SPRINGER NATURE

springernature.com

Tracking glacier response to climate change

Research Data

PRACTICAL CHALLENGES FOR RESEARCHERS IN DATA SHARING

ADVANCING DISCOVERY

White paper

BMC Springer nature research palgrave macmillan

Contents

Foreword 4
Executive summary 6
Key findings. 6
Subject differences 7
Regional differences 7
Discussion and Springer Nature perspectives 8
Limitations 8
Data underlying this whitepaper 8
Part 1: The Data Sharing Landscape 9
The data sharing status quo. 10
What are the challenges to sharing data? 11
Part 2: Data Sharing Norms and Challenges by Subject Area. 14
Subject differences in the size of the challenge. 14
Subject area differences in the challenges in sharing data 15
Biological sciences 17
Earth sciences. 17
Medical sciences 18
Physical sciences 19
Part 3: Regional Differences in Data Sharing Challenges and Norms . . 20
Subject influences on regional behaviour 23
Regional differences in the challenges in data sharing 24
Discussion and Springer Nature Perspectives. 25
Increased data management, support and education 26
Faster, easier routes to optimal ways of sharing data 27
Future research 27
Appendix 28
References 29

Authors
David Stuart, Stuart Information Research
Grace Baynes, Springer Nature
Iain Hrynaszkiewicz, Springer Nature
Katie Allin, Springer Nature
Dan Penny, Springer Nature
Mithu Lucraft, Springer Nature
Mathias Astell, Springer Nature

March 2018

This whitepaper and its underlying survey data have been made openly available in the Figshare repository.

Access whitepaper: https://doi.org/10.6084/m9.figshare.5975011

Access full survey dataset: https://doi.org/10.6084/m9.figshare.5971387

Foreword

We are in the midst of progress, and potentially exciting change, for open science and open access to research data. The world's funders are increasingly mandating good data practice, including data management plans and data sharing, and recognising the need for global collaboration on infrastructure and best practice. Across the research community, momentum is gathering in policy, strategy and working groups to achieve a future where research data are widely Findable, Accessible, Interoperable and Reusable (FAIR).

Open science should be about opening up all areas of research. Open access to research data can help speed the pace of discovery and deliver more value for funded research by enabling reuse and reducing duplication. The evidence is there that open data and good data management makes research studies more productive, more likely to be cited and unlocks innovation for the good of society including unexpected new discoveries and economic benefit.

Researchers do not need to be convinced further of these benefits, based on reported attitudes. Numerous studies show strong recognition of the benefits of data sharing, along with high motivation levels to share data and use that of others. Yet in 2017 only about half of research data were shared (according to surveys of researchers) and a much smaller proportion were shared openly or in ways that maximise discoverability and reuse. Why is this the case, when researchers' hearts and minds would seem to be in the right place? How do we move from positive attitudes to a change in behaviour where data sharing is the norm?

This survey aims to understand researcher activity around sharing data at a particular point in the research lifecycle – when they are preparing their work for publication. In this it builds on previously published studies that explore data sharing more generally during the research process. It explores attitudes briefly, but focuses on actions and challenges in sharing data. Responses from over 7,700 researchers enabled us to draw new insights across subject fields and, to a lesser extent, across geographies.

We find that researchers are sharing data associated with published works both in repositories and as supplementary information. Sharing data as supplementary information helps make it more available than if it were not shared at all, but does not optimise discoverability, accessibility or reuse. Springer Nature will continue our efforts to encourage sharing in repositories, and support data citations in published works.

Our findings confirm that researchers' efforts to archive, publish and share data continue to be hampered by time constraints and a lack of knowledge around data standards, metadata and curation expertise, repository options, and funder requirements. Subject and regional differences do exist, suggesting where targeted approaches may be helpful. But there are common global challenges that require concerted attention: the provision of more education and support for researchers, and faster, easier routes to sharing data optimally.

As with all market research of this kind, the survey results pose as many new questions as they answer. We offer one of the largest datasets exploring data sharing behaviours in Europe and North America, but the survey has relatively small sample sizes from Asia, Africa and South America. More research is needed to understand data sharing behaviours across the Global South and Asia-Pacific, and in specific fields, particularly in the humanities, social sciences and engineering. Springer Nature will undertake further research this year to understand data sharing activity in China and Japan. In the spirit of collaboration, and hoping that others will glean further insights from this data, the anonymised results are freely available on Figshare under a Creative Commons license.

