emerald PUBLISHING

Online Information F

Data sharing and reuse practices: Disciplinary differences and improvements needed

Journal:	*Online Information Review*
Manuscript ID	OIR-08-2021-0423.R2
Manuscript Type:	Research Paper
Keywords:	Data sharing, Data reuse, cross-sectional survey, Disciplinary difference

Data sharing and reuse practices: Disciplinary differences and improvements needed

Abstract

Purpose

This study investigates differences and commonalities in data production, sharing and reuse across the widest range of disciplines yet, and identifies types of improvements needed to promote data sharing and reuse.

Design

The first authors of randomly selected publications from 2018 and 2019 in 20 Scopus disciplines were surveyed for their beliefs and experiences about data sharing and reuse.

Findings

From the 3,257 survey responses, data sharing and reuse are still increasing but not ubiquitous in any subject area and are more common among experienced researchers. Researchers with previous data reuse experience were more likely to share data than others. Types of data produced and systematic online data sharing varied substantially between subject areas. Although the use of institutional and journal-supported repositories for sharing data is increasing, personal websites are still frequently used. Combining multiple existing datasets to answer new research questions was the most common use. Proper documentation, openness, and information on the usability of data continue to be important when searching for existing datasets. However, researchers in most disciplines struggled to find datasets to reuse. Researcher feedback suggested 23 recommendations to promote data sharing and reuse, including improved data access and usability, formal data citations, new search features, and cultural and policy-related disciplinary changes to increase awareness and acceptance.

Originality

This study is the first to explore data sharing and reuse practices across the full range of academic discipline types. It expands and updates previous data sharing surveys and suggests new areas of improvement in terms of policy, guidance, and training programs.

Keywords

Data sharing, data reuse, cross-sectional survey.

Introduction

Collecting and producing new data is an integral part of research in many disciplines and a good dataset can even count as a standard research output in the UK (REF, 2019). Subsequently sharing research data in a findable, accessible, and interoperable format (Wilkinson et al., 2016) supports reproducibility, efficiency, collaboration and interdisciplinarity (Borgman et al., 2019). Sharing research data also confers a citation advantage (Piwowar et al., 2007; Henneken & Accomazzi, 2011; Colavizza et al., 2020). Nevertheless, there are powerful reasons for not sharing data, such as the time needed to do it effectively and the perception or reality that shared data are rarely reused (Bezuidenhout, 2019; Hansson & Dahlgren, 2022).

Data sharing is increasingly mandated by funders (Kiley et al., 2017), but not always by journals (Wiley, 2018), allowing many researchers to avoid it, unless encouraged by organisational or other

factors (Mason et al., 2020). Although accreditation can be an incentive for data sharing (Dorta-González et al., 2021), few studies have asked researchers what else would incentivise data sharing and promote data reuse (Whitty et al., 2015; Rowhani-Farid et al., 2017; Devriendt et al., 2021). Moreover, current suggestions sometimes have limited scope. For example, Whitty et al. (2015) suggested that journal editors can play a key role in incentivising data sharing in a public health emergency by only publishing data-driven research when the data had already been shared in a timely fashion with relevant authorities. Because of the incomplete uptake of data sharing, it is important to understand the enablers and barriers to data sharing and reuse in different disciplines. This will help stakeholders and policy makers design effective, and possibly tailored, interventions to increase data sharing when and where relevant.

Whilst there have been many studies of data sharing and reuse in narrow contexts, the lack of substantial science-wide investigations is a problem because of likely sharp disciplinary differences. Previous studies are difficult to generalise because of their focus on one or a small number of disciplines (Piwowar, 2011; Wallis et al., 2013; Federer et al., 2015; Faniel & Yakel, 2017; Zenk-Möltgen et al., 2018; Sardanelli et al., 2018) or specific data repositories (Bishop & Kuula-Luumi, 2017; Coady et al. 2017; Borgman et al., 2019). In the few surveys of multiple disciplines, ad-hoc participant recruitment via email and social media have led to few responses from disciplines where data sharing is apparently less common, such as Business and Economics, Arts and Humanities (Tenopir et al., 2015), obscuring the general picture. Although the Kim and Stanton's (2016) large-scale survey did not have this problem, it was limited to Science, Technology, Engineering, and Mathematics (STEM). In contrast, secondary analyses of previously collected questionnaires (Sayogo & Pardo, 2013; Curty et al., 2017; Kim et al., 2018) have struggled to give timely findings. Moreover, the data sharing environment is evolving rapidly as funder mandates take hold and journal data sharing requirements increase in some fields. Understanding disciplinary differences and updating prior surveys are therefore important to develop new national, international, and disciplinary research policies.

This study addresses the above gaps by comparing data production, sharing and reuse practices across 20 disciplines, including understudied research areas in Arts and Humanities, and Business and Economics, driven by the following research questions.

1. How do types and formats of data produced by researchers differ between disciplines?
2. How do researchers share data on the web? Does data sharing differ between disciplines and research experience?
3. How do researchers find repositories to share data and what factors influence their choice of repositories?
4. How frequently do researchers reuse existing data in different disciplines and for which purposes? How does it compare to data sharing in those disciplines?
5. How do researchers find datasets to reuse? Which factors are considered important when searching for existing datasets? How easy is it to find relevant datasets for reuse?
6. What can be improved in current systems to encourage and promote data sharing and reuse?

Background

More than three decades ago, Ceci (1988) proposed a scheme for mandatory data sharing between social scientists. In two surveys on the issue, most respondents (87%) were willing to share data, but 59% claimed that their colleagues were not, even for funded research. Today, most research data are produced in digital format, with infrastructures and standards available to support sharing of these data. Rapidly increasing numbers of data repositories now allow for effective curation, storage, and long-term access to data (Pampel et al., 2013). Nevertheless, disciplinary cultures, sizes and data types affect how and whether researchers share their data (Bell et al., 2009; Tenopir et al., 2015; Faniel & Yakel, 2017). For example, scholars in qualitative research fields are less prepared to openly share research data than those in data intensive fields (Mozersky et al., 2020). Previous studies suggest that journal mandates, disciplinary norms, perceived career benefits and scholarly altruism are all important for data sharing (Kim & Stanton, 2016; Zenk-Möltgen et al., 2018). In comparison, perceived effort, trust in colleagues and a lack of incentives can all undermine it (Piwowar, 2011; Wallis et al., 2013; Sayogo & Pardo, 2013; Fecher et al., 2015).

In the last decade, sharing data directly with other researchers (Federer et al., 2015) or through personal data storage have been common, with only 11.3% using institutional repositories, 9.5% using disciplinary repositories, and 2.4% using publisher-related repositories (Tenopir et al., 2015). There have been significant differences between disciplines in terms of disciplinary repository usage. It has been more likely in Ecology (44.6%) than in Physical Sciences (7.1%) and Social Sciences (10.8%) (Tenopir et al., 2015). These numbers have subsequently increased slightly, but personal storage remained commonly used for sharing (Tenopir et al., 2020). A recent study suggests that standard data repositories can be rarely used even in research fields with relatively strong data sharing norms, such as genomics (Thelwall et al., 2020). Even when data sharing statements are included in journal articles, repository links to meaningfully access the data are often missing (Federer et al., 2018). More comprehensive disciplinary information about repository uptake is therefore needed as a key step to long term sustainable data sharing.

When datasets are made easily accessible, researchers are generally willing to reuse others' data (Wallis et al., 2013; Bishop & Kuula-Luumi, 2017; Tenopir et al., 2020). However, where and how researchers find datasets to reuse are less known. The social media and web recruitment survey of Kratz and Strasser (2015) suggested that scholars use multiple search strategies to find data, including checking article references, searching discipline-specific databases, and using a general-purpose search engine. Their study did not report disciplinary differences, however. In a recent survey (using multiple online recruitment methods), 52% of 728 respondents self-reported reusing others' research data, but difficulties in finding appropriate data were common (Hrynaszkiewicz et al., 2021). Since finding data is critical to reusability, it is important to understand disciplinary differences in the core issues. Furthermore, there is a lack of understanding about what type of data are frequently reused across different disciplines and for what purposes.

Method

Consulting active researchers is the most direct method to get insights into how data are currently shared and reused. A survey is the only practical way to get large-scale evidence of attitudes and practices of data sharing across many different disciplines, and allows comparisons with previous studies (Fink, 2003). Two key assumptions are that respondents are representative of the

population of researchers and that they are able to accurately describe their experiences. To get a wide range of perspectives, the survey was international and targeted all career stages.

Questionnaire design

Fifteen questions (supplement 1) were designed to address the six research questions above, informed by the existing literature and previous surveys (Kratz & Strasser, 2015; Tenopir et al., 2015), but with some new questions for important omissions. An open-ended question was included on the type of data produced by researchers, since specific subject knowledge would be required to design a comprehensive list of options. Questions about data sharing methods were adapted from Kratz and Strasser (2015) and Tenopir et al. (2015) with additional questions on how researchers find repositories to share data and the factors that influence their choice of repositories. Questions about data reuse purposes used the typology of Pasquetto et al. (2019). A multiple-choice question on how researchers find datasets to reuse was adapted from Kratz and Strasser (2015), with additional questions on important factors when searching for existing datasets and ease of finding datasets to reuse. Finally, an open-ended question was designed to explore what can be improved in current systems to encourage and promote data reuse. Prior to circulating the survey, a pilot study was conducted with the researchers at the University of Wolverhampton to test the questions and identify necessary adjustments.

Selection of subject areas

Since previous studies have focused on data intensive STEM disciplines (Kim & Stanton, 2016) or specific repositories (Pasquetto et al., 2019) or had relatively small samples of less represented disciplines (Tenopir et al., 2015), an overview and comparison of different qualitative and quantitative disciplines was missing. In contrast, this study compared subject classifications in both Scopus[1] and Web of Science[2] (WoS) for pre-selecting a wide range of disciplines. Scopus was selected since its All Science Journal Classification (ASJC) subfield codes were more granular (333 disciplines within 27 subject areas) than the subject classifications in WoS. In total, 20 Scopus disciplines in nine Scopus subject areas were selected (Table I) to partly replicate disciplines and subject areas addressed in previous studies and to include new disciplines that have not been previously reported.

Data collection

a) Population and sampling

To ensure systematic coverage of researchers in the selected disciplines, the survey used direct email (Kim & Stanton, 2016) instead of soliciting responses from professional channels and social media (Kratz & Strasser, 2015; Tenopir et al., 2015). For each of 20 selected Scopus disciplines, Scopus was searched using their ASJC code (Table I), limiting the results to journal articles published in 2018 and 2019 to focus on currently active researchers. Metadata from 8,000 randomly selected studies were collected, half from each year. First author email addresses were extracted, where available, resulting in 3,500, on average, per discipline. A total of 70,060 researchers were identified for the study. Due to the interdisciplinary nature of some disciplines and papers in Scopus, survey respondents were allowed to self-identify their discipline, if different from the one suggested by their article, by selecting 'Other' or only select a broader subject area.

[1] https://www.scopus.com/

[2] https://clarivate.com/webofsciencegroup/solutions/web-of-science/

b) Survey data

Ethical approval for survey data collection was received from the University of Wolverhampton Life Sciences Ethics Committee (LSEC/201920/MT/125) on June 12, 2020. The Jisc Online survey platform was used to send individual survey invitations. The survey opened on July 14, 2020 and closed on August 17, 2020. In total, 70,060 invitations were emailed and 3,257 responses received (response rate 4.65%) (Khan et al., 2022). 214 respondents only selected a broader subject area and did not report their specific disciplines. The survey platform does not record whether emails have been blocked or returned, so the underlying response rate may have been slightly higher.

Data analysis

Originally, 402 responses were reported under 'Other', outside of the nine categories defined. However, 149 were variations of the disciplines listed in the study and these were merged with the main categories, leaving 253 responses in 'Other'.

Four out of 20 disciplines received fewer than 30 responses: Organic Chemistry; Radiology, Nuclear Medicine and Imaging; Aerospace Engineering; and Biomedical Engineering (Table I). These disciplines were excluded when analysing disciplinary differences. The cut-off 30 was chosen as a common statistical sample size threshold, in the absence of a theoretical reason to pick a given number.

