeness, use of persistent identifiers (PIDs) and overall adherence to citation guidelines (Mayo et al., 2016; Mooney & Newton, 2012; van de Sandt et al., 2019). PIDs are particularly important, as they provide a sustainable mechanism for identifying and locating individual data objects (Peters et al., 2016).

In 2012, data citations in the social sciences and humanities lacked PIDs, publisher information, and electronic retrieval location (Mooney & Newton, 2012). Roughly ten years later, data citations to social science data in ICPSR included many traditional metadata elements, i.e. title, author and publication date, although the majority still lacked PIDs (van de Sandt, 2021). Moss and Lyle (2018, 2019) further identify a spectrum of data citations and data mentions which do not include PIDs, i.e. citations which are ‘almost complete’ to data mentions which are ‘barely there,’ consisting primarily of a dataset’s title. In many cases, this lack of identifiers and other citation elements stands in opposition to explicitly stated data citation guidelines from both data repositories and journals (Mayo et al., 2016; van de Sandt, 2021), which are designed to facilitate long-term data identification.

2.2.2. Data mentions and indirect data citations

A lack of data citations does not mean that data are not acknowledged in publications. Researchers refer to data throughout sections of traditional academic papers, as well as in figures, tables and captions (Mooney & Newton, 2012; Park et al., 2018; Pepe et al., 2014). Such indications of data use may not be directly traceable by automated indexes (Table 1) but may rather remain hidden, camouflaged in data mentions or indirect data citations, particularly to publications in which data have been previously analyzed (van de Sandt et al., 2019). Other indirect data citations reference *data papers,* papers dedicated to describing data and their contexts of creation (Callaghan et al., 2012). While data papers are increasingly cited within scholarly communication, initial evidence suggests that the number of citations to data papers varies by discipline and may not indicate actual data reuse (Jiao & Darch, 2020).

Further disciplinary patterns of data mentions and indirect data citations have also been observed. In an analysis of genetics and heredity data in the DCI, Park et al. (2018) demonstrate a strong tendency in biomedical fields to mention data within the main text of articles; fewer data mentions occur in other areas of a publication, i.e. in acknowledgements or supplementary material. This pattern was also observed in a study of three openly available oceanographic datasets (Belter, 2014) and an analysis of life science data published in Dryad (Mayo et al., 2016).

Van de Sandt (2021) analyzed data and software citation in the social sciences, using data from ICPSR, and in high-energy physics (HEP), using data from CERN. Mentions to data from ICPSR occur most frequently in the methodology or in a dedicated “data” section of a publication and often consist of the data title and year but do not have other identifying or descriptive elements, i.e. the study acronym or version number. Data mentions in HEP are more heterogeneous and their exact location more difficult to classify, reflecting the variety of publication structures in the sub-disciplines of HEP (van de Sandt, 2021).

When discussing bibliometric studies of disciplinary differences, it is important to note the role of classification systems when interpreting the results. Each data source and methodological approach uses a particular disciplinary or subject classification, complicating comparisons. For example, datasets and data studies within the DCI receive the subject classification of the repository in which they are published (Force & Robinson, 2014), whereas DataCite relies heavily on disciplinary metadata provided by data repositories, which can then be enhanced or mapped to other classifications (Garza et al., 2021)

Repository-based analyses, i.e. at ICPSR or CERN, subsume many disciplinary sub-groups within broad categories, such as “social sciences” or “high-energy physics.” We also make use of broad disciplinary categories to facilitate comparison in this study, but we recognize that such comparisons are challenging and that examining disciplinary (data) practices at a high level can potentially obscure differences in sub-fields and research communities (Ninkov et al., 2022).

2.3 Motivations for citing and mentioning data

Motivations for citing academic literature have long been studied and theorized in scientometrics and related fields (see Bornmann and Daniel (2008) and Tahamtan and Bornmann (2019) for reviews). While citations can be used to acknowledge intellectual and cognitive influences (Merton, 1973, 1988) citation motivations and practices are also socially situated and constructed (Collins, 2004; Knorr-Cetina, 1981). Citations are therefore made for a variety of reasons, including persuasion (Gilbert, 1977); authority claims (Moed & Garfield, 2004); paying homage to pioneers and colleagues or correcting and criticizing earlier work (Garfield, 1965).

Although data citations and mentions are largely taken as a sign of data use, there is a paucity of empirical evidence and conceptual development about motivations for citing or mentioning data. Existing literature synthesizes arguments made by those working to encourage data citation, rather than examining actual citation motivations of researchers themselves. Such arguments focus on motivating researchers to cite data as a way to connect data and literature; to facilitate data discovery and reproducibility; to understand the use and impact of data; and to recognize and reward data management work (Mayernik, 2012; Silvello, 2018).

Work on data citation undertaken from the perspective of research infrastructures often focuses on the practical uses of citations and metrics to demonstrate the value of the infrastructures themselves (Mayernik et al., 2017). In this context, data citations may be made in order to persuade others of the quality of data used in a particular study or to credit and reward data providers (Mayernik et al., 2017). While it is debatable if these motivations provide direct incentives for researchers to cite data in reference lists (Mayernik, 2012), recent surveys show that the vast majority of respondents believe that data citations would provide an important credit mechanism for sharing research data (Tenopir et al., 2020; Digital Science et al., 2022)

This belief reflects the current academic reward system, where citations to scholarly literature are traditionally viewed as the primary currency, what Merton calls “pellets of peer recognition” (1988, p. 621). Literature citations and making data (openly) available have also been shown to be linked. Piwowar and colleagues were among the first to demonstrate a citation advantage for articles which have openly available data within cancer research (Piwowar et al., 2007) and genetics (Piwowar & Vision, 2013), findings corroborated in an analysis of papers with data availability statements in publications in PLOS and BiomedCentral (Colavizza et al., 2019). These findings suggest that an increase in literature citations could be a means of incentivizing researchers to share their data, and to cite the data of others. It remains unclear, however, if accruing additional literature citations is in fact a motivating factor for sharing, citing or mentioning data in practice.

Data citation motivations are not often explored through a disciplinary lens. In a move towards studies in this direction, Banaeefar et al. (2022) classify the context and types of citations to data in ICPSR. They report that data citations are typically made in order to refer to findings from another study; provide a brief data point as background information; and acknowledge the use of a survey instrument, experimental measure, or comparison of methodological approaches.

3. Methods and Data

Asking researchers directly about their practices and motivations can add additional context to the literature reviewed in Section 2. Surveys have been increasingly used as a way of measuring data sharing and reuse practices within disciplines. Tenopir and colleagues conducted a series of survey studies (Tenopir et al., 2011, 2015, 2018, 2020), documenting that perceptions of data sharing vary significantly by discipline. Schmidt et al. (2016) controlled for disciplinary differences using a two-sample comparison approach in their survey. Annual surveys about disciplinary data practices are also conducted by academic publishers and private companies, e.g. Digital Science, owner of figshare (Digital Science et al., 2020, 2021, 2022).

Unlike our approach, these surveys relied on convenience samples and did not aim for representativity according to academic disciplines. The majority also focus on data sharing and reuse, rather than explicitly investigating data citation.

3.1. Questionnaire

The questionnaire (Gregory, Ninkov, Peters, et al., 2022) was designed and scripted in SurveyMonkey. It employed a branching design with two primary branches: one for researchers who reuse data and one for those who do not, who we term non-reusers. Researchers reusing data were asked a maximum of 28 questions; non-reusers were asked up to 22 questions. The questionnaire consisted of three sections i) Reusing and Citing Data, where participants were asked about their practices, preferences regarding their own data; and their citation motivations; ii) Rewarding Data Management; and iii) Demographics. Questions were designed based on past research in data reuse (e.g. (Gregory et al., 2020; Pasquetto et al., 2019), data citation (e.g. (Robinson-García et al., 2016; Silvello, 2018; van de Sandt, 2021), citation motivations (Garfield, 1965; Mayernik, 2012; Mayernik, 2017) and academic reward (National Information Standards Organization, 2016). Question types included binary, multiple choice, 5-point Likert scale, multiple response, and open-ended questions; the exact number of each question type varied by survey branch. This paper reports the results from questions in the first section of the questionnaire, Reusing and Citing Data, and excludes open-ended questions from the analysis.

We improved the understandability and accuracy of the questionnaire in two rounds of review. We first distributed the questionnaire to experts in scientometrics, research data management, and survey research to test the content, phrasing and overall design. We then conducted a pilot study with a stratified random sample of 1,000 researchers using the recruitment and sampling methodology described in Section 3.2. The pilot study yielded a 1.2% response rate; responses from the pilot study are not reported in our results.

3.2. Sampling and recruitment

Our population of interest consisted of researchers across disciplines who have published a paper indexed in WoS between 2016 and 2020. We aimed to create a representative sample of this population according to disciplinary domain. To do this, we used a two-step approach, incorporating the subject classification of journals in which authors have published and researchers' own disciplinary identification, a process detailed in our earlier work (Gregory, Ninkov, Ripp, et al., 2022) and outlined below.

In the first step, we determined the percentage of researchers by discipline according to journal subject classification. We queried the Observatoire des Sciences et des Technologies (OST) local WoS database for articles published between 2016-2020. The retrieved articles had both an associated email address for the corresponding author and a journal-level subject classification assignment, according to the National Science Foundation (NSF) journal-level classification. The result of this query was 5.8 million unique email addresses associated with 8.2 million articles. To avoid under-representation of humanities researchers in the email distribution, we used the distribution of articles in subsequent steps.

To facilitate comparison with past work, we mapped the NSF classification scheme for retrieved articles to the OECD's revised Field of Science and Technology (FOS) classification (Ninkov, et al., 2022). The FOS schema, with six high-level categories and 42 sub-categories, provides a balance between breadth and specificity (OECD, 2007). Using this distribution, we determined the needed number of respondents from each discipline to achieve a confidence interval of 0.025 in our statistical analysis. We then randomly sampled unique emails accordingly. A total of 158,600 recruitment emails were sent between 18 January, 2022 and 4 March, 2022 via SurveyMonkey. One reminder email was sent after two weeks to encourage participation.

Classifying researchers via journal-level subject classifications can be problematic. Researchers may publish in journals in multiple fields, and journal-level classifications may not accurately reflect the subject of individual articles. Participants therefore also selected their own FOS sub-disciplines in the questionnaire. We mapped participants' selected sub-disciplines to the six main FOS disciplines and compared this to our desired sampling distributions, as responses were received. We used the participants' classification to determine if our desired disciplinary distributions had been met. We sent an additional round of 5,000 recruitment emails to researchers in medical and health to match our desired number of respondents. Data collection stopped once the desired minimum number of respondents in all fields were met. Table 2 summarizes the results of our sampling and mapping methodology.

OECD class	Desired percentage of sample	Minimum sample size	Responses	Percentage of actual sample
Natural Sciences	38.9%	597	1037	41.3%
Engineering and Technology	16.9%	259	319	12.7%
Medical and Health Sciences	29.3%	450	488	19.5%
Agricultural Sciences	3.3%	51	106	4.2%
Social Sciences	8.9%	137	463	18.5%
Humanities	2.7%	42	96	3.8%
Total	100.0%	1536	2509	100.0%

Table 2. Summary of sampling researchers by disciplinary classification

3.3. Survey response, data preparation, and data analysis

In total, we received 3,632 responses, 2,509 of which were complete, yielding a survey completion rate of 68.6%. Of those who did not complete the survey, 65.2% of non-reusers dropped out after the third question and 74.6% dropped out after the fourth question. 63.8% of re-users with incomplete responses stopped responding after the fourth question. Incomplete responses were excluded from this analysis. During data cleaning, we identified and removed 17 respondents whose responses had been incorrectly recorded in the survey system, potentially because participants used the browser back button. This yielded a total of

2,492 complete responses and an uncorrected response rate of 1.57%. Controlling for invalid emails, bounced emails and opt-outs (n=5,201) produced a response rate of 1.62%, similar to a survey using comparable recruitment methods (Gregory et al., 2020). We re-coded ordinal variables and multiple-choice responses to account for the branching design of the survey. Codes, variables and data cleaning steps are further explained in the data dictionary and documentation published with the anonymized survey data (Ninkov et al., 2023).

Data were analyzed using Excel and SPSS. Normality testing indicated the use of non-parametric tests for significance. Table 3 summarizes the statistical tests used for each question, namely Kruskal-Wallis H test along with Chi-squared test coupled with Cramer's V to measure the substantive significance. Questions which were the same from both branches were combined for analysis (e.g. questions 7 and 15 or 8 and 16); only questions reported in this paper are included in Table 3. To analyze multiple response questions, we treated each possible variable as a single question and performed the appropriate statistical test for each variable.

Test	**Question number**	**Reporting statistic**
Chi-squared Test and Cramer’s V	[1], [3], [7+15], [8+16], [9+17], [10], [11], [12], [13], [18], [20], [21]	Percentage selecting an option[1]
Kruskal-Wallis H Test	[4], [5], [6], [19]	Mean rank

Table 3. Statistical tests of significance used in the analysis

As seen in Table 2, we received relatively more responses in some disciplines than others, particularly in the social sciences. We therefore weighted the number of responses to match our desired distribution when reporting descriptive statistics for the entire population.