Springer Nature is committed to supporting good research data management and data sharing. We want to help researchers adopt open approaches to their data wherever possible and to partner with funders, institutions and community initiatives to make it happen. By working together, we can unlock the huge potential of open data to improve our knowledge, the global economy, our health and environment. Our goal in sharing this further insight on the specific issues that are holding the research community back is to help build the case for the concrete actions needed now.

Grace Baynes
VP, Research Data & New Product Development, Open Research, Springer Nature

Executive summary

In one of the largest surveys about research data, we found widespread data sharing associated with published works and a desire from researchers that their data are discoverable. The survey confirms and extends recent findings on general data sharing attitudes and behaviour, including those published in the The State of Open Data 2017[i] report from Digital Science, to which Springer Nature contributed. The research presented here also reinforces previous findings on the challenges faced by researchers in sharing their data. The size of the results (with over 7,700 researchers responding) allowed us to explore these behaviours from regional and subject perspectives.

Key findings:

- When asked what they do with the data files generated by their research when submitting to a journal, 63% of respondents stated that they generally submit data files as supplementary information, deposit the files in a repository, or both.
- 76% of researchers rated the importance of making their data discoverable highly – with an average rating of 7.3 out of 10 and the most popular rating being 10 out of 10 (25%).
- The main challenge to data sharing was identified by respondents as 'Organising data in a presentable and useful way' (46%), with other challenges generally rated:
  - 'Unsure about copyright and licensing' - 37%
  - 'Not knowing which repository to use' - 33%
  - 'Lack of time to deposit data' - 26%
  - 'Costs of sharing data' - 19%
- This survey adds to previous research on how data sharing behaviour and challenges differ by subject and region[ii], finding that lack of time is a greater concern to researchers in Europe, North America and Australia, while costs of sharing data is recognised as more of a concern by those in Asia and South America.
- Similarly, there is a difference between how much time and knowledge is an issue depending on the seniority of researchers. When asked about barriers to data sharing, time is a bigger issue with more senior researchers (29% for most senior versus 23% of early career researchers), while 40% of early career researchers cite not knowing where to share data as a problem versus 30% for the most senior researchers; uncertainty about copyright and licensing is cited by 43% for early career researchers versus 33% for the most senior researchers.
- Concerns about cost stay reasonably low as a stated factor throughout different career stages (ranging between 18-20%), whereas concerns about organising data in a presentable and useful way stay high throughout (ranging between 48-49%).
- The size of datasets also has an impact on whether data are shared – respondents that generate the smallest data files (<20MB; n = 2,036) have the highest proportion of data that are neither shared as supplementary information nor deposited in a repository (42%). In contrast, 70% of those with data files greater than 50GB (n = 700) share their data, with a strong preference for sharing through repositories (59%).

Subject differences:

- The discoverability of data is rated as most important in the biological sciences (7.8 out of 10), followed by the Earth sciences (7.7), medical sciences (7.2), and physical sciences (6.6), which correlates with data sharing behaviour in these areas.
- The biological sciences had the highest proportion of respondents who share data relating to publications (75%), followed by the Earth sciences (68%), medical sciences (61%), and physical sciences (59%).
- Perceived barriers to data sharing also differs between subject areas. The problem of organising data in a useful way varied from 57% in the physical sciences to 40% in the medical sciences; copyright and licensing ranged from 44% in the medical sciences to 31% in the physical sciences; and not knowing which repository to use ranged from 37% in the medical sciences to 27% in the physical sciences.
- Even within subject communities where there are established norms for data sharing (in the form of funder mandates and the availability of community repositories), the survey shows a lack of awareness where data sharing is concerned. Only 54% of respondents who produce specific biological and medical data (e.g. DNA and RNA sequences), where dedicated community repositories exist, are using these repositories to share their data.