The survey included single-choice and multiple-choice questions with an optional 'Other' field. These answers were tallied for different groups and content analysis was conducted on open-text answers in 'Other' fields. Free text from the open-ended question on data types was analyzed to find term frequencies in broader subject areas. Chi-square tests were used to examine the independence between categorical variables. Binomial multiple logistic regression was used to explore the effect of research experience and disciplinary differences on data sharing and reuse experiences. The assumptions for binary logistic regression were met by the following: 1. Binary dependent variable, 2. Each observation is independent of each other, 3. There is no multicollinearity among the independent variables, and 4. Adequate sample size – minimum 10 cases for each independent variable. The glm function[3] in the stats package (version 3.6.2) in R was used to perform binomial logistic regression. A manifest content analysis with an inductive approach was used to analyse the final open-ended question (Bengtsson, 2016).

Results

The 3,257 respondents mostly had over 10 years of research experience (64.4%), followed by 6-9 years (15.5%), 3-6 years (13.8%) and 0-3 years (6.2%), with similar levels in all subject areas. More experienced researchers may be more familiar with the concepts of data sharing and data reuse, and therefore more inclined to respond.

[3] https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm

The Social Sciences had the most responses (22.5%) within the broader subject areas, and Medicine had the fewest (5.2%). The percentage of responses in specific disciplines ranged from 5% (Organic Chemistry) to 60% (Astronomy and Astrophysics). Many selected 'Other' disciplines under a broader subject area (on average 39%), with the most in Engineering (63%) and the least in Environmental Sciences (17%) (Table I). The number of responses in previously underreported disciplines was significantly higher than in previous studies, including for education, linguistics, visual and performing arts, literature, business, and economics.

Table I: Selected subject areas and disciplines for the survey and the number of responses

Types and formats of data produced in different disciplines

Term frequency analyses suggested that survey and observations were the most common types of data produced across all subject areas. Qualitative data, audio, and video were common in Social Sciences and Arts and Humanities. In comparison, samples, measurements, simulations, and images were common in Science and Engineering (Table A, supplement 2). Thus, unsurprisingly, there are substantial differences in the data types produced.

Data formats also varied between subject areas (Figure 1; participants could select multiple formats). Numerical data was overall popular across all subject areas except Arts and Humanities (25%). Text was the most common format in Social Sciences (74%) and Arts and Humanities (88%). These two research areas and Engineering were the top producers of multimedia (audio and video) data: Visual and Performing Arts (33%) and Linguistics and Language (38%). In contrast, Physical Sciences (45%), Engineering (40%), and Oceanography (45%) in Earth and Planetary Sciences (36%) generate many computer programs. Biomedical Sciences (55.3%) produces many images; this category was common overall in all subject areas except Social Sciences and Business and Economics.

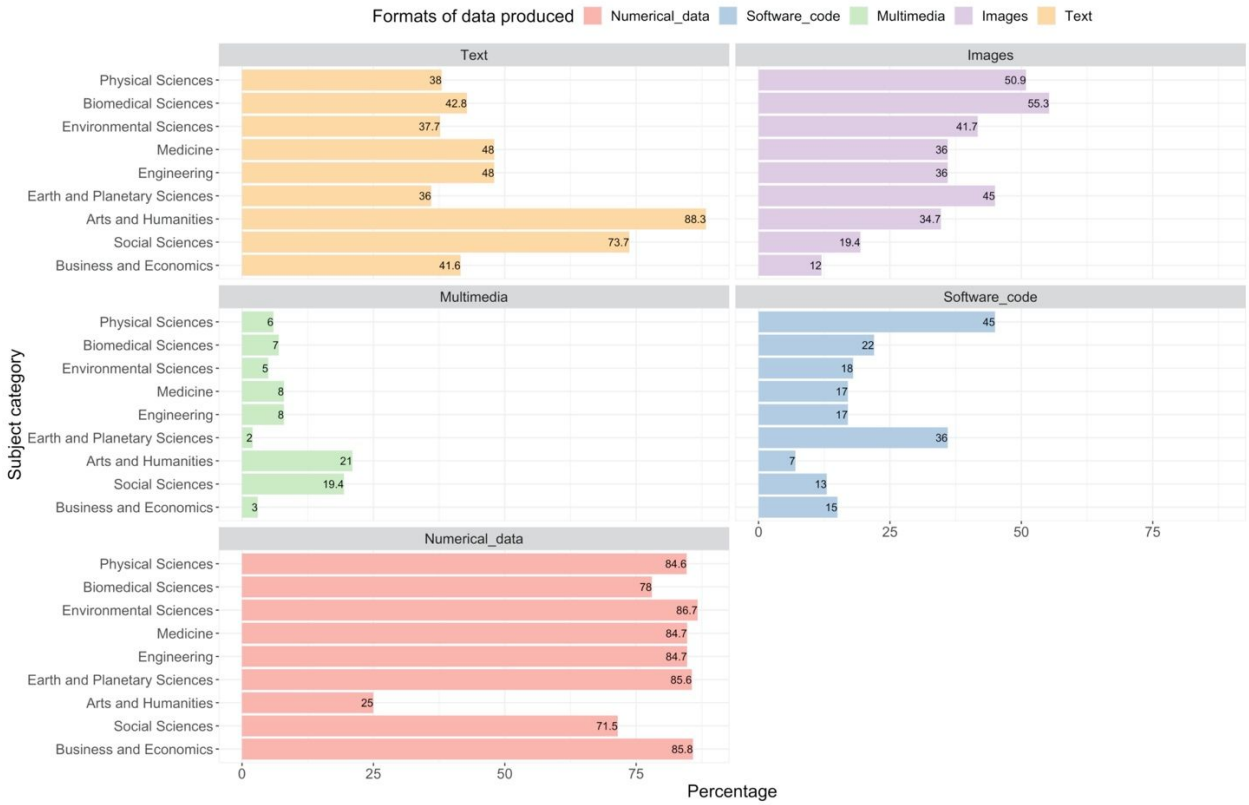

Figure 1. Formats of data produced in different subject areas

Data sharing across disciplines and research experience levels

Nearly half (46.8%, n=1,523) of the participants reported sharing data online. Self-reported data sharing experience varied among researchers in different stages of their research career (χ^2=36.85, p<0.001), and data sharing was more common with more research experience (from 39% during 0-3 years to 50% after 10+ years).

Differences in the prevalence of data sharing were significant between subject areas (χ^2=200.17, p<0.001). Physical Sciences (73%) shared most, followed by Earth and Planetary Sciences (70%) (Figure 2). In comparison, data sharing was less common in Business and Economics (33%) and Medicine (38%). A binomial logistic regression explored the effect of subject area on data sharing behavior while controlling for research experience. Compared to Arts and Humanities, being in Business and Economics, as well as Medicine significantly decreased the probability of data sharing by 0.5 and 0.63 times respectively. In contrast, the chances of data sharing increased by 2.9 times for researchers in Physical Sciences and 2.4 time for those in Earth and Planetary Sciences (Table B, supplement 2).

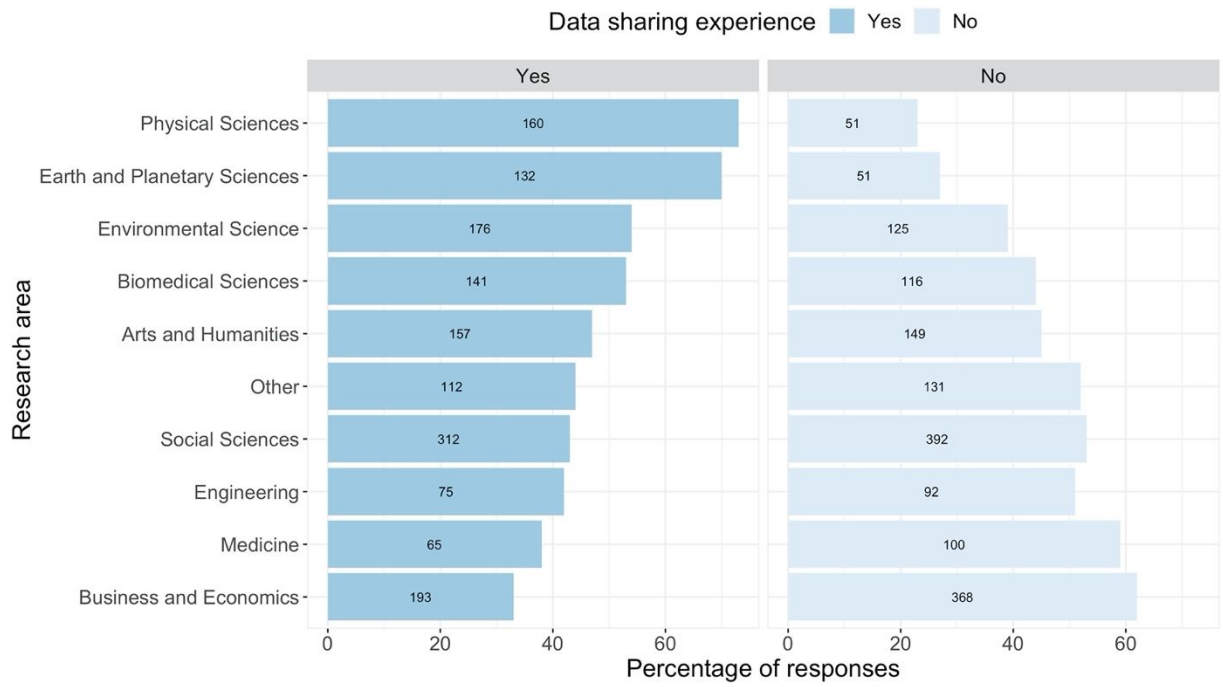

Figure 2. Data sharing by subject area (labels on bars represent number of responses)

Differences in the extent of data sharing exist within specific disciplines as well as between broad subject areas. For example, within the Business and Economics research area, only 26% (n=50) of 193 respondents in Business and International Management had previous data sharing experience, compared to 44% (n=110) of 251 respondents in Economics and Econometrics. Similarly, within the Social Sciences, data sharing was less common in Education (27%, n=31) compared to Library and Information Sciences (46%, n=117), and Linguistics and Language (46%, n=33). Despite data sharing being less common in Medicine, researchers in Infectious Disease (45%, n=17) more commonly shared data than do those in Radiology (26%, n=7).

Methods of sharing research data

From the different methods of sharing data on the web, over half of the respondents mentioned institutional repositories (53.4%, n=813), followed by journal-supported repositories (30%, n=457) and personal websites (24.5%, n=373) (participants could select multiple methods). Chi-square tests confirm the statistical significance of differences between subject areas in the types of method used for sharing data (Table II).

Table II. Data sharing methods in different subject areas

The use of disciplinary repositories was common in most STEM fields except Engineering (1%) and was rare in Arts and Humanities (7%), and Business and Economics (8%). Interdisciplinary repository usage was relatively common in Biomedical Sciences (26%), Earth and Planetary Sciences (24%) and Social Sciences (23%) and least common in Medicine (6%). Sharing data on personal websites was most common in Physical Sciences (37%) and least in Medicine (9%). Non-standard data deposit practices in the Social Sciences and Arts and Humanities include Academia.edu and Google drive, which are not ideal solutions for long-term retrieval.

Choice of data repositories

When asked how they first found repositories to share data, most researchers responded that they were already aware of them, even though this varied between disciplines (Table III). For example, in Physical Sciences and Biomedical Sciences, over 60% of respondents were already aware of relevant data repositories. This was followed by consulting with colleagues and consulting with experts, which was common across all disciplines. General web searches for repositories were more common in Engineering (29%) and Arts and Humanities (20%). Searching re3data, the registry of research data repositories, was not a preferred method by researchers in any discipline (5% or less), so this is perhaps a professional librarian's tool.

Table III. How researchers first found repositories to share data

Ease of use (53.8%, n=820), repository reputation (46.9%, n=714), disciplinary norms (41.1%, n=626), and appropriateness for the data type (40.5%, n=617) were the top reasons for choosing a data repository (Table C, supplement 2). Other factors that influence researchers' choices are requirement from funding bodies, journals and institutions, accessibility, privacy, security, zero cost, digital object identifier (DOI) assignment, interdisciplinary research support, and international reputation for collaborative project support. The following factors were dependent on disciplinary differences: Reputation of repository, cost, and appropriateness for data type. Cost and appropriateness for data type were important factors in disciplines where disciplinary repositories were more commonly used.