[1] Percentages are used in reporting to enhance readability. Chi-squared tests were conducted using observed counts. Observed counts as well as percentages are provided for questions with significant differences in Appendix A.

We report our results using visualizations in combination with descriptive and inferential statistics. To aid comparisons between disciplines, we begin each section of the findings with a figure visually summarizing results with significant differences between disciplines. We then provide figures summarizing our data at the level of the entire population in addition to narrative descriptions of overall trends and significant disciplinary differences. A synthesis figure with all statistically significant results is in the supplemental material (Appendix A of supplementary material).

3.6. Limitations

This study has limitations regarding a potential sampling bias, the questionnaire design, and our chosen analysis methods. While we used random sampling techniques to recruit a variety of researchers, respondents interested in the topics of data citation and reuse who are confident in their ability to complete an English-language survey would be more likely to respond. Our sample consists of researchers who have published in journals indexed in WoS, a database which has its own biases. Certain disciplinary domains are under-represented in WoS, e.g. the humanities, as are researchers from the Global South and those who do not publish in English (Mongeon & Paul-Hus, 2016; Petr et al., 2021; Sugimoto & Larivière, 2018). While these limitations are a source of sampling bias, drawing from this population also allowed us to target our desired population of researchers across domains. A further limitation in our analysis could be due to the lack of granularity in the FOS classification system which we use to report our results.

Responses indicate self-reported behaviors and attitudes, which could be affected by a desire to give socially acceptable answers. Responses were also influenced by the options to questions which we provided. To counter this, we designed our questions based on past research and provided open-ended response options for questions. Responses could also have been impacted by individual interpretations and the ordering of the questions. Our two-phase review of the questionnaire helped to address some of these limitations. Additionally, a list of definitions for terms was provided at the beginning of the questionnaire and was linked to on every page of the survey. Terms included *data reuse*, defined as ‘using data which others have created, for any purpose,’ and *data sharing*, defined as ‘making your data available to others, e.g. in a data repository.' The full list of terms is provided in the survey questionnaire (Gregory, Ninkov, Peters, et al., 2022).

3.7. Ethics and data availability

We received ethical approval from University of Ottawa for the study under number S-08-21-7283. The anonymized data from this survey are available under a CC-BY-4.0 license (Ninkov et al., 2023).

4. Findings

We begin by contextualizing our results with a description of the demographics of respondents and their reported data reuse practices. We then present our findings regarding data citation and mentioning practices; citation motivations; and respondents' preferences for their own data. To facilitate understanding our narrative results, we begin sections with tables summarizing statistically significant responses by discipline.

Reflecting our sampling and recruitment strategy, the majority of respondents are from the natural sciences and are in middle to senior career stages (Figure 1). Respondents primarily work in universities, followed by research institutions; most work in North America or Europe/Central Asia. Roughly two-thirds of respondents self-identify as men (66.2%) and one-third as women (31.5%)

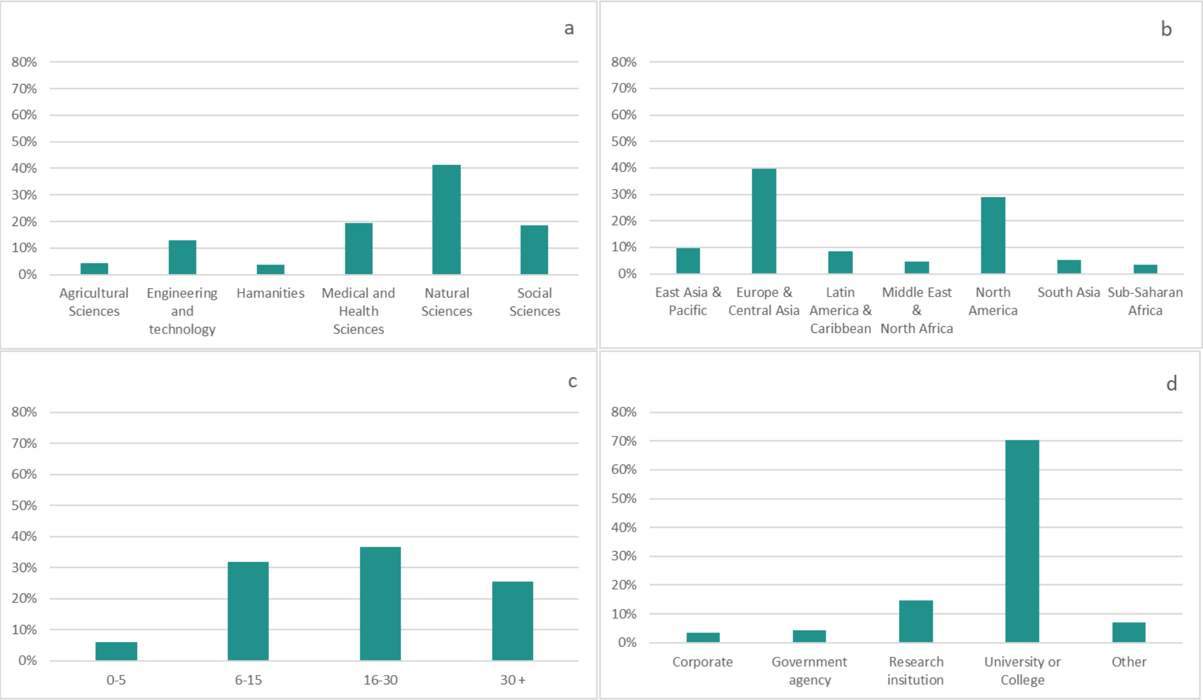


Figure 1. Percentage of respondents' a) disciplinary domains (n=2,492), b) geographic regions of employment, c) years of experience and d) employment institution. Percentages except for discipline have been weighted.

4.1 Data reuse practices

Figure 2 summarizes statistically significant results in questions related to data reuse practices.

Question	Nat. Sci.	Eng. & Tech.	Med. & Health Sci.	Agr. Sci.	Soc. Sci.	Hum.
1. Have you ever reused data which other people have created for any purpose (n=2,492)	84.4%	85.3%	76.7%	70.8%	78.5%	83.3%
4. How frequently do you reuse secondary data (n = 2,026):						
4.2. to prepare for a new project/proposal or to generate new ideas?	999.9	932.4	1,117.0	1,036.9	963.6	1,159.5
4.3. to integrate with other data?	1,092.3	937.6	936.8	947.7	967.5	1,042.8
4.4. as a model, algorithm, or system input?	1,107.6	1,219.8	941.7	938.9	793.9	686.8
4.5. to calibrate instruments or models?	1,071.8	1,257.9	932.3	1,029.9	835.5	715.3
4.6. to verify my own data?	1,058.8	1,187.0	965.3	1,114.6	794.7	1030.6
4.7. to identify trends, make comparisons or to make predictions?	1,024.6	1,081.2	984.1	998.6	949.6	1,100.9
4.8. to create visualizations or summaries?	1,051.5	960.2	1,046.9	963.8	964.2	896.8
13. What are your reasons for not reusing data created by other people (n = 466)?						
13.1. reusing data is not relevant to your research methods.	52.5%	51.1%	33.6%	37.7%	38.4%	81.3%
13.3. there are no available relevant data for your research.	25.6%	25.7%	16.8%	25.8%	37.4%	31.3%
13.6. you cannot find the data you need	12.5%	25.5%	9.7%	19.4%	26.3%	18.8%
13.7. you did not know that you could reuse data created by other people	7.5%	8.5%	18.6%	12.9%	6.1%	12.5%
20. Have you ever shared your own research data?	86.4%	82.1%	76.9%	75.5%	72.6%	75.0%
21. Have you ever reused your own research data?	71.4%	73.7%	68.6%	68.9%	81.1%	78.1%

Figure 2. Summary of statistically significant results by discipline for questions related to data reuse. Blue indicates a result greater than the average of the reporting statistic for each question; red indicates a result less than the average. Darker shades indicate larger deviation from the average. Reported mean ranks for Question 4 support the results of the Kruskal-Wallis tests.

The majority of respondents report reusing data (81.3%) and sharing their own data (81.0%). This indicates a potential self-selection bias in our sample towards people who share and reuse data. Roughly three quarters of respondents (71.9%) also reported reusing their own data multiple times. There is a significant difference but small association in data reuse according to academic discipline (X^2 (5, N=2,492)=27.18, p<.001, V=.104), with researchers in engineering and technology reusing data more and those in agricultural sciences less than expected, compared to other disciplines.

4.1.1. Types of data and types of data reuse

Across disciplines, there is a tendency for respondents to reuse *both* quantitative and qualitative data more than either data type alone (Figure 3). A significant difference with a medium association between disciplines was also detected (X^2 (15, N=2,026)=155.04, p<.001, V=.160), where social scientists reuse quantitative data more than expected, and researchers in the humanities use qualitative data more than expected.

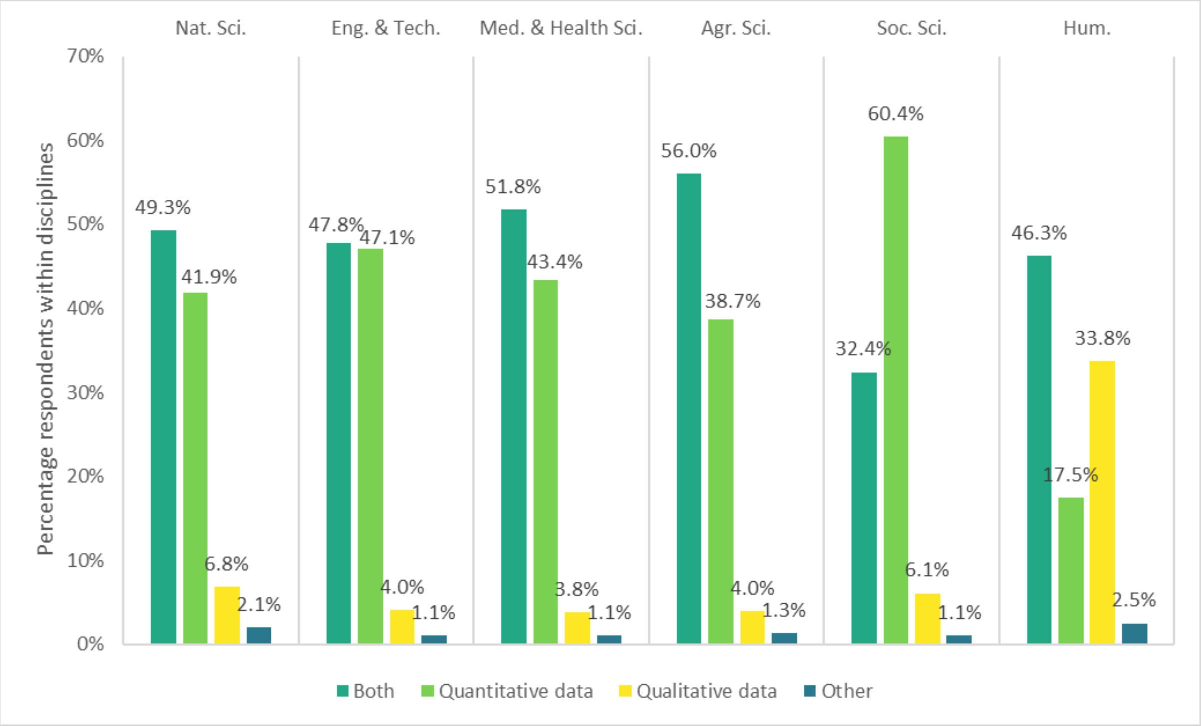


Figure 3. Types of data reused by participants, showing percentages of respondents in each discipline.

We also asked how frequently researchers reuse data for the different purposes proposed in Gregory et al. (2020). Significant differences to this question were detected (Figure 2, question 4), where researchers in engineering and technology more frequently reuse data as model, algorithm or system inputs; to calibrate instruments; or for verification purposes than do other disciplinary groups. Using data to identify trends and make comparisons or predictions is more frequently done by researchers in the humanities, natural sciences and engineering and technology. Researchers in the natural sciences and, to a lesser extent, the humanities more frequently integrate data to create new datasets than do researchers in other domains. These results suggest that data integration is influenced by a researcher's disciplinary domain and highlight that data integration is not something that every discipline engages in at the same frequency.

There was no significant difference between disciplines for two types of data reuse: using data as the basis for a new study (H(5)=7.115, p=.212) and using data in teaching

(H(5)=7.657, p=.176). This indicates that these types of data reuse are done with the same frequency levels (sometimes or often) across disciplines, which supports the preliminary findings of Gregory et al. (2020).