Regional differences:

- Data challenges differ considerably between regions, with respondents in Asia and South America reporting a higher level of data sharing than those in Europe, North America, and Australasia. This correlates with overall regional trends noted in The State of Open Data Report 2017 and elsewhere, but care should be taken with these findings as sample sizes in Asia and South America were small (n = 359 and n = 137, respectively) in comparison to North America (n = 2,215) and Europe (n = 4,692) and may represent a self-selecting, data-interested group. Further research is needed, particularly in China and Japan, to understand data sharing practice in more detail.
- In Asia, 77% of respondents shared data as supplementary information or in a repository when submitting a manuscript, compared to 67% in South America and Europe, 54% in North America, and 51% in Australasia. This largely correlates with the perceived importance of data discoverability in the different regions: Australasia, 6.9 out of 10; North America, 7.2; Europe, 7.3; Asia, 7.6; and South America, 7.7. Africa is omitted from the regional analysis due to the small sample size (n=65).
- 'Organising data in a presentable and useful way' is the most often stated reason for not sharing data in all regions: South America, 53%; North America, 52%; Europe, 44%; Asia, 43%; Australasia, 43%.
- The biggest regional variation in the barriers to data sharing is the trade-off between time and money. Time tended to be a bigger barrier for respondents from Australasia, North America and Europe (with up to 28% of researchers citing this as a barrier) when compared to South America and Asia (where as few as 20% cited it). Cost was perceived as a bigger barrier in South America and Asia (where up to 25% of researchers cited it as a barrier) in comparison to Australasia, North America and Europe (where as few as 17% cited it).

Discussion and Springer Nature perspectives:

This survey aimed to explore the practicalities of data sharing at a particular stage in the research lifecycle – during the publication process of journal articles. Our goal was to understand the status quo, so that Springer Nature and other stakeholders can continue to take practical steps to facilitate data sharing and good data practice.

The results suggest two areas of focus that could increase the sharing of data amongst researchers, regardless of subject specialism or location:

1. Increased education and support on good data management for all researchers, but particularly at early stages of researcher's careers. This should include readily available advice and support about good data practice, awareness-raising about the availability of repositories, and understanding of copyright and licensing of research data.

Awareness and attitudes to the benefits of data sharing were not directly addressed in this survey, as they have been investigated in previously published research. The survey does support and highlight key challenges noted in previous research, including: the widespread uncertainty about copyright and licensing; the problem of not knowing which repository to use (both of which were particularly seen as problems for early career researchers); along with the widespread concerns about organising data in a presentable and useful way.

2. Faster, easier routes to optimal ways of sharing data. The challenges of organising data, and the lack of time to do so, require readily available ways to organise and share data, which are easily accessible and usable by researchers.

In both cases, solutions require continued collaboration between researchers, institutes, funders, publishers, repositories and other research data infrastructure and service providers.

Limitations

Although the survey is one of the largest into data sharing behaviour, there is nonetheless a need for further investigation, especially in those regions where there were a limited number of responses: Africa, South America, Australasia, and Asia. In particular China, Japan and India are all major producers of research where there were not sufficient responses to be analysed at a country level. It is encouraging to see responses from the Global South including South America and Africa, but again small sample sizes preclude detailed analysis or conclusions.

Data underlying this whitepaper

The survey methodology, response rate and number of respondents are detailed in the Appendix of this report. The anonymised data will be freely available through the Figshare repository under a Creative Commons license. You can access the full dataset here: https://doi.org/10.6084/m9.figshare.5971387

1. The Data Sharing Landscape

Data sharing is increasingly required by funders[iii] and publishers[iv] to increase the return on investment, reuse and reproducibility of research, and in recent years there has been an increase in researchers' willingness to share data and embrace data sharing practices.[v,vi]

For example, the State of Open Data Report, published in October 2017, presented the second year of a longitudinal survey tracking researchers' data sharing behaviours and attitudes. It found a year-on-year increase in awareness of open datasets, and an increase in researchers' willingness to share. This supports earlier research exploring data sharing behaviours.[vii]

Policies that require or encourage data sharing are increasingly being adopted by research funders. To support this and encourage good practice, publisher actions have included the publication of data availability statements with research articles.[viii] In 2014, PLOS was one of the first publishers to introduce a data sharing mandate for publications in its journals. PLOS has reported a growing acceptance of data sharing in response to this policy, with few researchers refusing to share data.[ix] Springer Nature and many other publishers now have journal data policies that require or recommend data availability statements and data sharing.

Whilst previous research has identified peer to peer sharing as the most common route to making data available,[x,xi] the recent The State of Open Data report found that the most common method of data sharing was as an appendix to an article (just over 30% of respondents). To explore this finding further, the survey that formed the basis of this report asked respondents specifically about their data sharing behaviour at the point of submitting a manuscript, exploring what methods researchers are using to share their data.