Data reuse across disciplines and research experience

Overall, 54.3% (n=1,769) of the respondents had reused existing datasets. Data reuse frequency was dependent on researchers' experience (χ^2=8.88, p=0.03), increasing with research experience: 47% in 0-3 years, 49% in 3-6 years, 53% in 6-9 years, and 56% after 10+ years.

Data reuse experience significantly varied between subject areas (χ^2=152.03, p<0.001). Over 80% of respondents in Physical Sciences and Earth and Planetary Sciences, and 56~60% of respondents in Business and Economics, Environmental Sciences, and Engineering had reused existing data (Figure 3). This rate was lower among Arts and Humanities (42%), Medicine (44%), Social Sciences (47%), and Biomedical Sciences (49%). Only 1-3% of participants in all subject areas responded that their research does not use data; Arts and Humanities was an exception (16%). Outcomes of a binomial multivariable logistic regression (subject areas and research experience as predictors) indicate that when compared to data reuse in Arts and Humanities, the probability of data reuse increased by 5.86 times in Earth and Planetary Sciences; 5.56 times in Physical Sciences; 1.76 times in Engineering; 1.72 times in Environmental Sciences; and 1.5 times in Business and Economics (Table D, supplement 2).

Data reuse varied within specific disciplines in Business and Economics, as well as in Environmental Sciences. In contrast to 72% (n=180) researchers in Economics and Econometrics, only 38% (n=74) in Business and International Management reused secondary data. Within Environmental Sciences, data reuse was more common in Ecology (63%, n=93) than Pollution

(49%, n=53). This trend is similar to data sharing behaviour in these fields as data sharing was less common in the fields that predominantly rely on primary data.

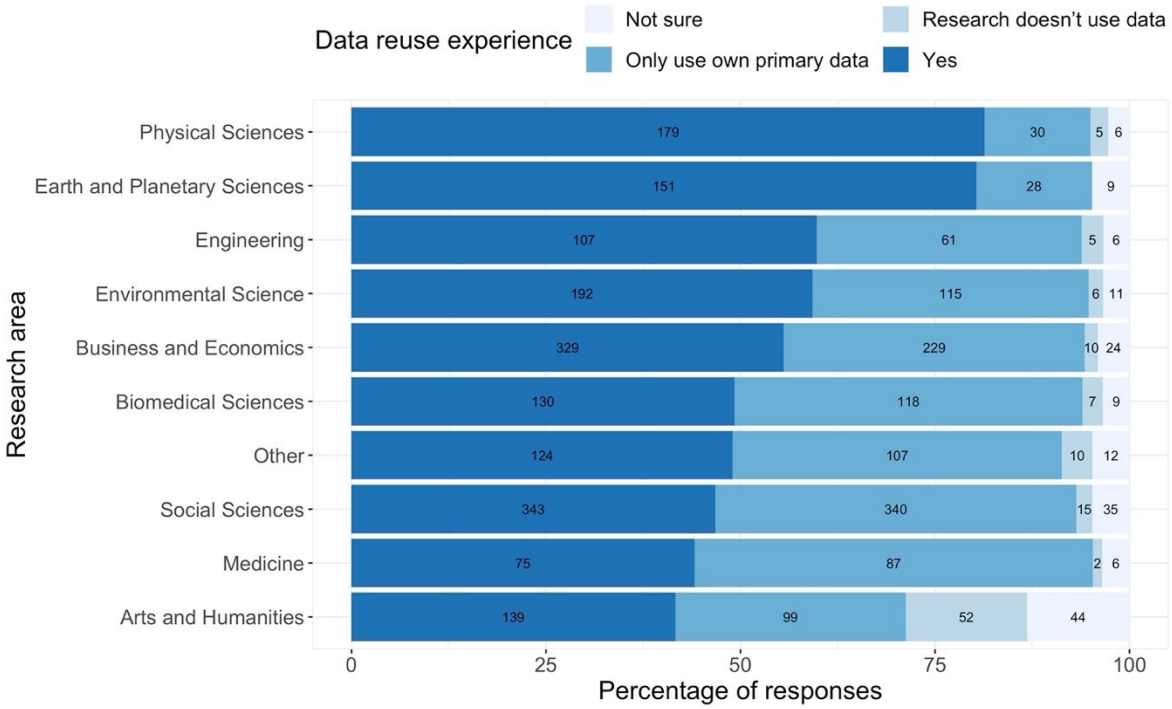

Figure 3. Data reuse across subject areas (labels on bars represent number of responses)

Data reuse purposes

Overall, 63.1% (n=1,116) of researchers reported that they combine multiple existing datasets to answer novel research questions; 50.7% (n=897) reused data for comparing or ground truthing, i.e., calibrate, compare, confirm; and 46.6% (n=825) analysed a single dataset to answer novel research questions. Data reuse types varied between subject areas ($p < 0.001$ across all three types) (Figure 4). From the dotted lines in Figure 4 (the average in each category of data reuse), analysis of a single dataset was most common in Medicine (59%) and least common in Environmental Sciences (29%). Combining multiple datasets to answer new research questions was common overall, especially in Earth and Planetary Sciences (80%), Physical Sciences (77%), and Environmental Sciences (71%). Comparative data analysis was most common in Engineering (71%) and least common in Business and Economics (28%).

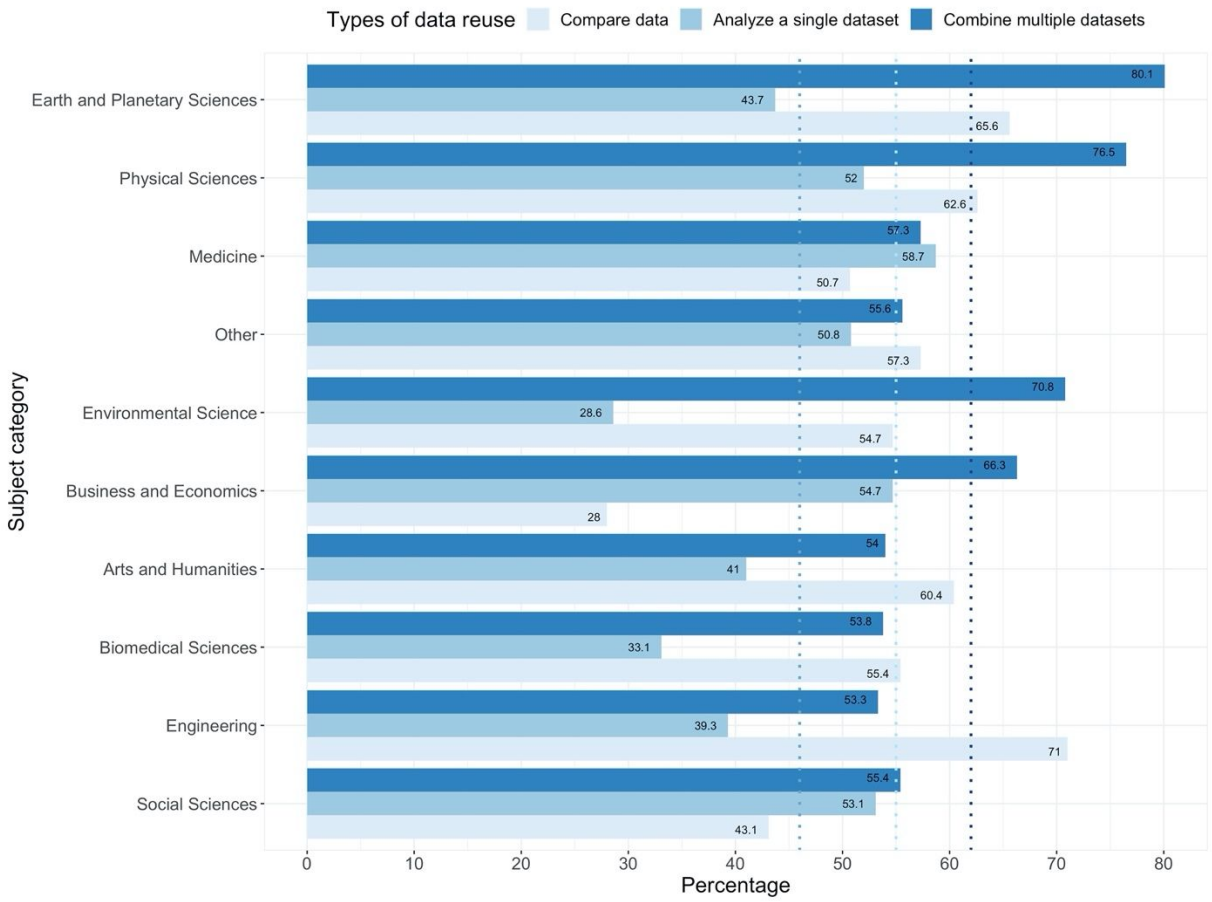

Figure 4 Data reuse types in different subject areas (dotted lines represent the average percentage in each area)

Other reuse types include testing and validating machine learning models, historical data analysis, teaching (e.g., student projects), evolution analysis, quantifying long-term climate conditions, applying new statistical methods on existing datasets, replicating findings in diverse populations, reusing existing linguistic corpora, systematic review and meta-analysis, using GIS data to correlate with image files, and discourse analysis.

Data sharing vs data reuse

Researchers who reuse data were more likely to share data (56.8%, n=1,004), compared to those who only used their own primary data (32.6%, n=396). In contrast, data reuse was more frequent in Engineering and Business and Economics than data sharing on the web. However, sharing and reuse of data were dependent overall (χ^2=181.11, $p < 0.001$). This was the same within all individual subject areas except Medicine and Engineering (Table E, supplement 2).

Data sharing among those who rely on their own data was relatively common in Earth and Planetary Sciences (50%), Arts and Humanities (44%), Physical Sciences (43%), and

Environmental Science (41%) (Figure B, supplement 3). It is possible that those who reuse data shared by other researchers are more aware of data sharing practices in their field, but those who only use their own primary data for research are less so. Alternatively, the common factor may be the importance of data sharing for particular specialties.

Finding datasets to reuse

Among the researchers who had reused datasets, 60.9% (n=1,078) found datasets by reading relevant papers. Also popular were web searches, such as Google Dataset Search (46.1%, n=816), and disciplinary repository searches (45.6%, n=806). However, all methods to find datasets varied in popularity between subject areas (Table F, supplement 2). Searching disciplinary repositories was more common in Physical Sciences (69.4%) and Earth and Planetary Sciences (50%), whereas interdisciplinary repository search was more common in Arts and Humanities (35%) and Social Sciences (28%). Similarly, web search was a majority choice in Engineering (63%), Arts and Humanities (60%), and Business and Economics (51.7%). These methods were not dependent on research experience.

The following factors were considered most important by researchers when searching for existing datasets to reuse: proper documentation (67%, n=2,195), open data (52%, n=1,678), and information on usability of data (42%, n=1,375). Availability of data in a universal standard format (36%, n=1163) and evidence that the dataset has an associated publication (34%, n=1,107) were of moderate importance. Evidence of prior reuse was rarely considered important (8.3%, n=270).

Despite evidence of increasing data reuse in all disciplines, most researchers reported difficulty finding datasets to reuse. Physical Sciences was an exception, where over 50% researchers could easily find datasets to reuse. This percentage was above average in Earth and Planetary Sciences (33%, n=57 out of 174) and Biomedical Sciences (29%, n=58 out of 199) as well.

Finding datasets becomes slightly easier with experience. 24% (n=511 out of 2,090) of researchers with over 10 years of research experience found it difficult to find datasets to reuse, compared to 26~28% of those with less experience (Figure C and D, supplement 3).

Future improvements

1,831 open-text responses suggested future improvements in current systems to promote data sharing and reuse. A content analysis of these responses identified 23 recommendations in eight themes within three categories: 1. Issues around data, 2. Technological solutions, and 3. Cultural and policy changes (Table IV).

Table IV Future improvements needed to promote data sharing and reuse

The most mentioned barrier to data reuse was a lack of knowledge about where and how to search for datasets. Therefore, a single trusted portal or federated search system across disciplines is needed that allows easy discovery of data:
"Perhaps more universal/federated searching mechanisms or portals--ArchiveGrid (https://researchworks.oclc.org/archivegrid/) was a game-changer for my research when it was released--now I no longer have to think "where might records about X person be?" and go to each individual institution and search."