4.1.2. Non-reusers of research data

Roughly one fifth of survey respondents (n=466) do *not* reuse data in their work. We specifically asked these respondents to indicate their reasons for not reusing data (Figure 4). Across disciplines, the most frequently selected option was that reusing data was not relevant to respondents' research methods, although significant differences with a medium association between disciplinary groups for this option were identified (X^2 (5, N=466)=20.989, p<.001, V=.212). 81.3% of non-reusers in humanities state that reusing data is not relevant to their research methods; this was not a reason selected as often by researchers in the social sciences, agricultural sciences or medical and health.

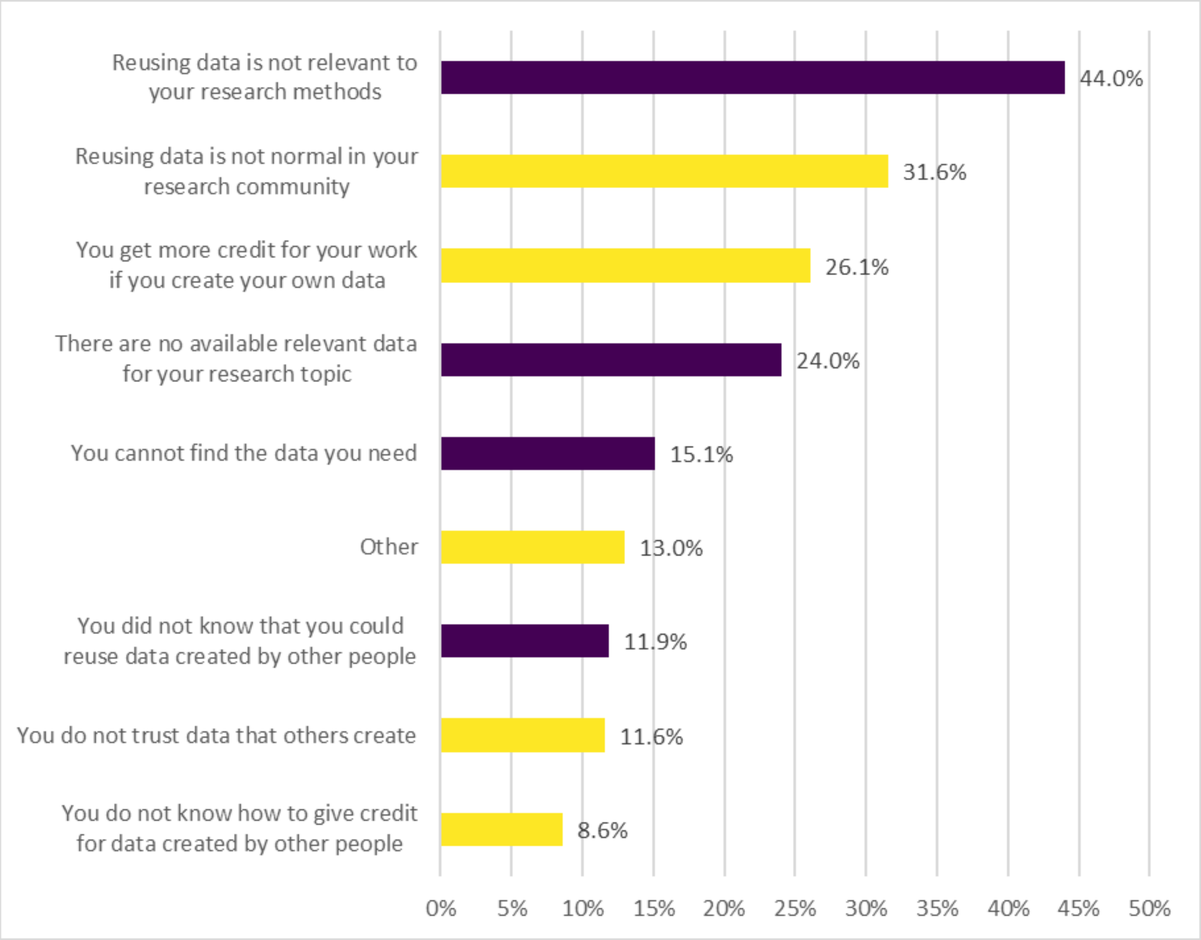


Figure 4. Reasons for ***not*** reusing data (multiple responses possible). Options with significant differences are indicated in purple. Percentages are based on weighted number of respondents answering this question (n=466).

Another significant difference between disciplines is tied to a lack of available relevant data for non-reusers. A medium association was detected for respondents in the social sciences, who selected this option more than other disciplines (X^2 (5, N=466)=11.768, p=.038, V=.159). Difficulties finding data are more of a barrier to reusing data in the social sciences and engineering and technology.

We did not detect significant disciplinary differences for many of the options to this question. One of the most common reasons (31.6%) to not reuse data across disciplines was that it is not normal practice in respondents' communities. Similarly, only slightly more than one quarter of respondents to this question indicated that they get more credit for creating their own data than for using other people's data. A lack of trust in data also does not appear to be a reason for many researchers not to reuse data, with only 11.6% selecting this option, nor does an awareness about how to credit the data of others.

4.2. Citing and mentioning data: practices, motivations and preferences

This section presents findings related to respondents' reported practices for citing and mentioning data; their data citation motivations; and their preferences for how others can acknowledge their own data.

4.2.1. Citing and mentioning practices

We asked respondents which data objects they refer to in publications, as well as to describe the methods with which they do so, i.e. by including a data citation, a mention in the footnote or body of text, or an indirect citation (Figure 5). We also asked respondents about their awareness and use of data citation standards. Significant differences for this question are reported in Figure 8.

Question	Nat. Sci.	Eng. & Tech.	Med. & Health Sci.	Agr. Sci.	Soc. Sci.	Hum.
5. When you reuse data, what do you usually cite or reference (n = 2,026):						
5.4. an article or publication analyzing the data?	1,063.0	1,044.1	998.8	1,104.5	862.5	1,037.1
5.5. a data paper?	1,053.2	1,048.6	1,000.7	981.4	928.4	937.0
6. When you reuse data, how do you cite or reference them? (n = 2,026)						
6.2. including a citation to a related paper in reference lists	998.5	1,026.7	1,049.5	1,009.8	980.8	1,118.0
6.3. mentioning data in a footnote	926.6	1,001.4	941.6	1,078.5	1,193.5	1,457.0
6.4. mentioning data in the body of text	996.2	969.9	1,008.0	933.2	1,119.0	973.7
6.6. mentioning data in captions, figures or tables	1,025.3	1,064.2	939.9	968.6	1,049.1	936.1

Figure 5. Summary of statistically significant results by discipline for questions related to data objects and citation/mentioning methods. Blue indicates a result greater than the average of the reporting statistic; red indicates a result less than the average. Darker shades indicate larger deviation from the average. Reported mean ranks for Question 6 support the results of the Kruskal-Wallis tests.

77.7% of data reusers indicated that they often or always cite or mention another publication in which the data have been analyzed (Figure 6). Respondents also frequently selected that they often or always refer to the source of the data (70.7%); referring to the data themselves was the third most frequently selected, with 58.3% of respondents across disciplines reporting that they either often or always cite or mention data.

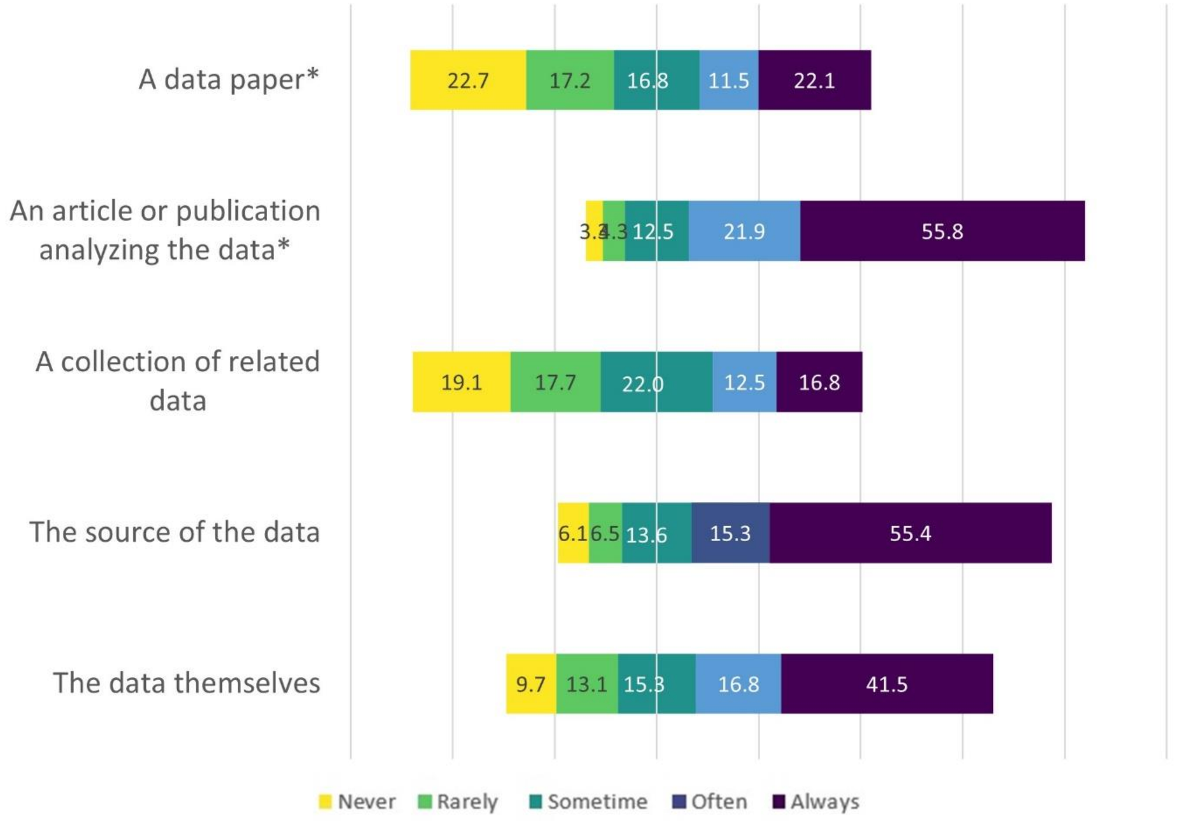


Figure 6. **Data objects.** Frequency of citing or mentioning various data objects. Options with significant differences are indicated with an asterisk. Percentages are based on weighted number of respondents per discipline answering this question (n=2,026). Bars are arranged around the middle (50% mark) of the ‘sometimes’ category.

Significant disciplinary differences were detected for how frequently respondents refer to two types of data objects: publications analyzing data and data papers. Social scientists cite or mention publications in which data have been previously analyzed less frequently than other disciplinary groups. Both social scientists and humanities researchers refer to data papers less frequently than other disciplines, particularly those in engineering and technology and natural sciences. No significant disciplinary differences were detected for how often respondents cite or mention the data themselves; referring to the data source is also a common practice for respondents across disciplinary groups.

Respondents indicated the frequency with which they employ various *methods* to refer to data (Figure 7). Across the sample, respondents report often including a citation to related papers, although a significant disciplinary difference was identified for this option (H(5)=61.877, p<.001). Researchers in engineering and technology more frequently cite or mention related papers, while social scientists engage in this practice the least, compared to other disciplinary groups. This tendency is supported by our previous finding regarding the types of data objects cited by social scientists (Figure 6).

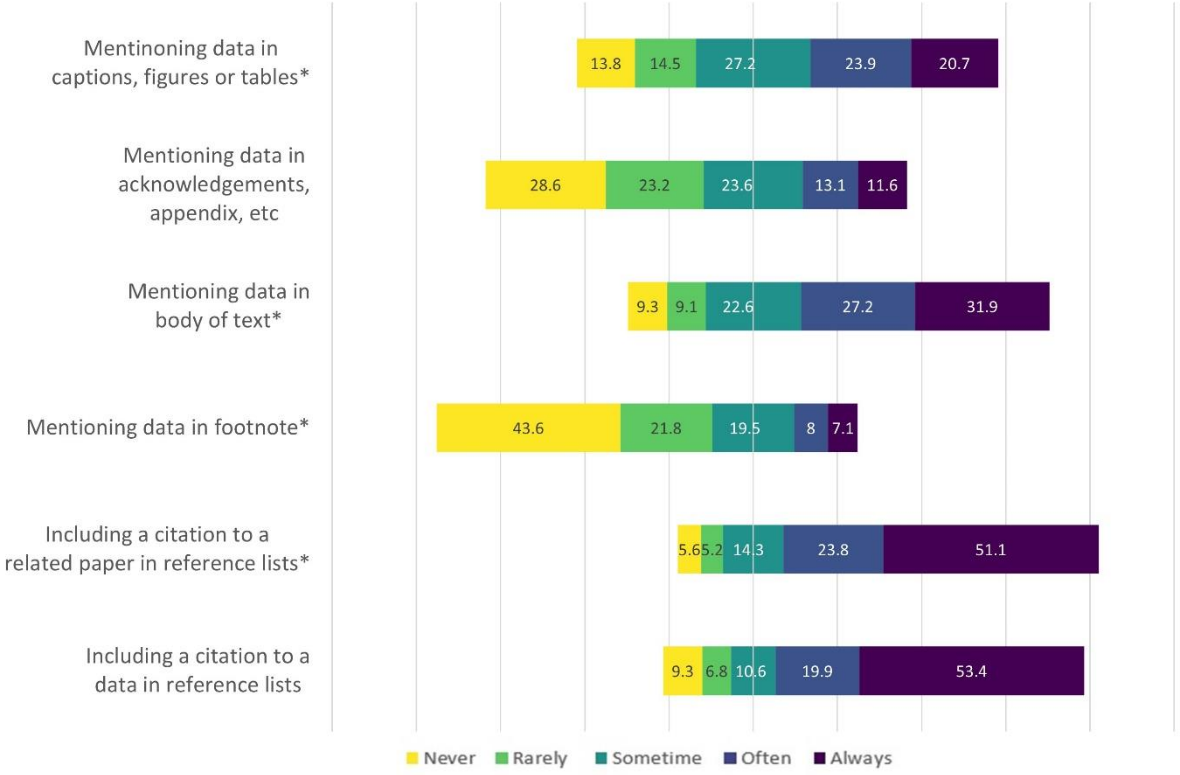


Figure 7. **Methods**. Frequency of citing or mentioning various data objects. Options with significant differences are indicated with an asterisk. Percentages are based on weighted number of respondents answering this question (n=2,026). Bars are arranged around the middle (50% mark) of the ‘sometimes’ category.