Previous research has explored researchers' willingness to share data,[xii] as well as the perceived motivations for data sharing.[xiii] Challenges or barriers to data sharing, global and subject variances have also been considered by previous studies (although not with the sample size of this report) and these have identified both lack of time and funding as key drivers for not sharing data.[xiv] Springer Nature's understanding of the challenges researchers face in sharing data has been enriched by tracking enquiries to Springer Nature's free research data Helpdesk,[xv] and this has informed the research reported here.

The data sharing status quo

The level of importance researchers attach to data sharing can be seen in responses to the survey's question: *How important is it to you that your data are discoverable?* On a scale of 1-10 the most popular rating was 10 (25% of respondents) – the average rating was 7.3 and 76% gave a rating greater than five (see Figure 1).

Overall 63% of respondents stated that when submitting a research manuscript, they generally submitted data files as supplementary information, deposited the files in a repository, or both (see Figure 2), with only 37% doing neither.

However when it comes to how data are shared, a slightly lower proportion of researchers share data in a repository (41%) than they do as supplementary information files (42%), which is consistent with earlier findings from a previous, smaller scale survey.[xvi]

Figure 1: Q - How important is it to you that your data are discoverable? (1 is the least important) (n=7,656)

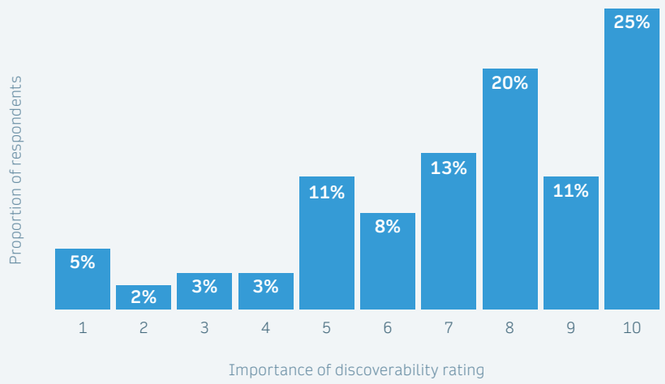
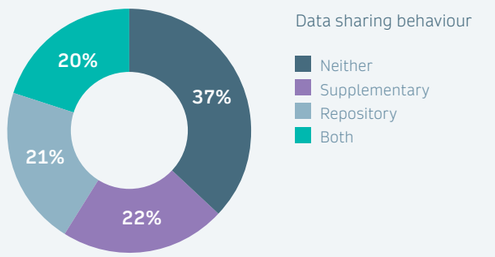


Figure 2: Q - Generally, when submitting a manuscript to a journal what do you do with the data files generated by your research? (n=7,697)



What are the challenges to sharing data?

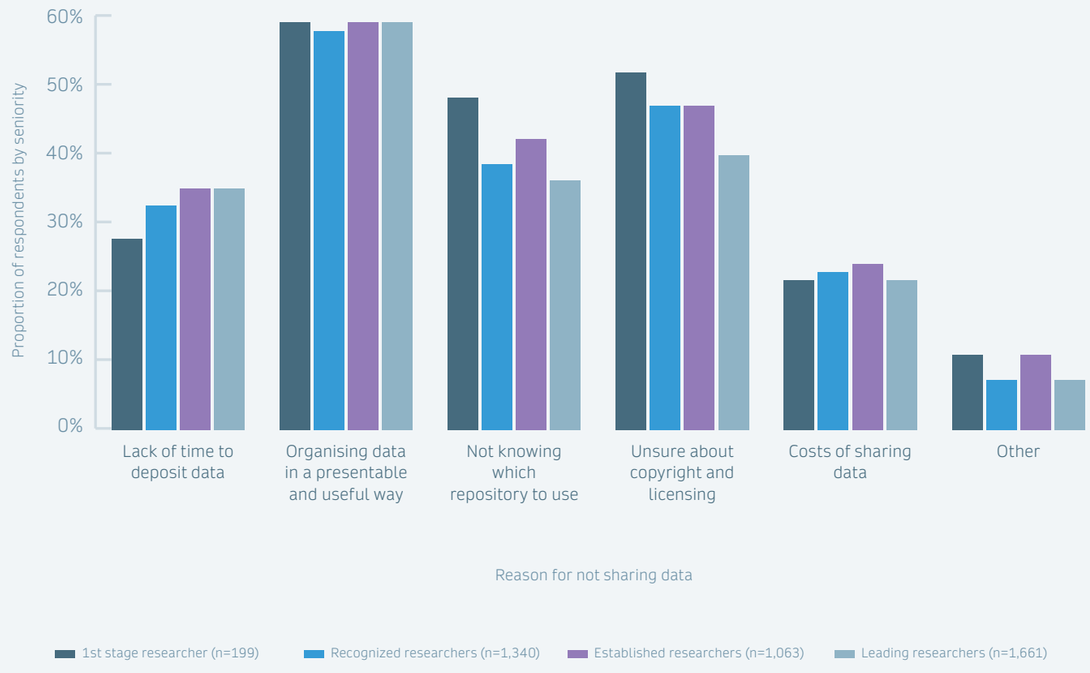
Overall the biggest challenge to data sharing is 'Organizing data in a presentable and useful way', selected by 46% of respondents in response to the question: *What problems do you have in sharing datasets?* This was followed by 'Unsure about copyright and licensing' (37%), 'Not knowing which repository to use' (33%), 'Lack of time to deposit data' (26%), and finally 'Costs of sharing data' (19%).