A few responses pointed out that not all datasets can have multiple use cases, because some are created for a single use only. Therefore, information on the applicability of data can be helpful to external users. Adequate contextual information is key to successful reuse of data, along with researchers' commitments to share data and proper data curation.

Streamlined institutional review board rules are critical to data sharing for reuse purposes in research where human participants are involved. Legal constraints about cultural data can be an impediment to data reuse in the Arts and Humanities. A response from a humanities researcher outlines different policy issues and the need for incentives:
"Before we can improve data reuse, we need to improve communications between disciplines, accept the resource costs of making data reusable, reward people who do make their data reusable, and of course work with legal systems and institutions (archives, libraries, publishers etc) who 'own' cultural data to make reuse for research more fluid".

Collaborations between data creators and reusers were recommended by multiple participants, as well as changes in research culture and policies. Participants mentioned that secondary data analysis may not be considered 'original enough' by journals to be published. Environmental Sciences researchers mentioned that their data are often very difficult to collect, and they can be reluctant to give away the fruits of their labour. Incentives such as data badges, data reuse indicators, more funding for secondary data analysis projects, and rescuing of historical data were recommended to promote more data sharing and reward data creators. As suggested by one participant:
"…data work is nowadays high-quality scientific work as well, i.e., the reputation for data work needs to be increased (co-authorship for data work; establish "data"-chairs at universities and research institutes, etc.)"

Limitations

The precision of the results is affected by differing subgroup sample sizes. The sample sizes of researchers in different experience groups varied, with over 60% in the 10+ years' experience group. This could be due to the topic of survey since we found that those with more experience tend to share and reuse datasets more frequently. In addition, four disciplines had fewer than 30 responses. Differences for these disciplines were not reported separately as the results may not accurately represent that group. The participant recruitment method may also have impacted this (i.e., sample selection bias) as more experienced researchers tend to publish more and are listed as the first author more frequently. The results also have an unknown survey self-selection bias related to the 4.65% response rate. Unlike similar studies (Unal et al., 2019; Tenopir et al., 2020),

researchers' geographic location was not considered in this survey due to its focus on web-based data sharing and reuse.

Discussion

Data sharing is known to be increasing in some disciplines to comply with funding body and institutional requirements. However, research data are not always shared in a meaningful way that can lead to long-term accessibility and reuse. In this study, both data sharing and reuse were dependent on researchers' experience; those with more than 10 years of experience tended to share and reuse data more often. This supports the positive association between data sharing and a longer career reported by Gregory et al. (2020) and Dorta-González et al. (2021). Disciplinary differences exist in how researchers share data on the web, presumably driven by the culture of data sharing in a discipline – Physical Sciences, Earth and Planetary Sciences, and Environmental Sciences are more likely, whereas Business and Economics, Medicine, and Engineering are less likely to share data. Institutional repositories were frequently used in all disciplines, with journal-supported repositories also being quite popular. This could be because of rapid growth of institutional repositories and research data services in higher education institutions to comply with funder mandates (Cragin et al., 2010; Cox et al., 2017). Many journals are also mandating data accessibility statements and have associated data repositories, such as Mendeley Data by Elsevier. These results extend the previously known patterns in Tenopir et al. (2015) to a wider range of disciplines, (e.g., Business and Economics) and demonstrate increased use of such repositories in recent years.

Disciplinary repositories have emerged to support domain specific data, such as in astronomy and astrophysics, zoology, and social science (Wallis et al., 2013; Faniel & Yakel, 2017). Data sharing and reuse are relatively common in these fields because researchers tend to be more aware of frequently used repositories in their specialty. This is in line with the good data practices reported by Tenopir et al. (2020) for Earth and Planetary Sciences and Environmental Sciences. However, the current results suggest that disciplinary repository usage has increased in Physical Sciences in recent years, compared to Tenopir et al. (2015). Personal websites were also frequently used for data sharing in many subject areas but not in Medicine, perhaps because of sensitive personal health data. This aligns with the findings of Tenopir et al. (2020). The examples of commonly used repositories reported by participants in this study demonstrate a lack of established data sharing methods in Engineering, Business and Economics, and Arts and Humanities, which could be one of the reasons for less data sharing in these subject areas. The Registry of Research Data Repositories (re3data.org) currently lists 951 repositories under Humanities and Social Sciences, including 207 for Economics, and 517 repositories for Engineering Sciences among other disciplines, so the infrastructure for sharing seems to be available.

This study shows that the previously reported growing data reuse in most disciplines (Bishop & Kuula-Luumi, 2017; Borgman et al., 2019; Khan et al., 2021) has continued and is highest in Physical Sciences and Earth and Planetary Sciences. Self-reported data reuse was significantly more common in Engineering and Business and Economics than data sharing. In contrast to Curty et al.'s (2017) secondary analysis, the evidence here suggest that data sharing and reuse are dependent, except for Engineering and Medicine. This suggests that the relationship between data sharing and reuse has evolved, perhaps due to a greater accumulation of data sharing experience over time.

Despite high levels of data reuse, researchers in most disciplines except Physical Sciences usually struggle to find datasets to reuse. Hrynaszkiewicz et al. (2021) reported similar findings for their overall study population. The study results also support the findings of Kratz and Strasser (2015) that most researchers read relevant papers to find reusable datasets, with web searches and disciplinary repository searches also being common. The findings here extend these prior findings to show disciplinary differences. Searching disciplinary repositories was common in Physical Sciences and Earth and Planetary Sciences, compared to other disciplines. Even though reading papers is opted for by over 60% of researchers, recent studies reported that only a small percentage of journal articles share data in a meaningful and accessible way (Federer et al., 2018; Thelwall et al., 2020). This may increase the difficulty of finding datasets from relevant articles. Formalizing data citation across all disciplines was suggested by the respondents. This will ensure that datasets are linked to associated articles and in turn increase visibility of data, as well as being an incentive for further data sharing (Dorta-González et al., 2021).

Conclusion

This study has revealed the extent to which data production, sharing and reuse varies between disciplines. While self-reported data sharing is increasing, significant disciplinary differences remain in the adoption of standard data sharing methods. Particularly for qualitative disciplines involving human participants, adequate guidance should be developed, and existing guidelines need to be reviewed by experts and funders to support best practices for de-identifying data. For example, in 2012 the U.S. Department of Health and Human Services published guidance for de-identification standards in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Such guidance is useful but may not fit all purposes and should be adapted for different country regulations and discipline specific rules.

In contrast to previous studies, widespread usage of institutional repositories in this study sample indicates that institutional support can play an important role in data sharing. At the institutional level, research data management training programs and curated resources (e.g., lists of relevant data repositories) can help researchers in all disciplines adopt best practices for data production, management and sharing. Early career researchers will especially benefit from this because data sharing and reuse were less common among less experienced respondents. Resources developed by the community of researchers, such as FAIRsharing.org[4] can be used in training to help find a suitable data repository, as can guidance provided in existing studies (Alter & Gonzalez, 2018; Figueiredo, 2017).

Standard data sharing and citation makes data findable and accessible in the long-term and can help reduce the burden of finding data to reuse. Although the results show that sharing data as supplementary materials in a journal or personal website is still common across disciplines, this does not ensure better discoverability and accessibility, and journal editors can ensure any related research data are deposited in a standard manner, adhering to FAIR principles. Since web search (e.g., Google Dataset Search[5]) was the second most common method used to find reusable datasets, data repository managers can help researchers by adopting Schema.org metadata standards to be indexed by Google Dataset Search and make their datasets more easily discoverable and reusable

[4] https://fairsharing.org/
[5] https://datasetsearch.research.google.com

(Patel, 2019). This will particularly help siloed institutional repositories as the survey results show that researchers are more likely to search well-known disciplinary repositories to find datasets to reuse.

Future studies can examine researchers' attitudes and needs in Arts and Humanities, Business and Economics, and Engineering to further explore why data sharing is particularly low in these subject areas despite relatively frequent data reuse. This will help to identify areas that need new policies, guidance, and infrastructure development. Furthermore, based on the researchers' responses, this study makes 23 recommendations related to data sharing, technological solutions and cultural and policy changes to support data sharing and promote data reuse. Incentives such as rewarding data creators in a formal manner similar to article publishing and implementing data reuse indicators or data badges to visualize impact of data sharing seem to be particularly useful. However, there are few incentives-based studies of data sharing (Rowhani-Farid et al., 2017; Devriendt et al., 2021) and more are needed to properly assess the impact of incentives on data sharing.

Data availability statement

Data collected from this study are available on figshare:
https://doi.org/10.6084/m9.figshare.19596967.v1

Conflicts of interest/Competing interests

Not applicable

References

1. Alter, G., & Gonzalez, R. (2018), "Responsible practices for data sharing", *American Psychologist*, 73(2), 146.
2. Bell, G., Hey, T., & Szalay, A. (2009), "Beyond the data deluge. Science", 323(5919), 1297-1298.
3. Bengtsson, M. (2016), "How to plan and perform a qualitative study using content analysis", NursingPlus, Open, 2, 8-14.
4. Bezuidenhout, L. (2019), "To share or not to share: Incentivizing data sharing in life science communities", *Developing World Bioethics*, Vol. 19 No. 1, pp. 18-24.
5. Bishop, L., & Kuula-Luumi, A. (2017), "Revisiting qualitative data reuse: A decade on", *Sage Open*, Vol. 7 No. 1, 2158244016685136.
6. Borgman, C. L., Scharnhorst, A., & Golshan, M. S. (2019), "Digital data archives as knowledge infrastructures: Mediating data sharing and reuse", *Journal of the Association for Information Science and Technology*, 70(8), pp. 888-904.
7. Ceci, S. J. (1988), "Scientists' attitudes toward data sharing", *Science, Technology, & Human Values*, 13(1-2), pp.45-52.
8. Coady, S. A., Mensah, G. A., Wagner, E. L., Goldfarb, M. E., Hitchcock, D. M., & Giffen, C. A. (2017), "Use of the national heart, lung, and blood institute data repository", *New England Journal of Medicine*, Vol. 376 No. 19, pp.1849-1858.
9. Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020), "The citation advantage of linking publications to research data", *PloS one*, Vol. 15 No. 4, e0230416.

10. Cox, A. M., Kennan, M. A., Lyon, L., & Pinfield, S. (2017), "Developments in research data management in academic libraries: Towards an understanding of research data service maturity", *Journal of the Association for Information Science and Technology*, Vol. 68 No. 9, pp. 2182-2200.
11. Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010), "Data sharing, small science and institutional repositories", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 368 No. 1926, pp. 4023-4038.
12. Curty, R. G., Crowston, K., Specht, A., Grant, B. W., & Dalton, E. D. (2017), "Attitudes and norms affecting scientists' data reuse", PloS one, 12(12), e0189288.
13. Devriendt, T., Shabani, M., & Borry, P. (2021), "Data sharing in biomedical sciences: a systematic review of incentives", Biopreservation and Biobanking, 19(3), 219-227.
14. Dorta-González, P., González-Betancor, S. M., & Dorta-González, M. I. (2021), "To what extent is researchers' data-sharing motivated by formal mechanisms of recognition and credit?", *Scientometrics*, *126*(3), 2209-2225.
15. Faniel, I. M., & Yakel, E. (2017), "Practices do not make perfect: Disciplinary data sharing and reuse practices and their implications for repository data curation", *Curating research data, volume one: Practical strategies for your digital repository*, 1, pp.103-126.
16. Fecher, B., Friesike, S., & Hebing, M. (2015), "What drives academic data sharing?", *PloS one*, 10(2), e0118053.
17. Federer L. M., Lu Y-L, Joubert D. J., Welsh J., Brandys B. (2015), "Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff", *PLoS ONE,* Vol. 10 No. 6, e0129506. https://doi.org/10.1371/journal.pone.0129506
18. Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y. L., Snyders, L. N., & Thompson, H. (2018), "Data sharing in PLOS ONE: an analysis of data availability statements", *PloS one*, 13(5), e0194768.
19. Figueiredo, A. S. (2017), "Data sharing: convert challenges into opportunities", *Frontiers in public health*, 5, 327.
20. Fink, A. (2003), *How to design survey studies*, Sage.
21. Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2020), "Lost or Found? Discovering Data Needed for Research", Harvard Data Science Review, 2(2).
22. Hansson, K., & Dahlgren, A. (2022), "Open research data repositories: Practices, norms, and metadata for sharing images", *Journal of the Association for Information Science and Technology,* Vol. 73 No. 2, pp. 303-316.
23. Henneken, E. A., & Accomazzi, A. (2011), "Linking to data-effect on citation rates in astronomy", arXiv preprint arXiv:1111.3618.
24. Hrynaszkiewicz, I., Harney, J., & Cadwallader, L. (2021), "A survey of researchers' needs and priorities for data sharing".
25. Khan, N., Thelwall, M. & Kousha, K. (2021), "Measuring the impact of biodiversity datasets: data reuse, citations and altmetrics", *Scientometrics*, pp.1-19.
26. Khan, N., Thelwall, M., & Kousha, K. (2022), "Survey data on disciplinary differences in data sharing and reuse practices", figshare. Dataset. https://doi.org/10.6084/m9.figshare.19596967.v1