Another oft-reported practice is to include citations to data in reference lists (Figure 7). We did not detect a significant disciplinary difference for this option. This finding is also supported by the results of a separate question, in which 69.0% of data reusers across disciplines stated that they cite data in a reference list and 24.0% stated that they sometimes do so.

Significant differences were identified for referring to data in footnotes (H(5)=116.581, p<.001) and for referring to data in the body of a publication (H(5)=16.980, p=.005). Perhaps reflecting common practices of citing academic literature, humanities researchers more frequently refer to data using footnotes than other disciplines. Social scientists most frequently refer to data throughout the body of a publication, which supports the findings from van de Sandt (2021).

All respondents are generally unaware of and do not use many citation standards which have been developed specifically for data, i.e. those developed by DataCite or scientific societies (Figure 8). Respondents report being most aware of data citation standards created by journals and publishers or those included in long-standing citation guidelines, i.e. APA or MLA. If respondents are aware of guidelines, they tend to use them.

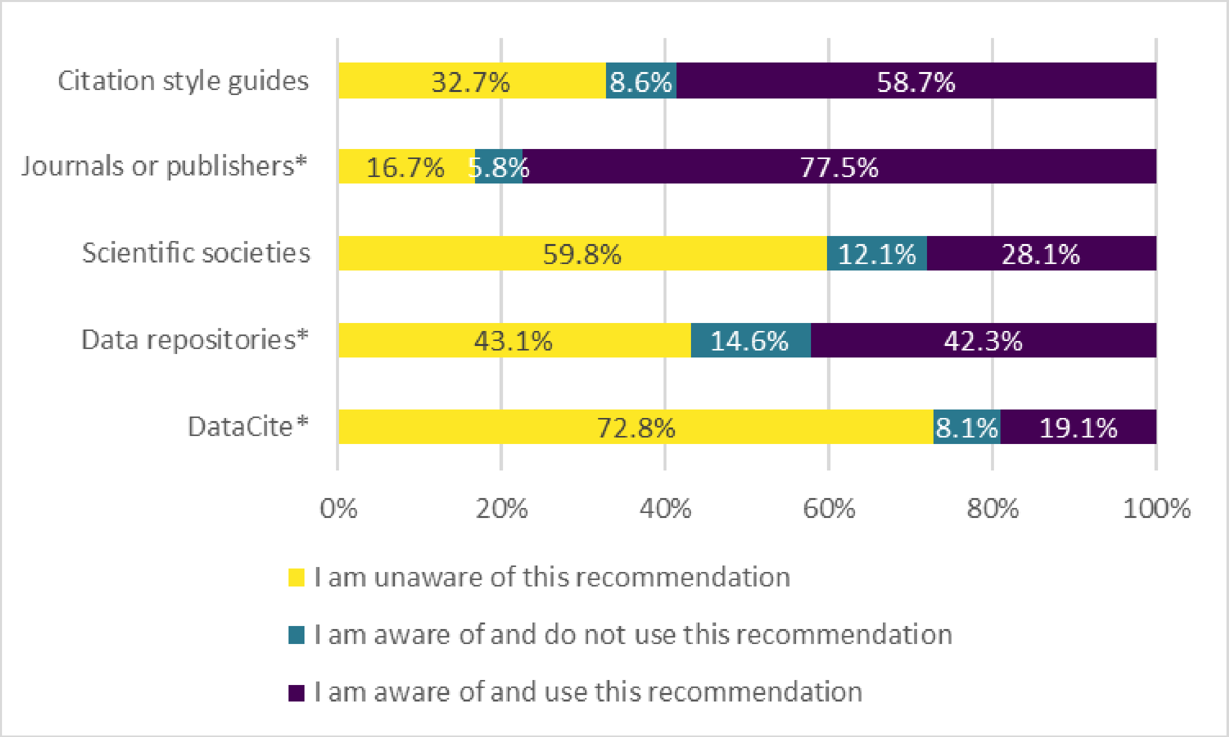


Figure 8. Awareness and use of data citation standards. Options with significant differences are indicated with an asterisk. Percentages are based on weighted number of all respondents (n= 2,492).

Significant disciplinary differences were identified for respondents' awareness and use of data citation standards. Social scientists, for example, were less aware of all citation standards, with the exception of standards from citation style guides, which they are aware of and use more than expected. Other disciplinary groups have greater awareness and use of other recommendations for data citation, particularly natural sciences and agricultural sciences, who are aware of and use recommendations issued by DataCite, repositories and scientific societies more than expected.

4.2.2. Motivations for citing data

We asked respondents who explicitly said that they cite data in a reference list about their motivations for doing so. Figure 9 summarizes statistically significant results for the relevant questions.

Question	Nat. Sci.	Eng. & Tech.	Med. & Health Sci.	Agr. Sci.	Soc. Sci.	Hum.
11. Why do you cite secondary data (n = 1,876)?						
11.1. as a way of showing intellectual debt to the data creator/data provider	78.9%	71.1%	74.2%	79.1%	86.0%	86.1%
11.3. as a way of helping others to locate and access the data you used	74.9%	69.5%	80.8%	68.7%	79.0%	81.9%
11.5. as a way of indicating that you have used the data in some way	66.0%	56.6%	71.9%	52.2%	64.6%	72.2%
12. Do you ever cite data for the following reason (n = 1,876)?						
12.1. to correct your own data (you cite your own data)	22.4%	23.0%	18.6%	22.4%	15.5%	33.3%

12.2. to build on or use data you have created (you cite your own data)	61.3%	52.7%	52.1%	53.7%	49.7%	61.1%
12.3. to criticize or correct the data of others	26.7%	26.2%	26.4%	17.9%	22.3%	56.9%
12.5. none of the above	27.5%	30.9%	33.8%	26.9%	39.9%	18.1%

Figure 9. Summary of statistically significant results by discipline for questions related to citation motivations. Blue indicates a result greater than the average of the reporting statistic; red indicates a result less than the average. Darker shades indicate larger deviation from the average.

Overall, motivations which reflect ideal scientific best practices, i.e. to show intellectual debt, to assist others in locating data, or to support the validity of research claims, were selected more frequently than external reasons (Figure 10). 8.4% of respondents to this question stated that they cite data because they were advised to, i.e. by journals or publishers.

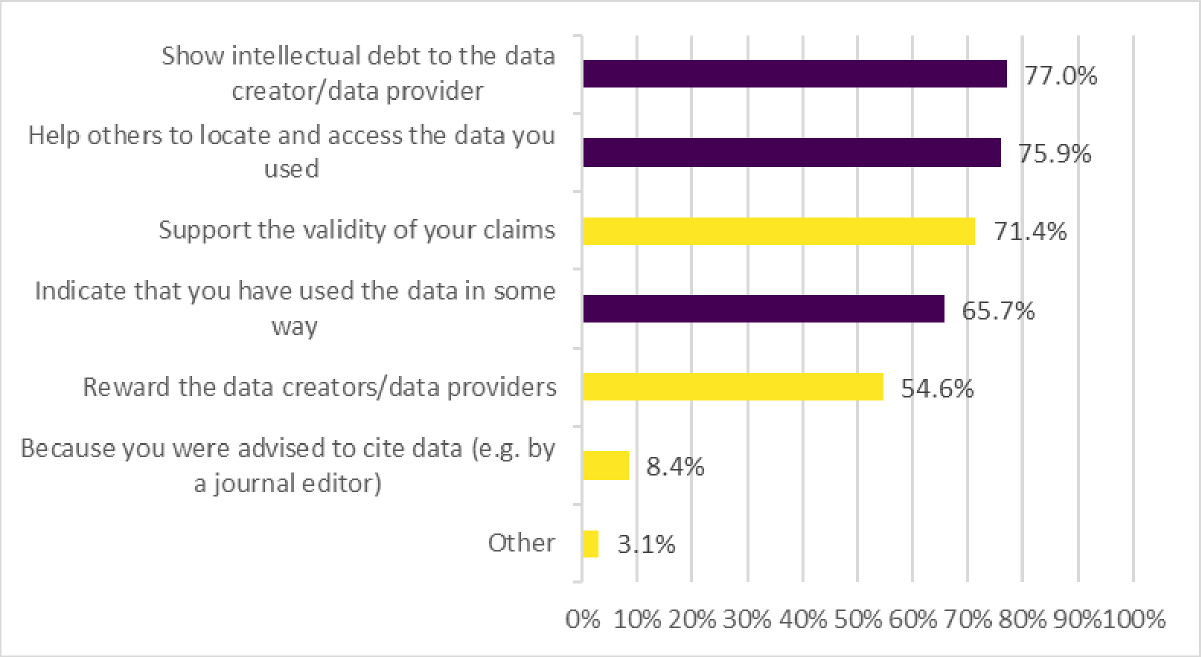


Figure 10. Motivations for citing data (multiple responses possible). Options with significant differences are indicated in purple. Percentages are based on weighted number of respondents answering this question (n=1,884).

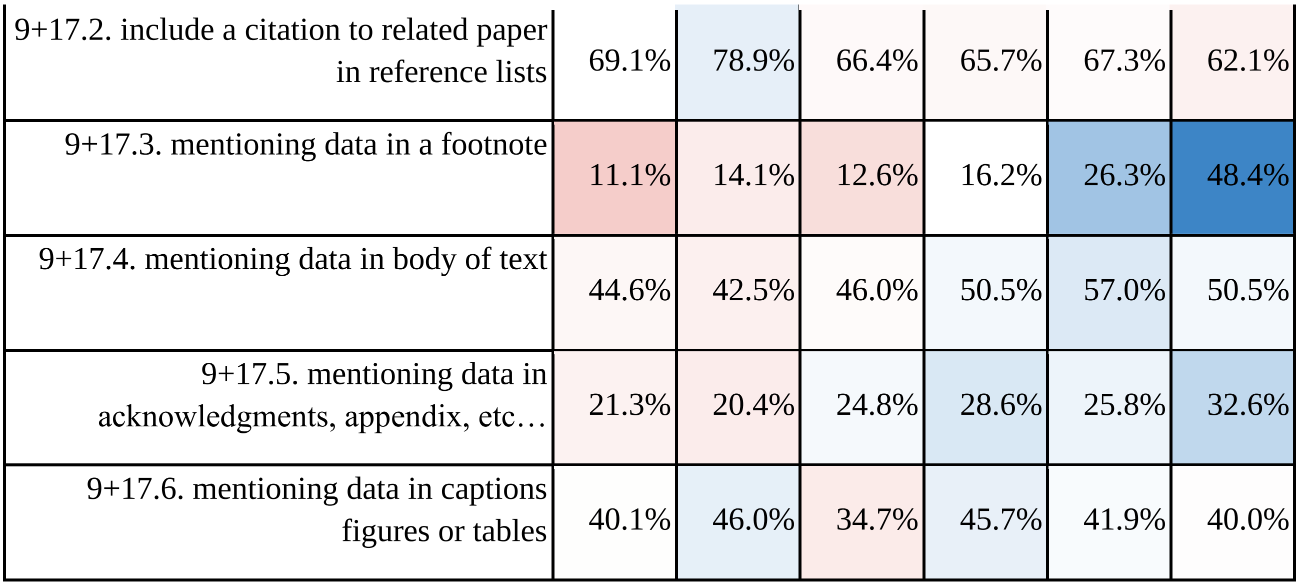
Significant disciplinary differences, although with small associations, were found for three motivations for data citation. Citing data as a way of demonstrating intellectual debt (X^2(5, 1,876)=25.497, p<.001, V=.117) was selected more frequently than expected by social scientists and humanities respondents**.** Facilitating data discovery (X^2(5, 1,876)=15.803, p=.007, V=.092) was selected more often than expected by researchers in the social sciences, medical and health sciences, and humanities. Using data citations to indicate data usage (X^2(5, 1,876)=22.062, p<.001, V=.108) was particularly important for researchers in the humanities and medical and health. No significant disciplinary difference was detected for respondents' who cite data to reward data providers; respondents across disciplines were roughly evenly split between those who selected this option and those who did not.