As may be expected, some problems are more prevalent at different stages of a researcher's career. This can be seen when classifying respondents' job titles where possible using the EURAXESS[xvii] classification of researchers:

- first stage researcher (e.g., PhD student) (n = 199);
- recognised researcher (e.g., research fellow) (n = 1,340);
- established researcher (e.g., senior research fellow) (n = 1,063);
- leading researcher (e.g., professor) (n = 1,661).

Overall, the same challenges are common at every career stage, with only small percentage differences. As can be seen in Figure 3, reporting 'lack of time to deposit data' rises with a researcher's seniority, whereas a lack of knowledge about 'copyright and licensing' is reported more often by early career researchers. A higher proportion of first stage researchers also state not knowing which repository to use is a problem. Although those identified as first stage researchers represent a small proportion of the total number of respondents (n=199), these findings imply that there is not as much understanding or awareness of good data practices at the earlier stages of a researcher's career.

Figure 3: Q - What problems do you have in sharing datasets? (separated by seniority) (n=4,263)



Those who identified 'Other' problems in sharing datasets were invited to specify the nature of the problems. A content analysis of the responses found that they could be broadly categorised into six types of problem: data sensitivity; intellectual property rights; organisational policy and culture; fear of data misuse and being scooped; technical issues (e.g., the large size of the dataset); and data issues (e.g., small datasets perceived as unsuitable for sharing).

Of these six other problems 'data sensitivity' was mentioned most often (see Figure 4). As may be expected, it was identified particularly often amongst those stating their subject related to the medical sciences – who may be dealing with patient data and other data related to human research participants. Data sensitivity accounted for 66% (117 out of 172) of 'Other' problems stated by medical researchers. As respondents noted, in many situations the anonymisation of medical data is difficult ("Difficult to completely anonymise (qualitative data)", "Data are often difficult to deidentify due to studying an uncommon disease in a small community") and data sharing is often restricted by law.

Overall the 'fear of data misuse and being scooped' was raised rarely. It accounted for only 12% (46 out of 385) of all 'Other' problems given. Although fear of misuse is often raised as an issue in the literature[xviii], based on the responses to this survey it seems to be of secondary importance to the more widely reported practicalities of sharing data.

It is notable that our respondents were skewed towards smaller datasets. 5,219 of respondents stated that their datasets were smaller than 1GB, whilst only 1,294 stated that they were bigger:

- <20MB (n = 2,036)
- 20MB-100MB (n= 1,859)
- 100MB-1GB (n = 1,324)
- 1GB-5GB (n = 594)
- 5GB-20GB (n = 280)
- 20GB-50GB (n = 134)
- >50GB (n = 286)

The size of the dataset can also have an impact on whether the data are shared or not (see Figure 5). Researchers that generated the smallest sized data files (<20MB; n = 2,036) had the highest proportion of data that were neither shared as supplementary information nor deposited in a repository (42%), with a clear preference for sharing data only through supplementary material when data are shared. In contrast, 70% of those with data files greater than 50GB (n = 700) shared their data, with a strong preference for sharing through repositories (59%).