27. Kiley, R., Peatfield, T., Hansen, J., & Reddington, F. (2017), "Data sharing from clinical trials—a research funder's perspective", *The New England Journal of Medicine*, 377:1990-1992.
28. Kim, J., Schuler, E. R., & Pechenina, A. (2018), "Predictors of data sharing and reuse behavior in academic communities", In Knowledge Discovery and Data Design Innovation: Proceedings of the International Conference on Knowledge Management (ICKM 2017), pp. 1-25.
29. Kim, Y., & Stanton, J. M. (2016), "Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis", *Journal of the Association for Information Science and Technology*, 67(4), pp.776-799.
30. Kim, Y., & Yoon, A. (2017), "Scientists' data reuse behaviors: A multilevel analysis", *Journal of the Association for Information Science and Technology*, Vol. 68 No. 12, pp. 2709-2719.
31. Kratz, J. E., & Strasser, C. (2015), "Making data count", *Scientific Data*, Vol. 2 No. 1, pp. 1-5.
32. Mason, C. M., Box, P. J., & Burns, S. M. (2020), "Research data sharing in the Australian national science agency: Understanding the relative importance of organisational, disciplinary and domain-specific influences", Plos one, 15(8), e0238071.
33. Meystre, S. M., Lovis, C., Bürkle, T., Tognola, G., Budrionis, A., & Lehmann, C. U. (2017), "Clinical data reuse or secondary use: current status and potential future progress", *Yearbook of medical informatics*, 26(01), 38-52.
34. Mozersky, J., Walsh, H., Parsons, M., McIntosh, T., Baldwin, K., & DuBois, J. M. (2020), "Are we ready to share qualitative research data? Knowledge and preparedness among qualitative researchers, IRB Members, and data repository curators", *IASSIST quarterly*, 43(4).
35. Office for Civil Rights (2012), "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule", available at: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf
36. 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).
37. Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.J., Gundlach, J., Schirmbacher, P. and Dierolf, U. (2013). Making research data repositories visible: the re3data. org registry. PloS One, 8(11), e78080.
38. Patel, D. (2019), "How Google's Dataset Search Engine Work", available at: https://towardsdatascience.com/how-googles-dataset-search-engine-work-928fa5237787 (accessed 31 March 2021).
39. Pinfield, S., Cox, A. M., & Smith, J. (2014), "Research data management and libraries: Relationships, activities, drivers and influences", *PLoS One*, Vol. 9 No. 12, e114734.
40. 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.

41. Piwowar, H. A. (2011), "Who shares? Who doesn't? Factors associated with openly archiving raw research data", *PloS one*, 6(7), e18657.
42. "re3data.org - Registry of Research Data Repositories", available at: https://doi.org/10.17616/R3D (accessed 17 November 2020).
43. REF (2019), "Guidance on submissions (2019/01) – REF 2021", available at: https://www.ref.ac.uk/publications/guidance-on-submissions-201901/ (accessed 13 July 2021).
44. Rowhani-Farid, A., Allen, M., & Barnett, A. G. (2017), "What incentives increase data sharing in health and medical research? A systematic review", *Research integrity and peer review*, 2(1), 1-10.
45. Sayogo, D. S., & Pardo, T. A. (2013), "Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data", *Government information quarterly*, 30, S19-S31.
46. 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.
47. 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.
48. Thelwall, M., Munafò, M., Mas-Bleda, A., Stuart, E., Makita, M., Weigert, V., Keene, C., Khan, N., Drax, K. and Kousha, K. (2020), "Is useful research data usually shared? An investigation of genome-wide association study summary statistics", *Plos One*, Vol. 15 No. 2, e0229578. https://doi.org/10.1371/journal.pone.0229578
49. Unal, Y., Chowdhury, G., Kurbanoglu, S., Boustany, J., & Walton, G. (2019), "Research data management and data sharing behaviour of university researchers".
50. 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*, Vol. 8 No. 7, e67332.
51. Whitty, C. J., Mundel, T., Farrar, J., Heymann, D. L., Davies, S. C., & Walport, M. J. (2015), "Providing incentives to share data early in health emergencies: the role of journal editors", *The Lancet*, *386*(10006), 1797-1798.
52. Wiley, C. (2018), "Data sharing and engineering faculty: An analysis of selected publications", *Science & technology libraries*, 37(4), pp.409-419.
53. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J. (2016), "The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, Vol. 3 No. 1, pp. 1-9.
54. Yoon, A. (2016), "Red flags in data: Learning from failed data reuse experiences", *Proceedings of the Association for Information Science and Technology*, Vol. 53 No. 1, pp.1-6.
55. Zenk-Möltgen, W., Akdeniz, E., Katsanidou, A., Naßhoven, V., & Balaban, E. (2018). Factors influencing the data sharing behavior of researchers in sociology and political science. *Journal of Documentation*.

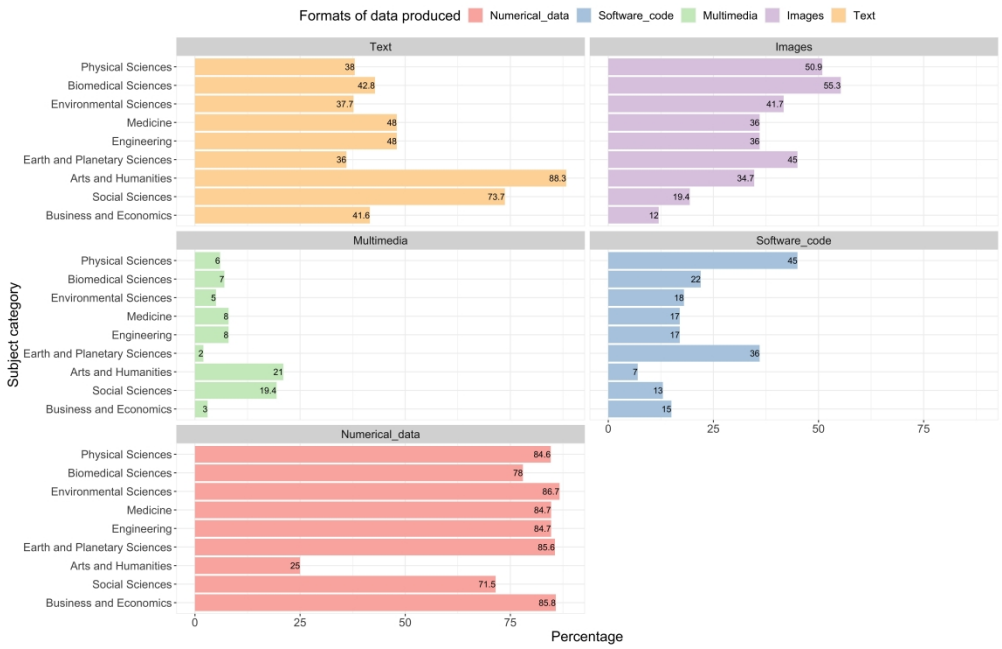

Figure 1: Formats of data produced in different subject areas

2777x1805mm (72 x 72 DPI)

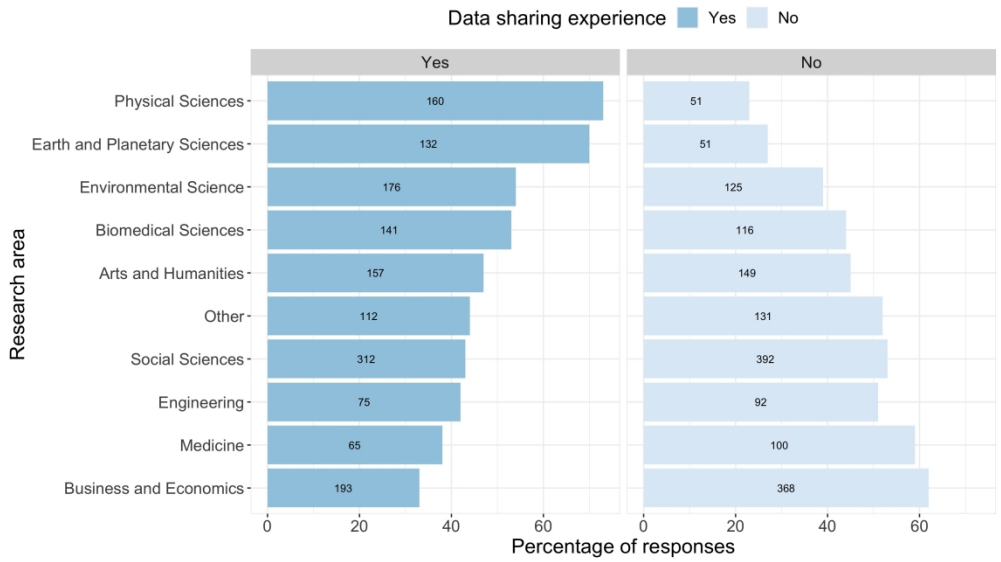

Figure 2. Data sharing by subject area (labels on bars represent number of responses)

1166x666mm (72 x 72 DPI)

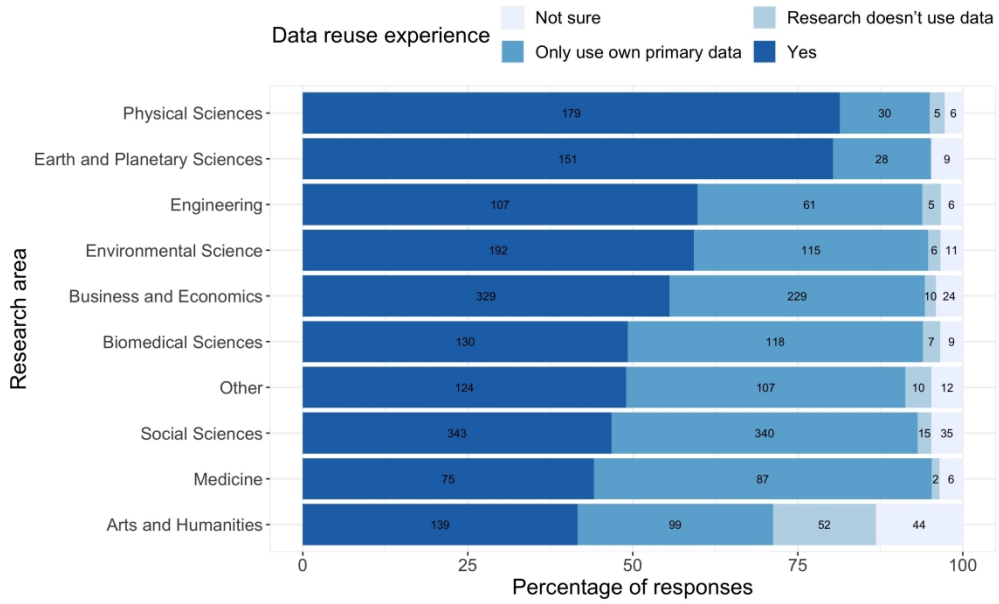

Figure 3. Data reuse across subject areas (labels on bars represent number of responses)

1083x666mm (72 x 72 DPI)