In a separate question, more than half of respondents across disciplines report citing their own data when they use data again. Respondents do not commonly cite data when criticizing or correcting the data of others (26.7%) or when correcting errors in their own data (21.2%). One notable exception to this is in the humanities, where respondents cite data in order to criticize the work of others much more than expected.

4.2.3. Preferences for respondents' own data

We asked all respondents a series of questions regarding their preferences for how they would like their own data to be cited or mentioned. Figure 11 summarizes statistically significant differences between disciplines for these questions.

Question	**Nat. Sci.**	**Eng. & Tech.**	**Med. & Health Sci.**	**Agr. Sci.**	**Soc. Sci.**	**Hum.**
8+16. What would you prefer that other people cite/reference when they use your data (n = 2,455)?						
8+16.1. the data themselves (e.g. a particular dataset or record)	46.0%	41.2%	46.0%	39.0%	58.5%	55.8%
9+17. How would you prefer other people to cite/reference your data (n = 2,455)?						



9+17.2. include a citation to related paper in reference lists	69.1%	78.9%	66.4%	65.7%	67.3%	62.1%
9+17.3. mentioning data in a footnote	11.1%	14.1%	12.6%	16.2%	26.3%	48.4%
9+17.4. mentioning data in body of text	44.6%	42.5%	46.0%	50.5%	57.0%	50.5%
9+17.5. mentioning data in acknowledgments, appendix, etc…	21.3%	20.4%	24.8%	28.6%	25.8%	32.6%
9+17.6. mentioning data in captions figures or tables	40.1%	46.0%	34.7%	45.7%	41.9%	40.0%

Figure 11. Summary of statistically significant results by discipline for questions related to citation preferences for respondents' own data. Blue indicates a result greater than the average of the reporting statistic; red indicates a result less than the average. Darker shades indicate larger deviation from the average.

The overwhelming majority of all respondents (98.5%) would like other people to refer to their data in some way. Mirroring the question design in Section 4.2.1, we asked respondents about their preferences for both types of *data objects* and referencing *methods*.

Across the sample, respondents prefer that others cite or mention a publication analyzing the data (84.3%) compared to other options, such as referring to the source of the data (55.3%) or the data themselves (46.3%) (Figure 12).

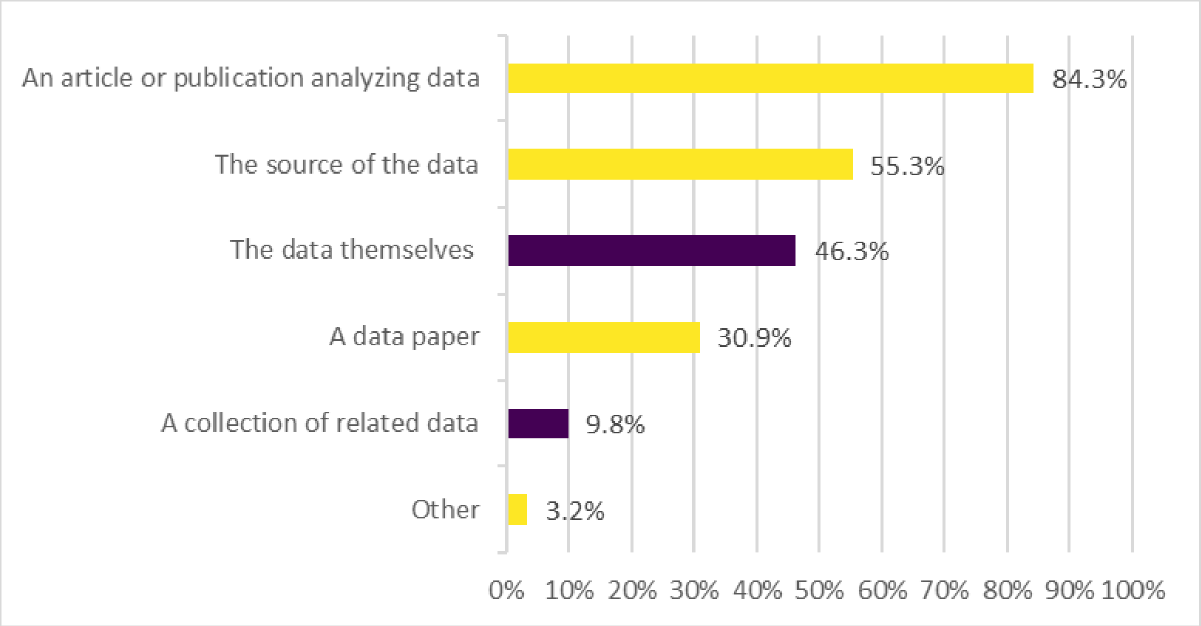
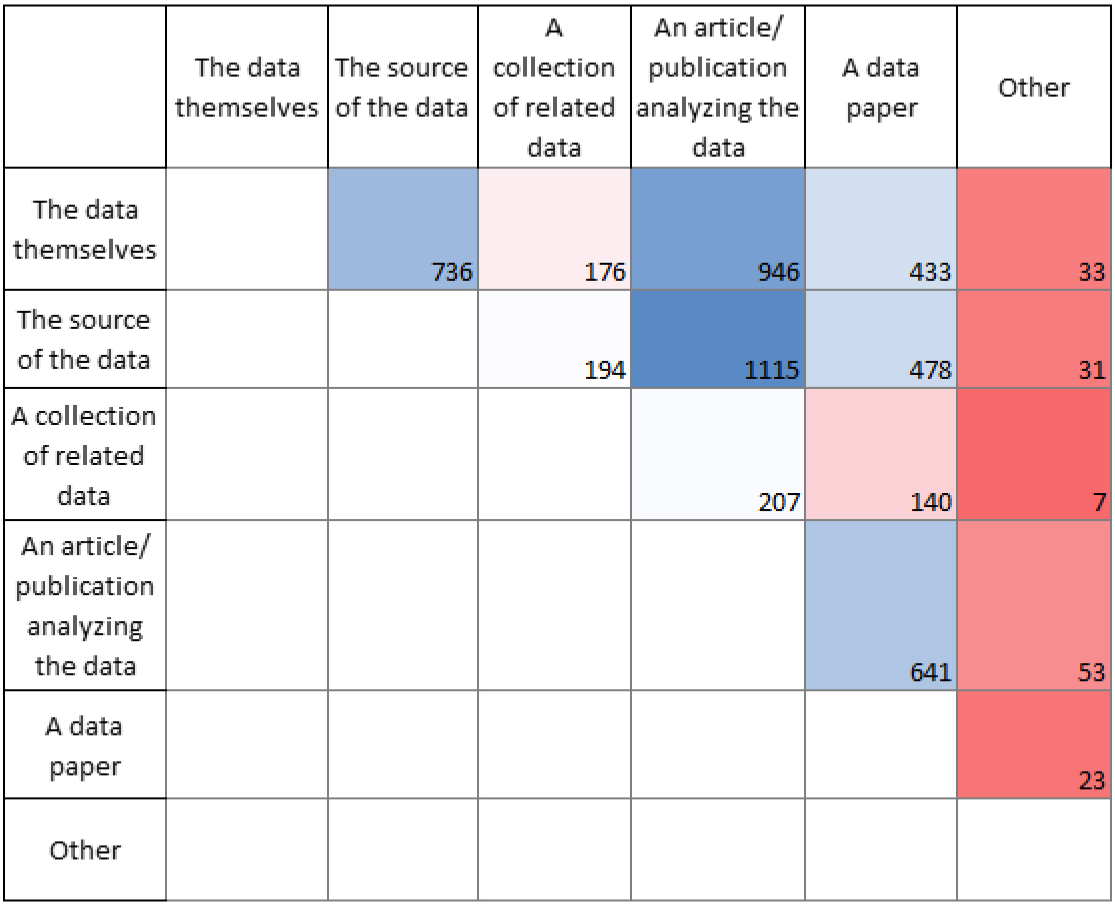


Figure 12. **Data objects.** Preferences for how respondents would like others to refer to their own data (multiple responses possible). Options with significant differences are indicated in purple. Percentages are based on weighted number of respondents (n=2,454).

Significant differences between disciplines were detected for the types of data objects which respondents prefer others to cite or mention (Figure 12). Social sciences and humanities are the only disciplines preferring that others cite or mention the data themselves more than expected**.** Respondents in all disciplines would like others to refer to a publication analyzing their own data; no significant difference was detected for this option.

72.5% of respondents chose more than one option for this question. Across disciplines, respondents frequently selected related publications and data sources together (Figure 13). In the medical and health, natural, and social sciences, related publications and data were also often chosen in conjunction. This suggests that respondents prefer that others cite multiple data objects to indicate the reuse of their data.



	The data themselves	The source of the data	A collection of related data	An article/ publication analyzing the data	A data paper	Other
The data themselves		736	176	946	433	33
The source of the data			194	1115	478	31
A collection of related data				207	140	7
An article/ publication analyzing the data					641	53
A data paper						23
Other						

Figure 13. **Multiple data objects**. Preferences for how respondents would like others to refer to their own data. Dark blue indicates objects most often selected together. Dark red indicates those least frequently selected together.

There is a preference among all respondents for others to include a citation of some sort in a reference list, be that a citation to the data themselves (71.3%) or to a related publication (69.5%) (Figure 14). This seems to stand in contrast to our findings about data objects. While respondents do not strongly prefer that others cite/mention the data themselves (Figure 12), they do want others to use a data citation (Figure 14). One explanation could be that respondents consider citations to other data objects, i.e. data sources, to constitute data citations. Findings in both Figure 12 and Figure 14 demonstrate that respondents across disciplines prefer others to cite related publications.

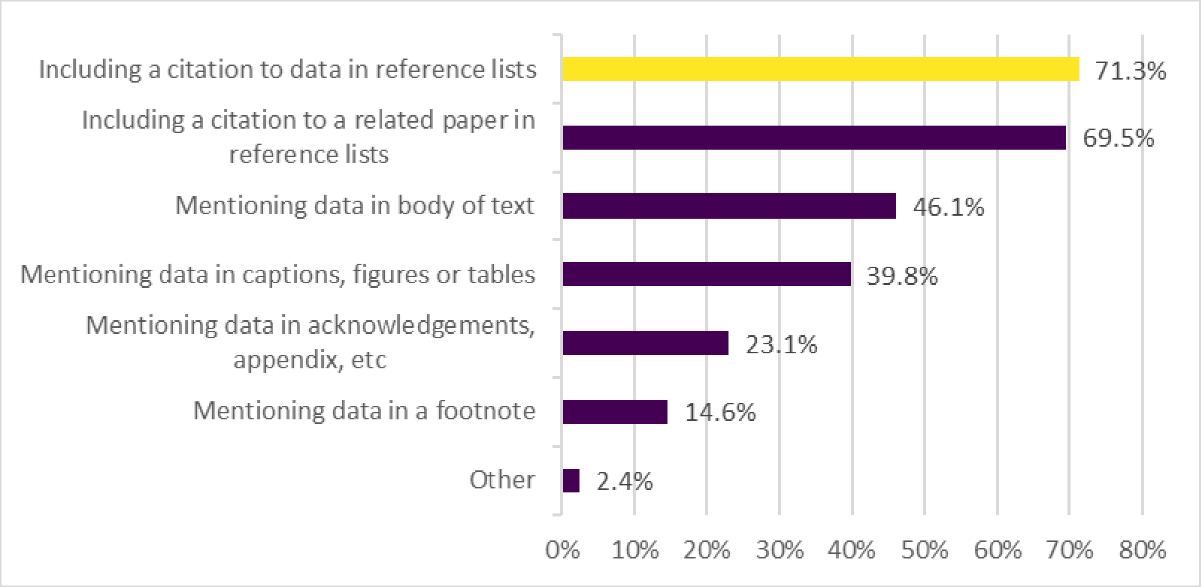


Figure 14. **Methods**. Preferences for how respondents would like others to refer to their own data (multiple responses possible). Options with significant differences are indicated in purple. Percentages are based on weighted number of respondents answering this question (n=2,454).

Significant disciplinary differences were identified for nearly every option to this question. Researchers in engineering and technology prefer data mentions in figures, captions and tables and to have indirect citations to related papers more frequently than expected. Researchers in the humanities do not. Humanities and social science respondents prefer the use of data mentions in footnotes (X^2(5, 2455)=131.739, p<.001, V=.232); this is the strongest association which we detected for this question. Social scientists also prefer that their data be mentioned in the body of publications more than other disciplinary groups.

5. Discussion

This paper presents findings from a survey explicitly investigating data reuse and citation practices using a carefully constructed, representative sample of researchers by discipline, as represented in WoS. We explored questions about the frequency of types of data reuse across disciplines and reasons why researchers do not reuse data. We examined researchers' reported practices and motivations for citing and mentioning data and also investigated respondents' preferences for how they would like their own data to be cited.