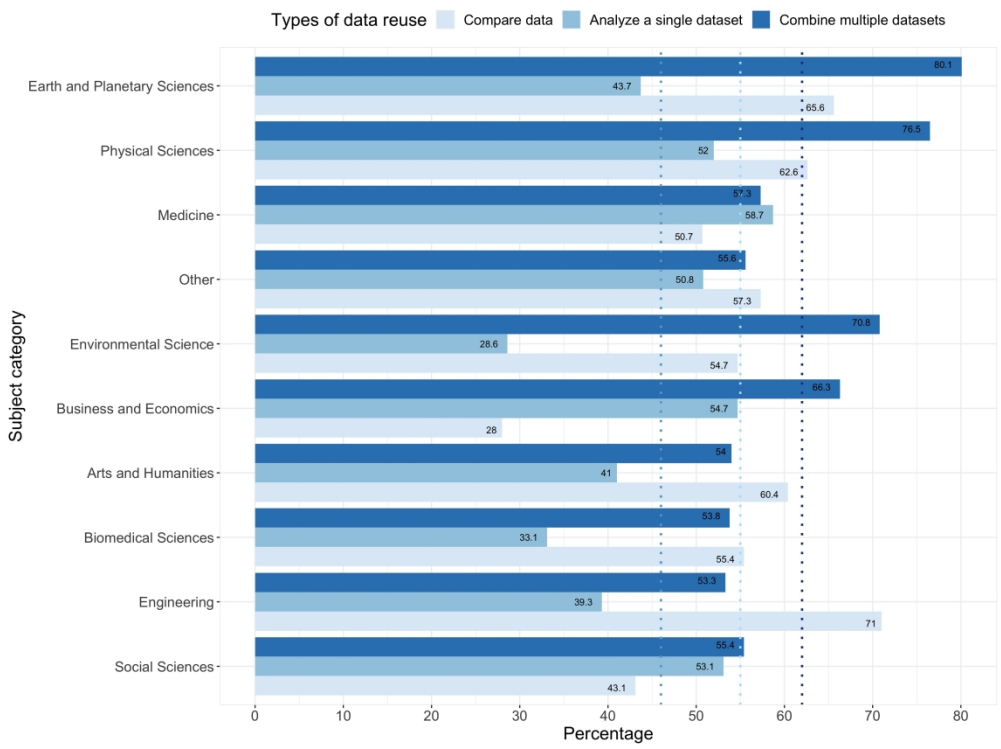

Figure 4 Data reuse types in different subject areas (dotted lines represent the average percentage in each area)

1333x999mm (72 x 72 DPI)

Table I: Selected subject areas and disciplines for the survey and number of responses

Subject area	*Discipline*	*Scopus subject code*	*Number of responses in each discipline**	*Percentage of responses within broader subject category (%)*
Social Sciences (*n*=733, 22.51%)	Linguistics and Language	3310	72	10%
	Education	3304	114	16%
	Library and Information Sciences	3309	252	34%
	'Other' in Social Sciences		211	29%
Arts and Humanities (*n*=334, 10.25%)	Visual and Performing Arts	1213	64	19%
	Literature and Literary Theory	1208	103	31%
	'Other' in Arts and Humanities		139	42%
Business and Economics (*n*=592, 18.18%)	Business and International Management	1403	193	33%
	Economics and Econometrics	2002	251	42%
	'Other' in Business and Economics		123	21%
Physical Sciences (*n*=220, 6.75%)	Astronomy and Astrophysics	3103	133	60%
	Organic Chemistry	1602	11	5%
	'Other' in Physical Sciences		63	29%
Biomedical Sciences (*n*=264, 8.11%)	Neurology	2808	39	15%
	Pharmacology	3004	46	17%
	'Other' in Biomedical Sciences		145	55%
Medicine (*n*=170, 5.22%)	Radiology, Nuclear Medicine and Imaging	2741	27	16%
	Infectious Diseases	2725	38	22%
	'Other' in Medicine		97	57%
Environmental Sciences (*n*=324, 9.95%)	Ecology	2303	147	45%
	Pollution	2310	108	33%
	'Other' in Environmental Sciences		56	17%
Earth and Planetary Sciences (*n*=188, 5.77%)	Geology	1907	56	30%
	Oceanography	1910	53	28%
	'Other' in Earth and Planetary Sciences		73	39%

Engineering (*n*=179, 5.5%)	Aerospace Engineering	2202	14	8%
	Biomedical Engineering	2204	19	11%
	Environmental Engineering	2305	30	17%
	'Other' in Engineering		113	63%
Other			253	8%

Table II. Data sharing methods in different subject areas

Subject category (n=previously data shared)	Institutional repository	Disciplinary repository	Interdisciplinary repository	Journal supported repository	Personal website	Commonly used repositories
Social Sciences (n=312, 42.6%)	169 (54.2%)	42 (14%)	71 (23%)	66 (21%)	75 (24%)	Academia.edu, Zenodo, DANS, ICPSR, Figshare
Arts and Humanities (n=157, 47%)	105 (66.9%)	11 (7%)	21 (13%)	34 (22%)	53 (34%)	Academia.edu, Google drive, Zenodo, Mendeley, OSF
Business and Economics (n=193, 32.6%)	89 (46%)	16 (8%)	18 (9%)	75 (39%)	62 (32%)	Data in Brief, Figshare, ICPSR, Dataverse, American Economic Association
Physical Sciences (n=160, 72.7%)	77 (48%)	44 (28%)	31 (19%)	58 (36%)	59 (37%)	Zenodo, CADC, GitHub, CCDC, NASA databases, SDSS, Sloan digital sky survey, SciFinder
Biomedical Sciences (n=141, 53.4%)	66 (47%)	34 (24%)	37 (26%)	50 (36%)	27 (19%)	DDBJ, OSF, Figshare, GenBank, MRI Image Consortium, NCBI, EMBL, PubMed, PubChem, The Cancer Imaging Archive, GitHub, GEO
Medicine (n=65, 38%)	35 (54%)	15 (23%)	4 (6%)	24 (37%)	6 (9%)	dbGAP, NCBI, GEO, GenBank, Zenodo, Dryad, EGA, IADR, fMRI database, PLoS ONE
Environmental Sciences (n=176, 54.3%)	90 (51%)	40 (23%)	30 (17%)	64 (36%)	22 (13%)	Dryad, GenBank, NCBI, PANGAEA, SeaBass, GitHub, MorphoSource, ForestPlots.NET, NASA, NSF Arctic Data Centre
Earth and Planetary Sciences	86 (65%)	32 (24%)	32 (24%)	33 (25%)	22 (17%)	biorXiv, arXiv, PANGAEA, GitHub, DeepBlue, GIRO,

(n=132, 70.2%)						NASA, NOAA, NCAR, NSF Arctic Data Centre, Zenodo
Engineering (n=75, 42%)	44 (59%)	1 (1%)	11 (15%)	23 (31%)	19 (25%)	Elsevier, Zenodo, Figshare, OSF, GitHub, Mendeley
Chi-square test result	$X^2 = 30.62$, $p = 0.00$	$X^2 = 66.35$, $p = 0.00$	$X^2 = 40.22$, $p = 0.00$	$X^2 = 36.03$, $p = 0.00$	$X^2 = 55.14$, $p = 0.00$	

Table III. How researchers first found repositories to share data

Subject category (n=previously data shared)	*Already aware*	*Search re3data.org*	*Web search*	*Consult with colleagues*	*Consult with experts*
Social Sciences (*n*=312)	174 (55.8%)	8 (3%)	57 (18%)	106 (34%)	80 (26%)
Arts and Humanities (*n*=157)	66 (42%)	1 (0.6%)	31 (20%)	52 (33%)	37 (24%)
Business and Economics (*n*=193)	87 (45%)	2 (1%)	31 (16%)	46 (24%)	32 (17%)
Physical Sciences (*n*=160)	108 (67.5%)	4 (3%)	17 (11%)	47 (29%)	25 (16%)
Biomedical Sciences (*n*=141)	86 (61%)	1 (0.7%)	26 (18%)	51 (36%)	24 (17%)
Medicine (*n*=65)	26 (41%)	3 (5%)	11 (17%)	24 (38%)	15 (24%)
Environmental Sciences (*n*=176)	87 (49%)	2 (1%)	27 (15%)	62 (35%)	32 (18%)
Earth and Planetary Sciences (*n*=132)	63 (48%)	2 (2%)	18 (14%)	52 (39%)	28 (21%)
Engineering (*n*=75)	34 (45%)	1 (1%)	22 (29%)	29 (39%)	12 (16%)
Chi-square test result	$X^2 = 38.18$, $p < 0.001$	$X^2 = 8.13$, $p = 0.42$	$X^2 = 15.61$, $p = 0.048$	$X^2 = 13.35$, $p = 0.1$	$X^2 = 13.05$, $p = 0.11$

Table IV. Future improvements needed to promote data sharing and reuse

Category	Theme	Recommendations
Data related issues	Availability of data	Increased data sharing with available code (where applicable)
		Data is easily available and accessible with a DOI
	Handling of data	Better data management during research lifecycles
	Data citation	Formalize data citation to ensure datasets and associated articles are linked
	Usability of data	Data quality - reliable data with adequate documentation and in a standard format supported in individual's discipline
		Publish data paper/ data descriptor articles to enhance the usability of datasets

		Information on usability of data with some case examples (some datasets are produced for a single use)
Technological solutions	Search system	A single trusted portal or federated search system to search across multiple repositories and disciplines
		Enhanced search system with better tagging feature
		User-friendly data repository interfaces with fast data retrieval (for disciplines producing big data)
	New search system feature	A recommendation system for datasets
		Availability of data extraction and analysis-support tools in the same platform used to access data
		Alert system to notify when relevant datasets are made available publicly
Cultural and policy changes	Awareness and acceptance	Readiness, awareness, and acceptance within the scientific community to support secondary data analysis and publish in journals
		Promotion of data and repositories within scientific communities via conferences, webinars, training for early career researchers
	Incentives	Credit data creators/ reward data sharing in a similar way to publishing journal articles
		Create incentives such as data badges, data reuse indicators to promote data reuse
		Increased funding for secondary data analysis projects and to rescue historical data
	Collaboration	Form collaborations between data creators and users and their institutions (in some cases data are not reusable without contextual explanation)
	Guidelines and documentation	Streamlined IRB rules on how to handle qualitative/ medical data to share at the end of research
		Adequate guidelines on how to anonymize qualitative and health data to ensure data privacy
		Adequate legal and copyright information in place to access and reuse data
		Reduce bureaucratic application procedure for data access to avoid extended waiting periods

Supplementary material 1: Survey questionnaire

Data production

Q1: Please indicate your research experience in terms of years (A research career would normally start when a PhD starts).

1. 0-3 years
2. 3-6 years
3. 6-9 years
4. 10+ years

Q2: Please select your main current research area from the options below. If your research area is not included in the list, then please include it under 'Others'.

1. Physical Sciences – a. Astronomy and Astrophysics, b. Organic Chemistry
2. Biomedical Sciences – a. Neurology, b. Pharmacology
3. Social Sciences – a. Linguistics and Language, b. Education, c. Library and Information Science
4. Arts and Humanities – a. Visual and Performing Arts, b. Literature and Literary Theory
5. Earth and Planetary Sciences – a. Geology, b. Oceanography
6. Engineering – a. Aerospace Engineering, b. Biomedical Engineering, c. Environmental Engineering
7. Environmental Science – a. Ecology, b. Pollution
8. Medicine – a. Radiology, Nuclear Medicine and Imaging, b. Infectious Diseases
9. Business and Economics – a. Business and International Management, b. Economics and Econometrics
10. Other (Please specify)

Q3: What type of data do you produce in your research? [Please give specific examples, e.g., survey data, type of samples/ observations] (open text)

Q4: What are the most important formats of data that you produce in your research? [Select all that apply]

1. Text
2. Images
3. Multimedia (Audio/Video)
4. Software/ code
5. Numerical data (Any type of quantitative measurements)
6. None/ not sure
7. Other (Please scpecify)

Data sharing

Q5. Have you ever shared your research data by posting it on the web (e.g., in a data repository)? [If no, then skip to data reuse question 7]

1. Yes,
2. No,
3. I don't know/not sure

Q5a (If yes): How do you usually share your data? [Please select all that apply]

1. Institutional repository (e.g., university repository)
2. Discipline-specific repository (e.g., Inter-university Consortium for Political and Social Research (ICPSR), PANGAEA)
3. Interdisciplinary repository (e.g., Zenodo, UCLA Center for Embedded Networked Sensing (CENS))
4. A journal supported repository (e.g., PLOS ONE)
5. Personal website
6. Other

(i) Please specify repositories other than institutional

(ii) Other repositories

Q6. How did you first find a repository to share your data?