Although we found many disciplinary differences, our results particularly highlight differences in the social sciences and humanities (SSH). We therefore discuss our findings in two broad clusters, beginning with SSH researchers and then addressing other disciplinary groups.

5.1. Social Sciences and Humanities

5.1.1. Commonalities between SSH researchers

Social sciences and humanities researchers share some practices and preferences regarding data reuse and citation. Compared to other surveyed disciplines, SSH respondents are slightly more likely to reuse their own data than to share them with others. The reuse of one's own data or 'material' is common practice in the humanities and some areas of qualitative social sciences, where a particular object, corpus, or ethnographic study can be used as data throughout a researcher's career (Borgman, 2015). Our results also indicate that it may be common for social scientists to reuse their own quantitative data (Figure 2).

SSH are also the only disciplinary groups who prefer that others cite or mention their own data, as opposed to other data objects. This preference contrasts with the citation practices of SSH researchers documented in scientometric work, where SSH scholars cite 'data studies', rather than individual data files (Robinson-García et al., 2016), likely representing the varied ways in which researchers define data (Borgman, 2015; Leonelli, 2015).

Both disciplines also less frequently cite or mention data papers. This reflects the slower emergence of data papers and journals in SSH (Candela et al., 2015) and possibly a history of using data from governmental sources, where data papers may not be as relevant. There is some evidence that the landscape of data papers in SSH may be changing, and that data papers may have an effect on metrics of associated papers and data (McGillivray et al., 2022).

Both disciplinary groups cite data using footnotes and prefer to have their own data mentioned in footnotes more than other disciplines, although this practice is stronger in the humanities. This reflects the long-standing practice of using footnotes as a way of referencing, particularly in the humanities (Hammarfelt, 2012; Ochsner et al., 2016), and the tendency among social scientists in our results to mention data throughout a publication, which has also been documented in previous studies (Moss & Lyle, 2019; van de Sandt, 2021).

Respondents across the sample indicated that they cite data in order to acknowledge intellectual debt; our results suggest that this is a particularly important motivation for SSH researchers. Referring to data as ‘intellectual building blocks’ may be a factor of the purposes for which SSH respondents reuse data, i.e. as the basis for a new study or to integrate (literature) sources to build an argument. Acknowledging intellectual debt via citation is an established motivation for citing literature (Garfield, 1965; Merton, 1973); it could be that when a researcher’s data *is* literature, as is the case in some areas of humanities research, literature and data citation motivations are also intertwined.

SSH researchers, as well as those in medical and health, cite data as a way to help others to locate and access data. While this may be a motivation for data citation, the actual practice of many social science researchers may impede this goal. Mentioning data throughout the body of a publication or using incomplete data references (Banaeefar et al., 2022) may hinder automated forms of data discovery, which rely on or recommend the use of data citations with PIDs in reference lists (Data Citation Synthesis Group, 2014; Wilkinson et al., 2016). Recent efforts exploring alternative methods for automatically linking and discovering data from within the body of publications (see Lane et al., 2020) are more in line with the practices of social science researchers.

5.1.2. Social Sciences: Unique practices

While our findings demonstrate similarities among SSH researchers, we also find differences between social scientists and those in the humanities. Social science researchers report most often reusing quantitative data. This supports the findings of Fear (2013), documenting the prevalence of reusing numerical or statistical data created through social research methods and harkens the longstanding debate about the reuse of qualitative data within the social sciences (Bishop & Kuula-Luumi, 2017; Curty, 2016). Compared to other respondents who do *not* reuse data, social scientists indicated that challenges with data discoverability and a lack of available relevant data on certain research topics may inhibit data reuse.

In contrast to the other disciplines, social scientists cite or mention publications in which data have been previously analyzed less frequently. Instead, social science researchers tend to cite data objects throughout a publication, as found in studies at ICPSR (Banaeefar et al., 2022; Moss & Lyle, 2019; van de Sandt, 2021) and prefer that others do this as well. Social scientists are also most aware of and use data citation standards issued by long-standing citation style guides, i.e. APA. This could indicate a tendency for researchers to use standards with which they are already familiar, or it could signal a conflation among respondents between data and literature citation standards, given the recency of APA data-specific guidelines (American Psychological Association, 2022)

5.1.3. Humanities: Unique practices

Although the majority of humanities respondents reuse both quantitative and qualitative data, humanities researchers reuse qualitative data much more than other disciplinary groups. We also see that in many cases, humanities respondents indicated doing the opposite of other disciplines, as also noted by Cannon et al. (2022), engaging in practices which may be rooted in specific research methodologies.

Along with the natural sciences, humanities researchers reuse data to integrate different data sources, identify trends, and make comparisons more frequently than other disciplinary groups. This could indicate the use of digital methods among humanities respondents, but it could also be a sign of a tradition of bringing together and comparing different sources, both digital and analog, to make research claims.

Scientometric studies have suggested that self-citation may be common in existing data citations (Park et al, 2017). More than any other disciplinary group, humanities researchers cite their own data in order to make corrections; along with natural sciences researchers, humanities scholars also cite their own data in order to build on their past work. Again, perhaps reflecting critical research and discourse methods, respondents from the humanities cite data in order to both criticize and correct the data of others more than other disciplines.

5.2. Agricultural Science, Natural Sciences and Engineering and Medical and Health Science

We have discussed many of the disciplinary differences identified in our results from the standpoint of SSH researchers. The practices, preferences and motivations of researchers in other disciplinary groups also share commonalities and have some differences. Agricultural science, natural sciences, engineering and technology and medical and health sciences are similar when it comes to the type of data they reuse, all reusing both quantitative and qualitative data. While the majority of all survey respondents report sharing their own data, those in natural sciences do so more compared to other disciplines, supporting scientometric work in this area (Ninkov et al., 2022; Robinson-Garcia et al., 2017).

Building on the results of earlier work (Gregory et al., 2020), we also see strong reflections of disciplinary methodologies in the *frequency* of reusing data. In addition to the differences discussed in Section 5.1.1, our results show that natural sciences and engineering and technology researchers most frequently reuse data to calibrate instruments, to verify their own data or as model, algorithm, or system inputs.

Across our sample, respondents report most frequently citing or mentioning an article analyzing the data, compared to other data objects (Figure 6). Researchers in agricultural sciences, natural sciences and engineering and technology engage in this practice more than other disciplines. Engineering and technology researchers also report citing data in figures, tables and graphs, a practice which mirrors how these researchers would prefer that other people refer to their own data. This reflects a link to how researchers discover data from the literature (Pepe et al., 2014), where they also draw data for reuse from figures or captions.

While we found disciplinary differences in motivations for citing data, the strength of the detected associations was small. Engineering and technology is often situated on the opposite side of the spectrum from humanities in terms of citation motivations. Engineering and technology researchers, as well as those in agricultural sciences, do not cite data as often to acknowledge intellectual debt, to help others locate data, or as a sign of data use. We hypothesize that these differences in motivations could be linked to different reasons for reusing data and to associated research methods. Common data uses in these disciplinary groups, i.e. as model, algorithm or system inputs; to calibrate instruments; or for verification purposes, may not be seen as meriting an acknowledgement of 'intellectual debt,' but may rather be so standard that they are seamlessly integrated into research workflows.

6. Conclusion: Considerations for tracing data reuse

This study sheds light on relationships between data citation and data reuse, while also providing insight into why researchers cite, or do not cite, data in their academic work. Our results contextualize the broader development of research data services and have implications for efforts to trace signals of data reuse, e.g. in the development of data metrics. We conclude by highlighting three points for consideration when tracing data.

Data 'citation' is varied and differently interpreted.

Our results show that respondents from all disciplines reuse data for various purposes in research and teaching. However, the survey also reveals that this reuse of data is reflected via a variety of mechanisms in publications, including data mentions and indirect citations to related literature. At the same time, the vast majority of data reusers responding to our survey state that they *cite* data in reference lists, suggesting that researchers may have different interpretations of what it means to 'cite' data, and that they may construe these different mechanisms as valid and appropriate forms of data citation. We also see signs that researchers prefer that others reference a combination of different *data objects* to indicate data reuse (Figure 13).

These types of variations in practice and preference contrast with efforts which have gained momentum in the scholarly infrastructure space, i.e. those of data repositories and organizations such as DataCite which encourage the standardized citation of data in one location - reference lists - and the use of PIDs for individual datasets. Relying solely on data citations to trace signs of data reuse potentially disadvantages researchers who are engaging in what they see as a best practice, particularly if such signals are incorporated into systems of academic recognition and reward.

Data citation is rooted in other practices. When do we meet researchers where they are?

As seen in our findings, citing and mentioning data are shaped by discipline-specific practices, standards and research cultures. These practices seem to be rooted in long-standing traditions of indicating use in certain ways, i.e. via footnotes, and of referring to certain objects, particularly academic publications. The power of disciplinary and academic norms, including those of reward systems based on literature citation metrics, may impede citing data in reference lists. As literature citations are the primary currency in academic reward systems, researchers may loath to cite data rather than publications. Researchers may also find that their current practices meet the needs and expectations of their disciplinary communities. At the same time, survey responses from individuals who do not reuse data suggest that data citations could help to counter some barriers to data reuse, i.e. in facilitating data discovery. This juxtaposition raises a central question. When should research practice be adapted to current technical requirements and recommendations for data citation, and when should requirements and recommendations be adapted to reflect actual practice? Addressing this question requires long-term engagement with research communities and disciplinary debate.

Acknowledging data reuse is complex.

Another key insight derived from our results is that acknowledging data reuse in academic work may be more complex than acknowledging the reuse of ideas, methods and other knowledge present in academic publications. Research data are extremely diverse and exist at different levels of granularity in different formats (Peters et al., 2017). While different formats for communicating scholarly knowledge have been developed (Priem, 2013), such knowledge is often transmitted via standardized formats (i.e., journal articles, book chapters, conference proceedings), perhaps facilitating more homogeneous methods of citation.

Our results suggest that researchers cite or mention data for reasons related to ideal good research practices and that they are not motivated by external recommendations, i.e. from journal publishers. We also see that if they are aware of such citation guidelines that they tend to use them. As suggested by Banaeefar et al. (2022), this could indicate that while researchers are willing to cite data, they are still developing norms about when it is appropriate to acknowledge data reuse via a citation and when it is not.

Taken together, these points for consideration highlight that data citation is complex, local to different disciplines and communities, and tied to existing research practices and systems of recognition and reward. While we have explored data citation practices, preferences and motivations across disciplines, there is still much work to be done. Future work is needed to examine data citation practices according to other characteristics, i.e. by academic career stage, as is conducting more in-depth qualitative studies. Additionally, it is important to consider how to advance and adapt the development of metrics, policies and recommendations which incorporate data citations, particularly those related to rewarding individual data sharing and reuse.

Acknowledgements

We are grateful to Anthony Artino, the data management experts at the Data Archiving and Networked Services (DANS), Daniela Lowenberg, Mario Malicki, and Felicity Tayler for their input into the questionnaire design. We are especially grateful to Yoon Soo Park for his detailed comments regarding the sampling and recruitment strategy.

Funding

The Alfred P. Sloan Foundation (Sloan Grant G-2020-12670) supported this work.

Competing interests statement

We have no competing interests to declare.

Data availability

The cleaned and anonymized survey data is published in the Zenodo repository under a CC-BY 4.0 license. A citation to the dataset (Ninkov et al., 2023) is listed in the reference list.

References

American Psychological Association. (2022). *Data set references*. https://apastyle.apa.org/style-grammar-guidelines/references/examples/data-set-references

Baker, M. (2016). *1,500 scientists lift the lid on reproducibility: Nature News & Comment*. https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Banaeefar, H., Burchart, S., Moss, E., & Palvolgyi-Polyak, E. (2022). *Best Practice May Not Be Enough: Variation in Data Citation Using DOIs*. ICPSR.

Belter, C. W. (2014). Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets. *PLoS ONE, 9*(3), e92590. https://doi.org/10.1371/journal.pone.0092590

Bishop, L., & Kuula-Luumi, A. (2017). Revisiting Qualitative Data Reuse: A Decade On. *SAGE Open*, *7*(1), 2158244016685136. https://doi.org/10.1177/2158244016685136

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

Borgman, C. L. (2016). Data citation as a bibliometric oxymoron. In C. R. Sugimoto (Ed.), *Theories of informetrics and scholarly communication* (pp. 93–116). De Gruyter. https://doi.org/10.1515/9783110308464-008

Borgman, C. L., Wofford, M. F., Golshan, M. S., & Darch, P. T. (2021). Collaborative qualitative research at scale: Reflections on 20 years of acquiring global data and making data global. *Journal of the Association for Information Science and Technology*, *72*(6), 667–682. https://doi.org/10.1002/asi.24439

Bornmann, L., & Daniel, H. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80.

https://doi.org/10.1108/00220410810844150

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P., Bowie, R., Leadbetter, A., Lowry, R., Moncoiffé, G., Harrison, K., Smith-Haddon, B., Weatherby, A., & Wright, D. (2012). Making data a first class scientific output: Data citation and publication by NERC's environmental data centres. *International Journal of Digital Curation*, *7*(1), 1. https://doi.org/10.2218/ijdc.v7i1.218

Candela, L., Castelli, D., Manghi, P., & Tani, A. (2015). Data journals: A survey: Data Journals: A Survey. *Journal of the Association for Information Science and Technology*, *66*(9), 1747–1762. https://doi.org/10.1002/asi.23358

Cannon, M., Grant, R., & McKellar, K. (2022). Understanding and Supporting Data Sharing in the Humanities. In *State of Open Data 2022* (pp. 24–27). Digital Science, Springer Nature, Figshare. https://doi.org/10.6084/m9.figshare.21276984.v2

Clarivate. (2022). *Data Citation Index*. Web of Science Group. https://clarivate.com/webofsciencegroup/solutions/webofscience-data-citation-index/

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2019). The citation advantage of linking publications to research data. *ArXiv:1907.02565 [Cs]*. http://arxiv.org/abs/1907.02565

Collins, H. (2004). *Gravity's shadow: The search for gravitational waves*. University of Chicago Press.

Curty, R. G. (2016). Factors influencing research data reuse in the social sciences: An exploratory study. *International Journal of Digital Curation*, *11*(1), 96–117. https://doi.org/10.2218/ijdc.v11i1.401

Data Citation Synthesis Group. (2014). *Joint Declaration of Data Citation Principles*.

Force11. https://doi.org/10.25490/A97F-EGYK

DataCite. (n.d.). *DataCite*. Retrieved September 30, 2022, from https://datacite.org/

Digital Science, Goodey, G., Hahnel, M., Zhou, Y., Jiang, L., Chandramouliswaran, I., Hafez, A., Paine, T., Gregurick, S., Simango, S., Palma Peña, J. M., Murray, H., Cannon, M., Grant, R., McKellar, K., & Day, L. (2022). *The State of Open Data 2022* [Report]. Digital Science. https://doi.org/10.6084/m9.figshare.21276984.v5

Digital Science, Hahnel, M., McIntosh, L. D., Hyndman, A., Baynes, G., Crosas, M., Nosek, B., Shearer, K., van Selm, M., Goodey, G., & Research, N. (2020). *The State of Open Data 2020* [Report]. Digital Science. https://doi.org/10.6084/m9.figshare.13227875.v2

Digital Science, Simons, N., Goodey, G., Hardeman, M., Clare, C., Gonzales, S., Strange, D., Smith, G., Kipnis, D., Iida, K., Miyairi, N., Tshetsha, V., Ramokgola, M., Makhera, P., & Barbour, G. (2021). *The State of Open Data 2021* [Report]. Digital Science. https://doi.org/10.6084/m9.figshare.17061347.v1

Dudek, J., Mongeon, P., Bergmans, J., Meijer, I., European Commission, & Directorate-General for Research and Innovation. (2019). *Which role does DataCite play in researchers' data sharing and data (re)use practices?.* http://publications.europa.eu/publication/manifestation_identifier/PUB_KI0119255ENN

Fear, K. M. (2013). *Measuring and Anticipating the Impact of Data Reuse.* [Thesis]. http://deepblue.lib.umich.edu/handle/2027.42/102481

Federer, L. M. (2019). *Who, what, when, where, and why? Quantifying and understanding biomedical data reuse*. https://drum.lib.umd.edu/handle/1903/21991

Force, M. M., & Robinson, N. J. (2014). Encouraging data citation and discovery with the

Data Citation Index. *Journal of Computer-Aided Molecular Design*, *28*(10), 1043–1048. https://doi.org/10.1007/s10822-014-9768-5

Garfield, E. (1965). *Can Citation Indexing be Automated?* 7. http://garfield.library.upenn.edu/essays/V1p084y1962-73.pdf

Garza, K., Strecker, D., Ninkov, A., Schabinger, R., & Gregory, K. (2021). *DFG to OECD subject classification Mapping*. https://doi.org/10.5281/zenodo.5176122

Gilbert, G. N. (1977). Referencing as Persuasion. *Social Studies of Science*, *7*(1), 113–122.

Gregory, K. (2021). *Findable and reusable?: Data discovery practices in research* [Maastricht University]. https://cris.maastrichtuniversity.nl/en/publications/findable-and-reusable-data-discovery-practices-in-research

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

Gregory, K., Ninkov, A., Peters, I., & Haustein, S. (2022). *Questionnaire: A survey on data citation and reuse practices*. https://doi.org/10.5281/zenodo.6505207

Gregory, K., Ninkov, A., Ripp, C., Peters, I., & Haustein, S. (2022). Surveying practices of data citation and reuse across disciplines. *Proceedings of the 26th International Conference on Science and Technology Indicators*. International Conference on Science and Technology Indicators, Granada, Spain. https://doi.org/10.5281/ZENODO.6951437

Hammarfelt, B. (2012). *Following the Footnotes: A Bibliometric Analysis of Citation Patterns in Literary Studies*. Skrifter utgivna av Institutionen för ABM vid Uppsala universitet 5.

Jiao, C., & Darch, P. T. (2020). The role of the data paper in scholarly communication.

Proceedings of the Association for Information Science and Technology, *57*(1), e316. https://doi.org/10.1002/pra2.316

Kim, Y., & Yoon, A. (2017). Scientists' Data Reuse Behaviors: A Multilevel Analysis. *Journal of the Association for Information Science and Technology*, *68*(12), 2709–2719. https://doi.org/10.1002/asi.23892

Knorr-Cetina, K. (1981). *The manufacture of knowledge: An essay on the constructivist and contextual nature of science*. Pergamon Press.

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

Lafia, S., Fan, L., Thomer, A., & Hemphill, L. (2022). *Subdivisions and Crossroads: Identifying Hidden Community Structures in a Data Archive's Citation Network* (arXiv:2205.08395). arXiv. https://doi.org/10.48550/arXiv.2205.08395

Lane, J. I., Mulvany, I., & Nathan, P. (Eds.). (2020). *Rich Search and Discovery for Research Datasets*. SAGE Publications, Ltd. https://study.sagepub.com/richcontext/student-resources/e-book-files/web-pdf

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

Lowenberg, D., Chodacki, J., Fenner, M., Kemp, J., & Jones, M. B. (2019). *Open Data Metrics: Lighting the Fire*. Zenodo. https://doi.org/10.5281/zenodo.3525349

Make Data Count. (n.d.). *Make Data Count*. Make Data Count. Retrieved December 8, 2022, from https://makedatacount.org/

Mayernik, M. S. (2012). Data citation initiatives and issues. *Bulletin of the American Society* *for Information Science and Technology*, *38*(5), 23–28. https://doi.org/10.1002/bult.2012.1720380508

Mayernik, M. S., Hart, D. L., Maull, K. E., & Weber, N. M. (2017). Assessing and tracing the outcomes and impact of research infrastructures. *Journal of the Association for Information Science and Technology*, *68*(6), 1341–1359. https://doi.org/10.1002/asi.23721

Mayo, C., Vision, T. J., & Hull, E. A. (2016). The location of the citation: Changing practices in how publications cite original data in the Dryad Digital Repository. *International Journal of Digital Curation*, *11*(1), 1. https://doi.org/10.2218/ijdc.v11i1.400

McGillivray, B., Marongiu, P., Pedrazzini, N., Ribary, M., Wigdorowitz, M., & Zordan, E. (2022). Deep Impact: A Study on the Impact of Data Papers and Datasets in the Humanities and Social Sciences. *Publications*, *10*(4), 4. https://doi.org/10.3390/publications10040039

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago Press.

Merton, R. K. (1988). The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property. *Isis*, *79*(4), 606–623. https://doi.org/10.1086/354848

Moed, H. F., & Garfield, E. (2004). In basic science the percentage of “authoritative” references decreases as bibliographies become shorter. *Scientometrics*, *60*(3), 295–303. https://doi.org/10.1023/B:SCIE.0000034375.39385.84

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Mooney, H., & Newton, M. P. (2012). The Anatomy of a Data Citation: Discovery, Reuse, and Credit. *Journal of Librarianship and Scholarly Communication*, *1*(1). https://doi.org/10.7710/2162-3309.1035

Moss, E., & Lyle, J. (2018). *Opaque data citation: Actual citation practice and its implication for tracking data use*. http://deepblue.lib.umich.edu/handle/2027.42/142393

Moss, E., & Lyle, J. (2019, May 29). *Informal Data Citation: The Impact on Tracking Shared Data Reuse*. IASSIST 2019: Data down under: Exploring "data firsts," Sydney, Australia. https://doi.org/10.5281/zenodo.3605607

National Information Standards Organization, N. (2016). *Outputs of the NISO Alternative Assessment Metrics Project*. National Information Standards Organization.

National Institutes of Health. (2023). *NOT-OD-21-013: Final NIH Policy for Data Management and Sharing*. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

Ninkov, A., Gregory, K., Jambor, M., Garza, K., Strecker, D., Schabinger, R., Peters, I., & Haustein, S. (2022, September 7). Mapping Metadata—Improving Dataset Discipline Classification. *Proceedings of the 26th International Conference on Science and Technology Indicators*. International Conference on Science and Technology Indicators, Granada, Spain. https://doi.org/10.5281/ZENODO.6948238

Ninkov, A., Gregory, K., Peters, I., & Haustein, S. (2021). *Datasets on DataCite—An Initial Bibliometric Investigation*. 837–842. https://doi.org/10.5281/ZENODO.4730857

Ninkov, A., Ripp, C., Gregory, K., Peters, I., & Haustein, S. (2023). *A dataset from a survey investigating disciplinary differences in data citation (Version v2)* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7853477

Ochsner, M., Hug, S. E., & Daniel, H.-D. (Eds.). (2016). *Research Assessment in the Humanities*. Springer International Publishing. https://doi.org/10.1007/978-3-319-29016-4

Organisation for Economic Cooperation and Development. (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding*. https://doi.org/10.2481/dsj.6.OD4

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

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the reuse of scientific data. *Data Science Journal*, *16*(8), 1–9. https://doi.org/10.5334/dsj-2017-008

Pepe, A., Goodman, A., Muench, A., Crosas, M., & Erdmann, C. (2014). How do astronomers share data? Reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among US astronomers. *PLoS ONE*, *9*(8), e104798. https://doi.org/10.1371/journal.pone.0104798

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2016). Research data explored: An extended analysis of citations and altmetrics. *Scientometrics*, *107*(2), 723–744. https://doi.org/10.1007/s11192-016-1887-4

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. I. (2017). Zenodo in the Spotlight of Traditional and New Metrics. *Frontiers in Research Metrics and Analytics*, *2*. https://www.frontiersin.org/articles/10.3389/frma.2017.00013

Petr, M., Engels, T. C. E., Kulczycki, E., Dušková, M., Guns, R., Sieberová, M., & Sivertsen, G. (2021). Journal article publishing in the social sciences and humanities: A comparison of Web of Science coverage for five European countries. *PLOS ONE*, *16*(4), e0249879. https://doi.org/10.1371/journal.pone.0249879

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE*, *2*(3), e308. https://doi.org/10.1371/journal.pone.0000308

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. https://doi.org/10.7717/peerj.175

Priem, J. (2013). Beyond the paper. *Nature*, *495*(7442), 7442. https://doi.org/10.1038/495437a

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

Robinson-Garcia, N., Mongeon, P., Jeng, W., & Costas, R. (2017). DataCite as a novel bibliometric source: Coverage, strengths and limitations. *Journal of Informetrics*, *11*(3), 841–854. https://doi.org/10.1016/j.joi.2017.07.003

Schmidt, B., Gemeinholzer, B., & Treloar, A. (2016). Open Data in Global Environmental Research: The Belmont Forum’s Open Data Survey. *PLOS ONE*, *11*(1), e0146695.

https://doi.org/10.1371/journal.pone.0146695

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

Sugimoto, C. R., & Larivière, V. (2018). *Measuring research: What everyone needs to know*. Oxford University Press.