1. I was already aware of the popular/ relevant repositories in my field
2. Searched re3data.org (Registry of Research Data Repositories)
3. Web search
4. Consulted with colleagues or senior researchers
5. Consulted with the experts in my institution, e.g., Research data support services
6. Others (Please specify)

Q7: Which of these factors influence your choice of repositories to share your data from? [Please select all that apply]

1. Discipline norms,
2. Cost,
3. Ease of use,
4. Reputation of the repository,
5. Appropriateness for data type,
6. Data curation services offered,
7. None of the above
8. Other factors [Please specify]

Data Reuse

Q8: Have you ever reused existing datasets created by other people in your research?

1. Yes [Please select the best option that applies]
    a. I use my own primary data but sometimes combine it with data from existing data sources
    b. I never use my own primary data but only ever use data from existing sources (e.g., datasets published in repositories) to answer new research questions
2. I only ever use my own primary data for my research
3. My research doesn't use data
4. I don't know/ Not sure

(Those who answer 'Yes' proceed to next questions. Skip to question 12 for option 2, 3,4)

Q9: How do you find datasets to reuse? [check all that apply]

1. Search disciplinary repositories
2. Search interdisciplinary repositories
3. Web search (e.g., Google Dataset Search)
4. Read relevant papers and then check if the authors shared data
5. By accident – I noticed the dataset (e.g., in the original paper) and decided to use it.
6. I don't know/can't remember
7. Other [Please specify]

Q10: For which purposes do you reuse existing data? [Please select all that apply]

1. Ground truthing: calibrate, compare, confirm (Comparative reuse)
2. Analyze a single existing dataset to answer novel research questions (Integrative reuse)
3. Combine multiple existing datasets to answer novel research questions (Integrative reuse)
4. I don't know/can't remember
5. Other (Please specify)

Q10b:(When selected 1-3) Please describe the type of data and how you used it (open text)

Measuring data reuse

Q11: Would you like to know whether someone else has reused your published data?

1. Yes,
2. No,
3. Not sure

Q12: Do you ever actively promote your published datasets?

1. Yes,
2. No
3. Not sure

Q12 (If Yes): How do you promote your datasets?

1. In classrooms
2. Using social media platforms – (i) Twitter, (ii) Facebook, (iii) Blog posts
3. Promote within research groups and collaborators' channels
4. Other (specify)

Incentive

We are investigating the type of incentives that can improve the search experience and usage of research data. The following questions identify the factors that may assist in such decision-making process.

Q13: When searching for existing datasets in a repository, which of the following factors you consider important for the decision to use one? [Please select all that apply]

1. Proper documentation for the dataset
    a. Type of data
    b. Subject of data
    c. Data collection method
    d. Other (Please specify)

2. The data is open (no application procedure)
3. Information on the usability of the data
4. Evidence that the data is from an associated publication
5. The data is in a universal standard format
6. Evidence that the data has been reused
7. Other (Please specify)
8. Not applicable

Q14: How easy is it for you usually to find relevant datasets for reuse? [Likert scale]

1. Extremely, 2. Very, 3. Neutral, 4. Difficult, 5. Very difficult or often impossible, 6. I don't know/does not apply

Q15: What can be improved in current systems to encourage and promote data sharing and reuse? (open-ended)

Following definitions were added as reference:

Research data: Any information that has been collected, observed, generated, or created to answer novel research questions and validate research findings. Data may include any form of raw data, multimedia files, such as images, audio, video, codes, and software.

Dataset: A single file or a collection of data produced as a part of research and its associated metadata, such as an abstract, license, and any other relevant information that enables understanding and usage of the data in a legal way.

Data repository: A data repository or data archive is a web-based infrastructure that hosts data in a secure manner and provides long term access to data. A repository can be a part of an academic institution or hosted independently (e.g., Zenodo).

Data reuse: Any secondary use of data by users other than the data collectors.

Supplementary material 2: Tables

Table A: Top 10 types of data produced in different subject areas with the term frequencies

Subject categories	*Data types*
Social Sciences	Survey (481), interviews (234), observations (113), qualitative (98), transcripts (52), quantitative (44), audio (38), video (36), recordings (34), experimental (34)
Arts and Humanities	Survey (60), observations (43), texts (37), images (33), interviews (33), video (28), audio (26), qualitative (24), literary (23), historical (22)
Business and Economics	Survey (345), secondary (83), interviews (67), observations (31), qualitative (26), experimental (18), financial (16), quantitative (16), economic (15), time series (14)
Environmental Sciences	Survey (105), samples (54), observations (48), water (35), field data (30), experimental (29), measurements (23), images (22), soil (22), species (19)
Earth & Planetary Sciences	Models (40), Observations (32), Samples (25), Survey (25), Measurements (24), Water (17), Field data (16), Numerical (16), Chemical (15), Temperature (12)
Biomedical Sciences	Survey (45), images (32), behavioral (30), samples (26), experimental (21), imaging (19), recordings (13), clinical (12), EEG (12), brain (11)
Medicine	Survey (60), Clinical (35), Observations (26), Images (21), Imaging (11), Qualitative (11), Medical (10), Samples (10), Measures (9), Trials (8)
Physical Sciences	Images (50), observations (48), simulations (45), spectra (36), survey (27), software (24), astronomical (22), numerical (18), catalogues (13), physical objects (13)
Engineering	Survey (33), Experimental (27), Simulations (26), Numerical (16), Images (15), Samples (15), Observations (12), Software (12), Measurements (10), system (8)

Table B: Logistic regression of data sharing outcomes (*n*=3,098)

Predictor	*Estimate (β)*	*Std. Error*	*z-value*	*p-value*
Intercept	-0.15238	0.18822	-0.810	0.4182
Biomedical Sciences	0.10910	0.17032	0.641	0.5218
Business and Economics	-0.69173	0.14524	-4.763	1.91e-06 ****
Earth and Planetary Sciences	0.89144	0.20112	4.432	9.32e-06 ****
Engineering	-0.25182	0.19368	-1.300	0.1935

Environmental Sciences	0.27336	0.16406	1.666	0.0957 *
Medicine	-0.46562	0.19683	-2.366	0.0180 **
Other	-0.18376	0.17283	-1.063	0.2877
Physical Sciences	1.07162	0.19778	5.418	6.02e-08 ****
Social Sciences	-0.26114	0.13775	-1.896	0.0580 *
10+ years	0.30144	0.16074	1.875	0.0607 .
3-6 years	-0.08549	0.18425	-0.464	0.6426
6-9 years	0.11700	0.17991	0.650	0.5155
Overall model evaluation				
Likelihood ratio test	$X^2 = 15.225$	df = 13		0.00163**
Significance codes: 0 '****'; 0.001 '***'; 0.01 '**'; 0.05 '*'				

Table C: Factors that influence choice of repositories in different subject areas

Subject category (n=previously data shared)	*Disciplinary norms*	*Cost*	*East of use*	*Reputation of a repository*	*Appropriateness for data type*	*Data curation services offered*
Social Sciences (*n*=312)	145 (46.5%)	104 (33.3%)	178 (57.1%)	156 (50%)	124 (39.7%)	48 (15%)
Arts and Humanities (*n*=157)	67 (43%)	54 (34%)	86 (55%)	73 (46%)	57 (36%)	18 (11%)
Business and Economics (*n*=193)	79 (41%)	56 (29%)	87 (45%)	79 (41%)	54 (28%)	17 (9%)
Physical Sciences (*n*=160)	66 (41%)	65 (41%)	98 (61%)	74 (46%)	73 (46%)	27 (17%)
Biomedical Sciences (*n*=141)	60 (43%)	61 (43%)	79 (56%)	77 (55%)	78 (55%)	23 (16%)
Medicine (*n*=65)	21 (32%)	23 (35%)	28 (43%)	35 (54%)	27 (42%)	9 (14%)
Environmental Sciences (*n*=176)	63 (36%)	66 (38%)	87 (49%)	73 (41%)	71 (40%)	27 (15%)
Earth and Planetary Sciences (*n*=132)	49 (37%)	57 (43%)	68 (52%)	50 (38%)	53 (40%)	20 (15%)
Engineering (*n*=75)	24 (32%)	26 (35%)	41 (55%)	41 (55%)	27 (36%)	5 (7%)

Chi-square test result	X^2 = 12.87, p = 0.16	X^2= 18.33, p = 0.03	X^2 = 16.38, p = 0.06	X^2 = 17.27, p = 0.04	X^2= 31.2, p < 0.001	X^2= 11.59, p = 0.24

Table D: Logistic regression of data reuse outcomes (*n*=3,095)

Predictor	*Estimate (β)*	*Std. Error*	*z-value*	*p-value*
Intercept	-0.08281	0.11754	-0.704	0.48114
Biomedical Sciences	0.12203	0.17178	0.710	0.47748
Business and Economics	0.40240	0.14505	2.774	0.00553 ***
Earth and Planetary Sciences	1.76788	0.23697	7.460	8.62e-14 ****
Engineering	0.56598	0.19368	2.891	0.00383 ***
Environmental Sciences	0.54451	0.16520	3.296	0.00098 ****
Medicine	-0.08834	0.19592	-0.451	0.65206
Other	0.14091	0.17444	0.808	0.41919
Physical Sciences	1.71484	0.21903	7.829	4.91e-15 ****
Social Sciences	0.04842	0.13982	0.346	0.72912
Overall model evaluation				
Likelihood ratio test	X^2 = 5.96	df = 10		0.1136
Significance codes: 0 '****'; 0.001 '***'; 0.01 '**'; 0.05 '*'				

Table E: Comparison between data sharing and reuse across different subject areas

Subject category	*Previously shared data*	*Previously reused data*	*Chi-square test results (Data sharing vs reuse)*
Social Sciences	312 (43%)	343 (47%)	X^2 = 34.594, p < 0.001
Arts and Humanities	157 (47%)	139 (42%)	X^2 = 7.839, p = 0.02
Business and Economics	193 (33%)	329 (56%)	X^2 = 19.175, p < 0.001
Physical Sciences	160 (73%)	179 (81%)	X^2 = 13.559, p < 0.001
Biomedical Sciences	141 (53%)	130 (49%)	X^2 = 35.29, p < 0.001
Medicine	65 (38%)	75 (44%)	X^2 = 0.008, p = 0.93
Environmental Sciences	176 (54%)	192 (59%)	X^2 = 10.737, p = 0.001
Earth and Planetary Sciences	132 (70%)	151 (80%)	X^2 = 4.813, p = 0.03
Engineering	75 (42%)	107 (60%)	X^2 = 2.082, p = 0.149

Table F: How researchers find datasets to reuse in different subject areas

Subject category (n=previously reused data)	*Search disciplinary repositories*	*Search inter-disciplinary repositories*	*Web search (e.g., Google Dataset Search)*	*Read relevant papers*	*By accident*
Social Sciences (*n*=343)	148 (43.1%)	96 (28%)	164 (47.8%)	177 (51.6%)	66 (19%)
Arts and Humanities (*n*=139)	66 (47%)	49 (35%)	84 (60%)	81 (58%)	38 (27%)
Business and Economics (*n*=329)	142 (43.2%)	76 (23%)	170 (51.7%)	181 (55%)	54 (16%)
Physical Sciences (*n*=170)	118 (69.4%)	22 (13%)	62 (36%)	142 (83.5%)	30 (18%)
Biomedical Sciences (*n*=130)	58 (45%)	35 (27%)	44 (34%)	79 (61%)	28 (22%)
Medicine (*n*=75)	28 (37%)	13 (17%)	23 (31%)	39 (52%)	9 (12%)
Environmental Sciences (*n*=192)	69 (36%)	40 (21%)	77 (40%)	114 (59.4%)	24 (13%)
Earth and Planetary Sciences (*n*=151)	75 (50%)	33 (22%)	71 (47%)	115 (76.2%)	17 (11%)
Engineering (*n*=107)	39 (36%)	25 (23%)	67 (63%)	73 (68%)	17 (16%)
Chi-square test result	$X^2 = 46.94$, $p < 0.001$	$X^2 = 31.38$, $p < 0.001$	$X^2 = 55.41$, $p < 0.001$	$X^2 = 63.12$, $p < 0.001$	$X^2 = 21.36$, $p = 0.01$