Tahamtan, I., & Bornmann, L. (2019). What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics*, *121*(3), 1635–1684. https://doi.org/10.1007/s11192-019-03243-4

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. https://doi.org/10.1371/journal.pone.0021101

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., 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

Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., Grant, B., Olendorf, R., & Sandusky, R. J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLOS ONE*, *15*(3), e0229003. https://doi.org/10.1371/journal.pone.0229003

van de Sandt, S. (2021). *The Tracking of Research Data and Software (re)Use in Scholarly Works*. Humboldt-Universität zu Berlin.

van de Sandt, S., Nielsen, L. H., Ioannidis, A., Muench, A., Henneken, E., Accomazzi, A., Bigarella, C., Lopez, J. B. G., & Dallmeier-Tiessen, S. (2019). Practice meets Principle: Tracking Software and Data Citations to Zenodo DOIs. *ArXiv:1911.00295 [Cs]*. http://arxiv.org/abs/1911.00295

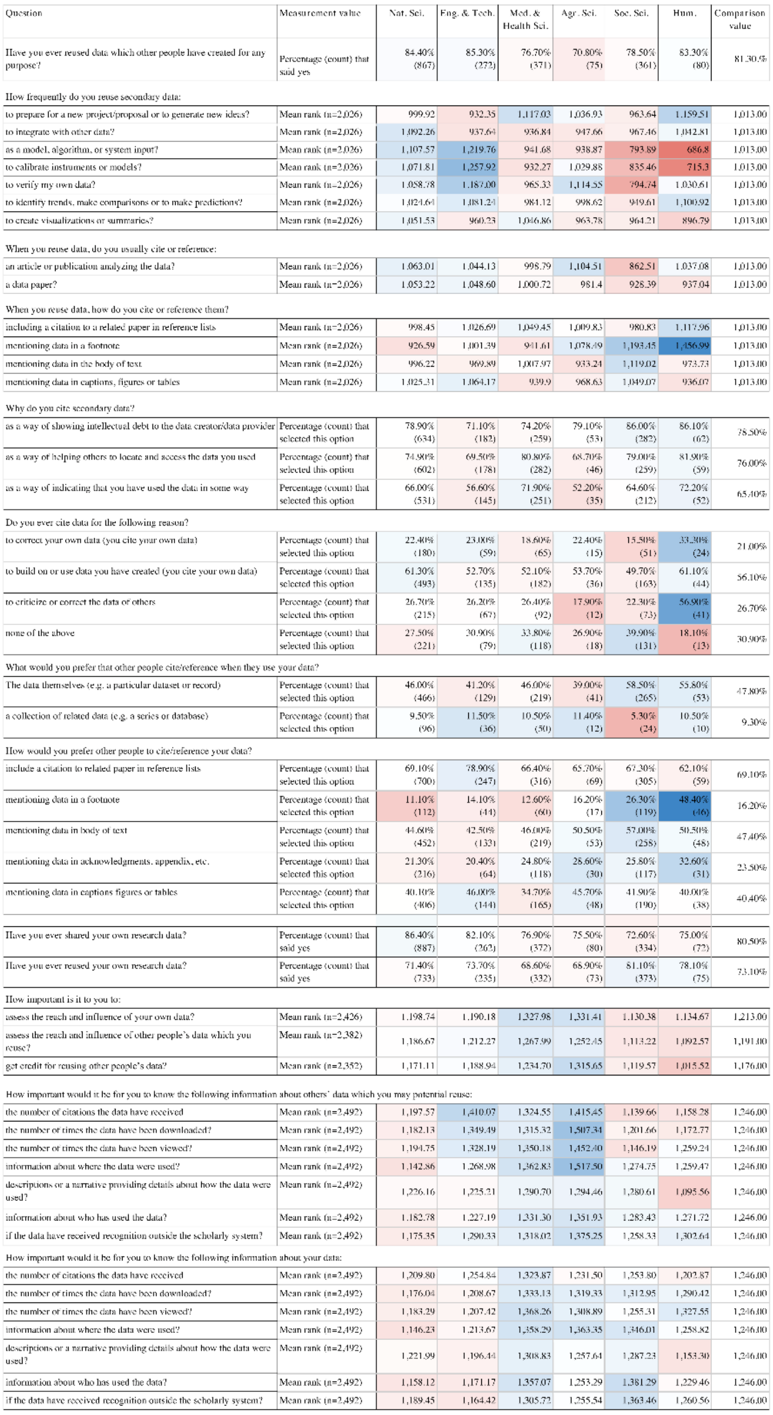
Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLoS ONE*, *8*(7), e67332. https://doi.org/10.1371/journal.pone.0067332

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., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wynholds, L. A., Wallis, J. C., Borgman, C. L., Sands, A., & Traweek, S. (2012). Data, data use, and scientific inquiry: Two case studies of data practices. *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, 19–22. https://doi.org/10.1145/2232817.2232822

Appendices

Appendix A. Full summary table with all significantly different results



Question	Measurement value	Nat. Sci.	Eng. & Tech.	Med. & Health Sci.	Agr. Sci.	Soc. Sci.	Hum.	Comparison value
Have you ever reused data which other people have created for any purpose?	Percentage (count) that said yes	84.40% (867)	85.30% (272)	76.70% (371)	70.80% (75)	78.50% (361)	83.30% (80)	81.30.%
How frequently do you reuse secondary data:								
to prepare for a new project/proposal or to generate new ideas?	Mean rank (n=2,026)	999.92	932.35	1,117.03	1,036.93	963.64	1,159.51	1,013.00
to integrate with other data?	Mean rank (n=2,026)	1,092.26	937.64	936.84	947.66	967.46	1,042.81	1,013.00
as a model, algorithm, or system input?	Mean rank (n=2,026)	1,107.57	1,219.76	941.68	938.87	793.89	686.8	1,013.00
to calibrate instruments or models?	Mean rank (n=2,026)	1,071.81	1,257.92	932.27	1,029.88	835.46	715.3	1,013.00
to verify my own data?	Mean rank (n=2,026)	1,058.78	1,187.00	965.33	1,114.55	794.74	1,030.61	1,013.00
to identify trends, make comparisons or to make predictions?	Mean rank (n=2,026)	1,024.64	1,081.24	981.12	998.62	949.61	1,100.92	1,013.00
to create visualizations or summaries?	Mean rank (n=2,026)	1,051.53	960.23	1,046.86	963.78	964.21	896.79	1,013.00
When you reuse data, do you usually cite or reference:								
an article or publication analyzing the data?	Mean rank (n=2,026)	1,063.01	1,044.13	998.79	1,104.51	862.51	1,037.08	1,013.00
a data paper?	Mean rank (n=2,026)	1,053.22	1,048.60	1,000.72	981.4	928.39	937.04	1,013.00
When you reuse data, how do you cite or reference them?								
including a citation to a related paper in reference lists	Mean rank (n=2,026)	998.45	1,026.69	1,049.45	1,009.83	980.83	1,117.96	1,013.00
mentioning data in a footnote	Mean rank (n=2,026)	926.59	1,001.39	941.61	1,078.49	1,193.45	1,456.99	1,013.00
mentioning data in the body of text	Mean rank (n=2,026)	996.22	969.89	1,007.97	933.24	1,119.02	973.73	1,013.00
mentioning data in captions, figures or tables	Mean rank (n=2,026)	1,025.31	1,064.17	939.9	968.63	1,049.07	936.07	1,013.00
Why do you cite secondary data?								
as a way of showing intellectual debt to the data creator/data provider	Percentage (count) that selected this option	78.90% (634)	71.10% (182)	74.20% (259)	79.10% (53)	86.00% (282)	86.10% (62)	78.50%
as a way of helping others to locate and access the data you used	Percentage (count) that selected this option	74.90% (602)	69.50% (178)	80.80% (282)	68.70% (46)	79.00% (259)	81.90% (59)	76.00%
as a way of indicating that you have used the data in some way	Percentage (count) that selected this option	66.00% (531)	56.60% (145)	71.90% (251)	52.20% (35)	64.60% (212)	72.20% (52)	65.40%
Do you ever cite data for the following reason?								
to correct your own data (you cite your own data)	Percentage (count) that selected this option	22.40% (180)	23.00% (59)	18.60% (65)	22.40% (15)	15.50% (51)	33.30% (24)	21.00%
to build on or use data you have created (you cite your own data)	Percentage (count) that selected this option	61.30% (493)	52.70% (135)	52.10% (182)	53.70% (36)	49.70% (163)	61.10% (44)	56.10%
to criticize or correct the data of others	Percentage (count) that selected this option	26.70% (215)	26.20% (67)	26.40% (92)	17.90% (12)	22.30% (73)	56.90% (41)	26.70%
none of the above	Percentage (count) that selected this option	27.50% (221)	30.90% (79)	33.80% (118)	26.90% (18)	39.90% (131)	18.10% (13)	30.90%
What would you prefer that other people cite/reference when they use your data?								
The data themselves (e.g. a particular dataset or record)	Percentage (count) that selected this option	46.00% (466)	41.20% (129)	46.00% (219)	39.00% (41)	58.50% (265)	55.80% (53)	47.80%
a collection of related data (e.g. a series or database)	Percentage (count) that selected this option	9.50% (96)	11.50% (36)	10.50% (50)	11.40% (12)	5.30% (24)	10.50% (10)	9.30%
How would you prefer other people to cite/reference your data?								
include a citation to related paper in reference lists	Percentage (count) that selected this option	69.10% (700)	78.90% (247)	66.40% (316)	65.70% (69)	67.30% (305)	62.10% (59)	69.10%
mentioning data in a footnote	Percentage (count) that selected this option	11.10% (112)	14.10% (44)	12.60% (60)	16.20% (17)	26.30% (119)	48.40% (46)	16.20%
mentioning data in body of text	Percentage (count) that selected this option	44.60% (452)	42.50% (133)	46.00% (219)	50.50% (53)	57.00% (258)	50.50% (48)	47.40%
mentioning data in acknowledgments, appendix, etc.	Percentage (count) that selected this option	21.30% (216)	20.40% (64)	24.80% (118)	28.60% (30)	25.80% (117)	32.60% (31)	23.50%
mentioning data in captions figures or tables	Percentage (count) that selected this option	40.10% (406)	46.00% (144)	34.70% (165)	45.70% (48)	41.90% (190)	40.00% (38)	40.40%
Have you ever shared your own research data?	Percentage (count) that said yes	86.40% (887)	82.10% (262)	76.90% (372)	75.50% (80)	72.60% (334)	75.00% (72)	80.50%
Have you ever reused your own research data?	Percentage (count) that said yes	71.40% (733)	73.70% (235)	68.60% (332)	68.90% (73)	81.10% (373)	78.10% (75)	73.10%
How important is it to you to:								
assess the reach and influence of your own data?	Mean rank (n=2,426)	1,198.74	1,190.18	1,327.98	1,331.41	1,130.38	1,134.67	1,213.00
assess the reach and influence of other people's data which you reuse?	Mean rank (n=2,382)	1,186.67	1,212.27	1,267.99	1,252.45	1,113.22	1,092.57	1,191.00
get credit for reusing other people's data?	Mean rank (n=2,352)	1,171.11	1,188.94	1,234.70	1,315.65	1,119.57	1,015.52	1,176.00
How important would it be for you to know the following information about others' data which you may potential reuse:								
the number of citations the data have received	Mean rank (n=2,492)	1,197.57	1,410.07	1,324.55	1,415.45	1,139.66	1,158.28	1,246.00
the number of times the data have been downloaded?	Mean rank (n=2,492)	1,182.13	1,349.49	1,315.32	1,507.34	1,201.66	1,172.77	1,246.00
the number of times the data have been viewed?	Mean rank (n=2,492)	1,194.75	1,328.19	1,350.18	1,452.40	1,146.19	1,259.24	1,246.00
information about where the data were used?	Mean rank (n=2,492)	1,142.86	1,268.98	1,362.83	1,517.50	1,274.75	1,259.47	1,246.00
descriptions or a narrative providing details about how the data were used?	Mean rank (n=2,492)	1,226.16	1,225.21	1,290.70	1,294.46	1,280.61	1,095.56	1,246.00
information about who has used the data?	Mean rank (n=2,492)	1,182.78	1,227.19	1,331.30	1,351.93	1,283.43	1,271.72	1,246.00
if the data have received recognition outside the scholarly system?	Mean rank (n=2,492)	1,175.35	1,290.33	1,318.02	1,375.25	1,258.33	1,302.64	1,246.00
How important would it be for you to know the following information about your data:								
the number of citations the data have received	Mean rank (n=2,492)	1,209.80	1,254.84	1,323.87	1,231.50	1,253.80	1,202.87	1,246.00
the number of times the data have been downloaded?	Mean rank (n=2,492)	1,176.04	1,208.67	1,333.13	1,319.33	1,312.95	1,290.42	1,246.00
the number of times the data have been viewed?	Mean rank (n=2,492)	1,183.29	1,207.42	1,368.26	1,308.89	1,255.31	1,327.55	1,246.00
information about where the data were used?	Mean rank (n=2,492)	1,146.23	1,213.67	1,358.29	1,363.35	1,346.01	1,258.82	1,246.00
descriptions or a narrative providing details about how the data were used?	Mean rank (n=2,492)	1,221.99	1,196.44	1,308.83	1,257.64	1,287.23	1,153.30	1,246.00
information about who has used the data?	Mean rank (n=2,492)	1,158.12	1,171.17	1,357.07	1,253.29	1,381.29	1,229.46	1,246.00
if the data have received recognition outside the scholarly system?	Mean rank (n=2,492)	1,189.45	1,164.42	1,305.72	1,255.54	1,363.46	1,260.56	1,246.00

Figure 15. Summary of statistically significant results by discipline for all questions. Blue indicates a result greater than the average of the reporting statistic; red indicates a result less than the average. Darker shades indicate larger deviation from the average.