Supplementary material 3: Figures

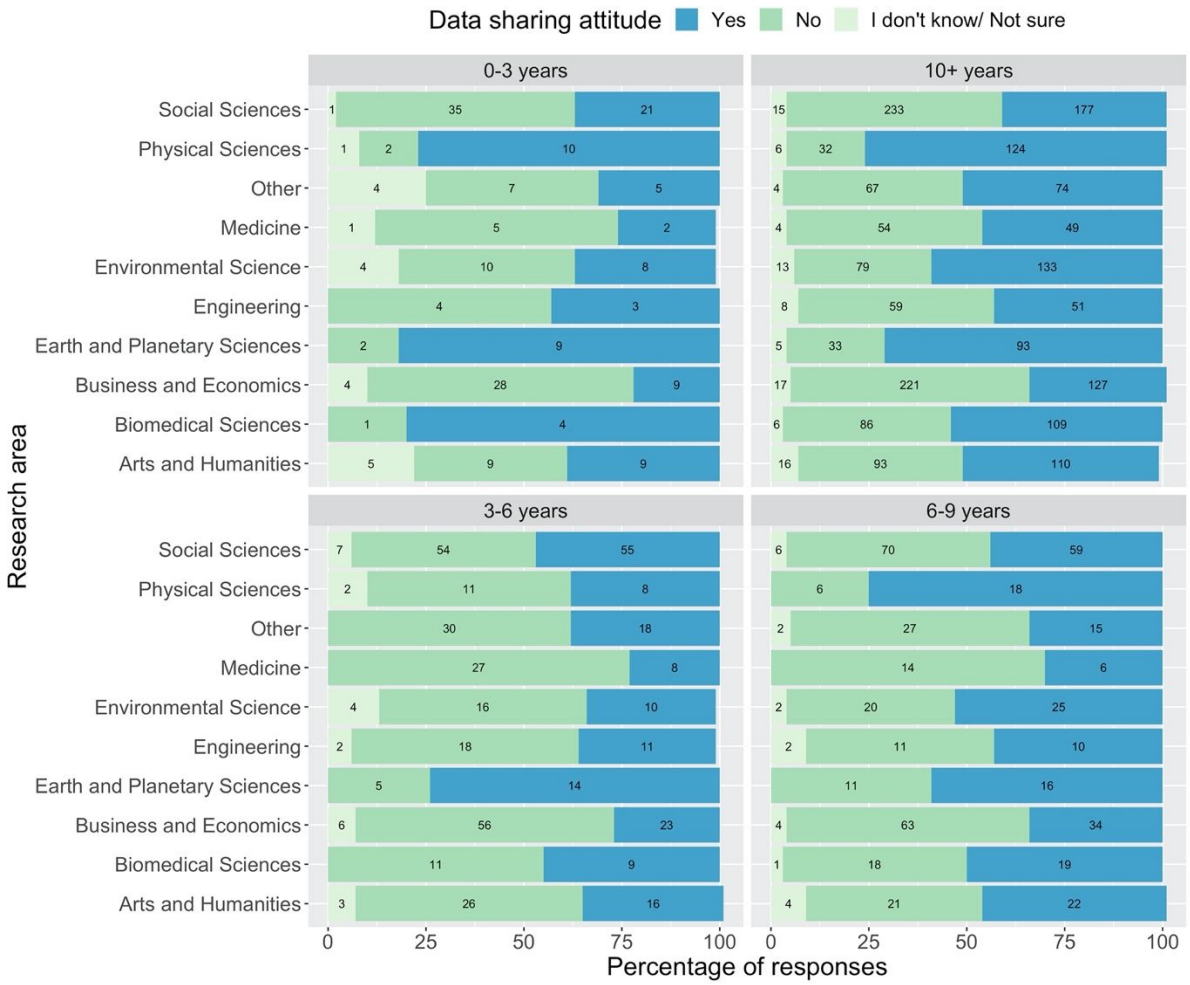

Figure A. Data sharing in groups with different research experiences across subject areas (labels on bars represent numbers of responses)

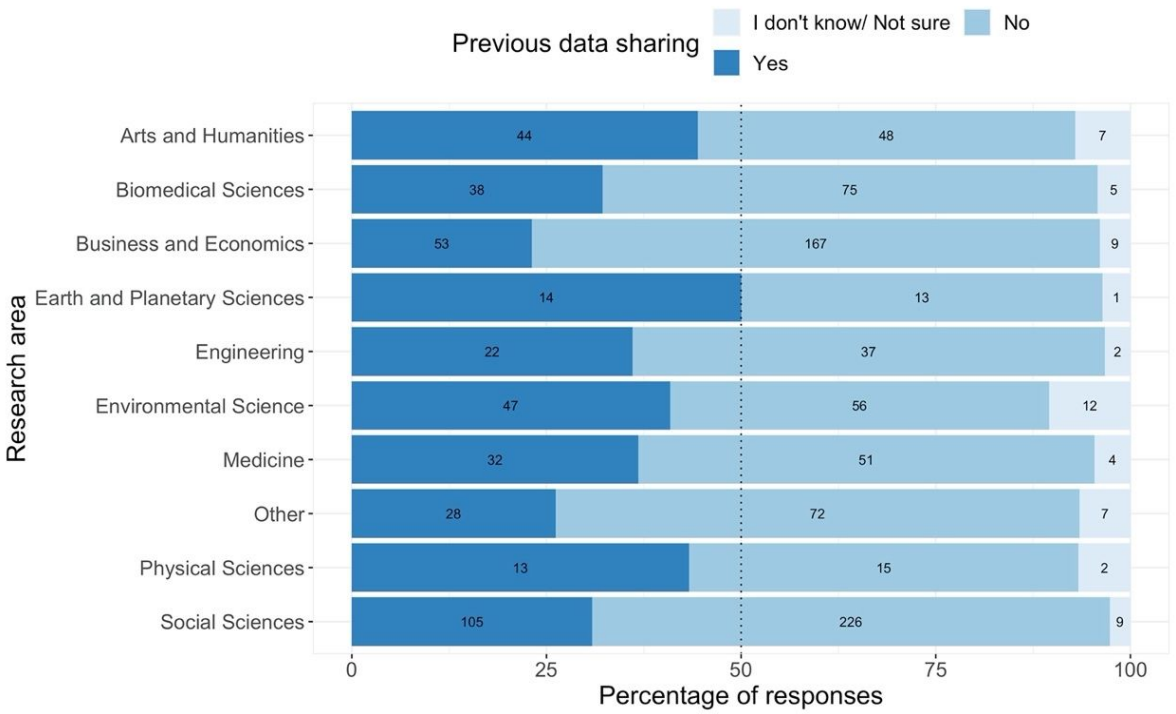

Figure B: Data sharing among those who only use own primary data (labels on bars represent numbers of responses)

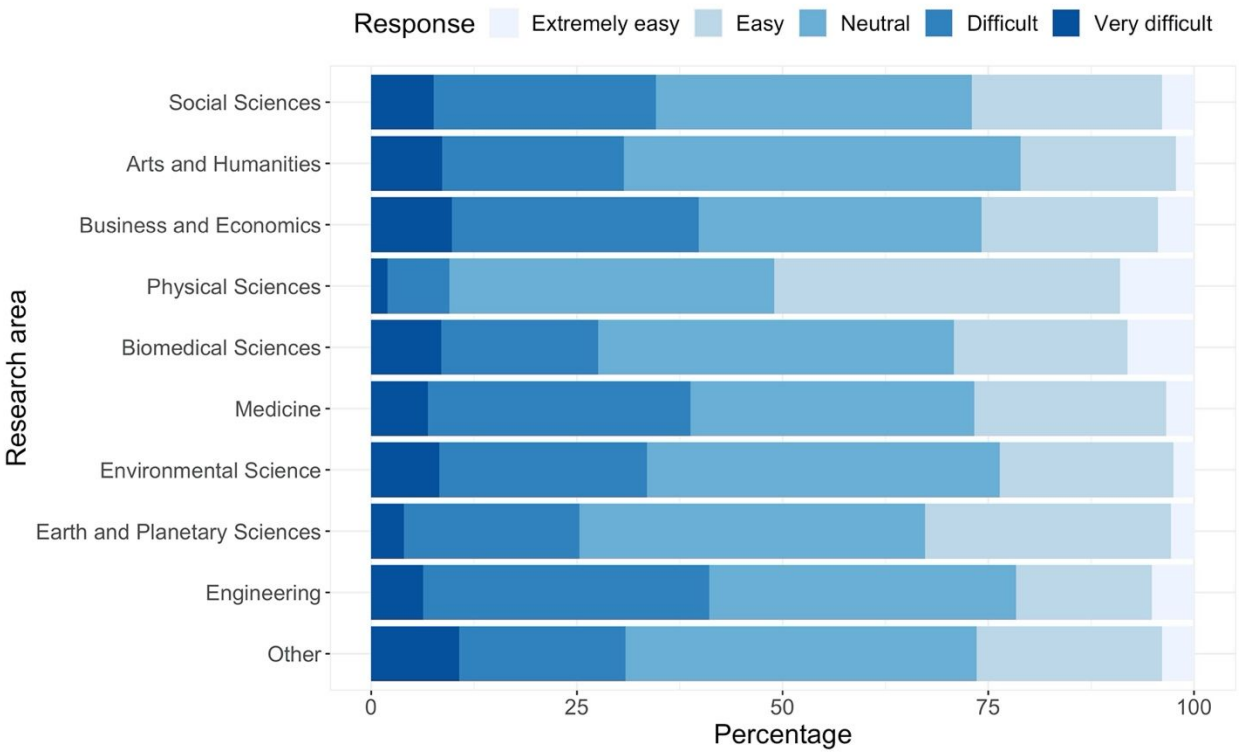

Figure C: Ease of finding datasets to reuse by subject areas

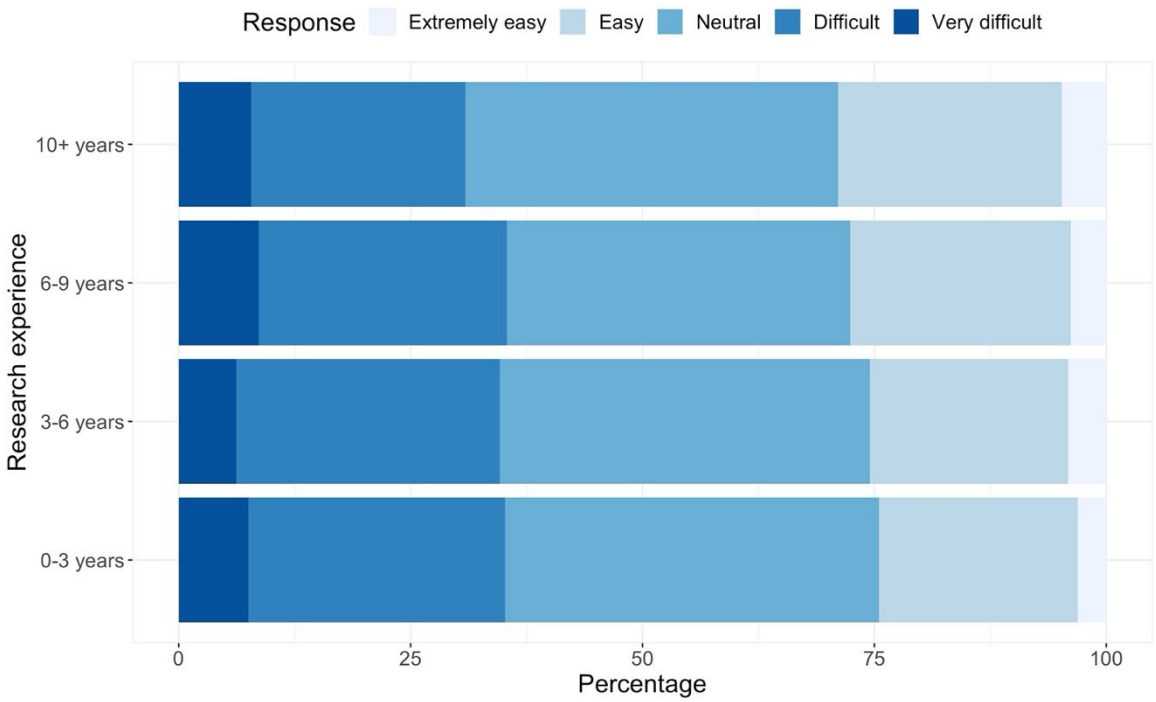

Figure D: Ease of finding datasets to reuse by research experience