This clearly highlights that there is a sense among some researchers that sharing data is about size of data or "big data". Respondents further confirm this in the open text comments provided around problems in data sharing, stating that data are not shared because they are perceived as too small, too easily replicated, or too idiosyncratic to have widespread interest: "Unsure about usefulness in sharing small datasets", "Querying relevance to others", "Mine are easy to replicate experiments, why sharing *(sic)* the data?".

Figure 4: 'Other' problems mentioned in sharing datasets mentioned (n=385)

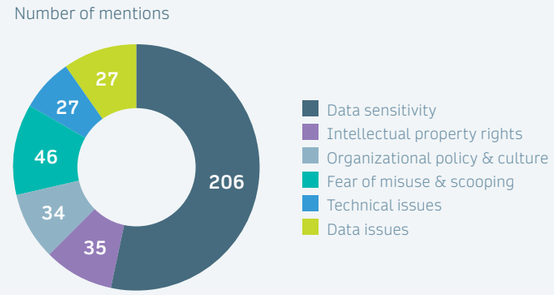
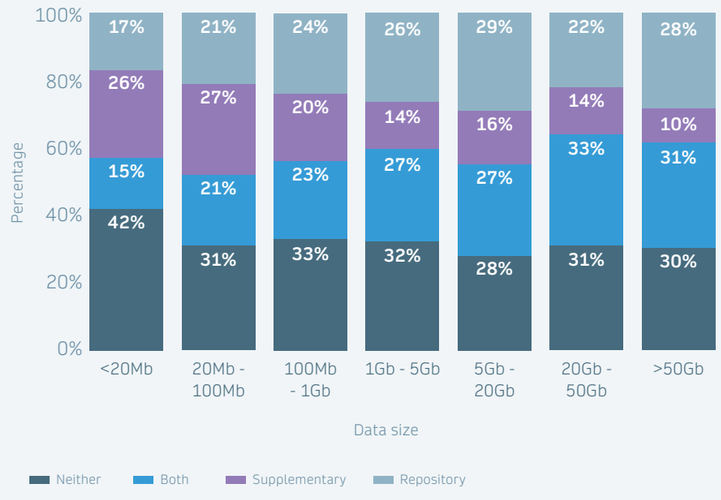


Figure 5: Data sharing behaviour by size of dataset (n=6,513)

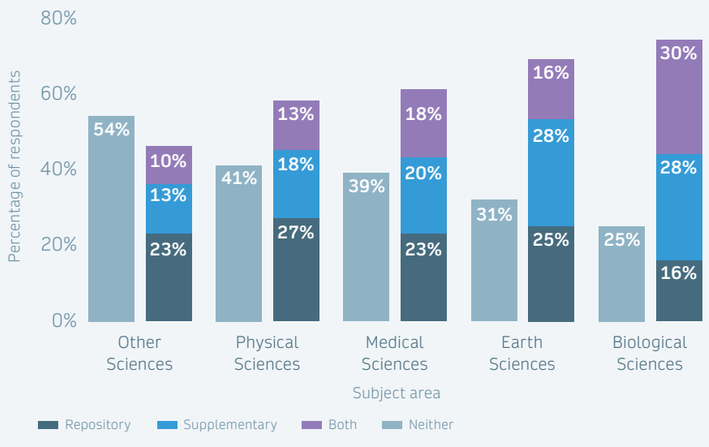


2. Data Sharing Norms and Challenges by Subject Area

Subject differences in the size of the challenge

Data sharing behaviours and challenges differ considerably between subject areas. The proportion of respondents who, on submitting a manuscript, shared their data as supplementary information files or in a repository ranged from 75% in the biological sciences to 59% in the physical sciences. 'Other sciences', where only 46% of respondents said they shared their data through supplementary information or repositories, was the only subject area group not to have a majority of respondents sharing their data in these ways (see Figure 6). This group covers many disparate fields, including the social sciences, computer science, humanities and mathematics, and we would anticipate a wide range of data sharing behaviours between these subject communities. Subject area specialism for "Other Sciences" was not collected in the current survey, precluding specific analysis, which means this is an area that would benefit from further research.

Figure 6: The depositing of data in different subject areas (n=7,664)



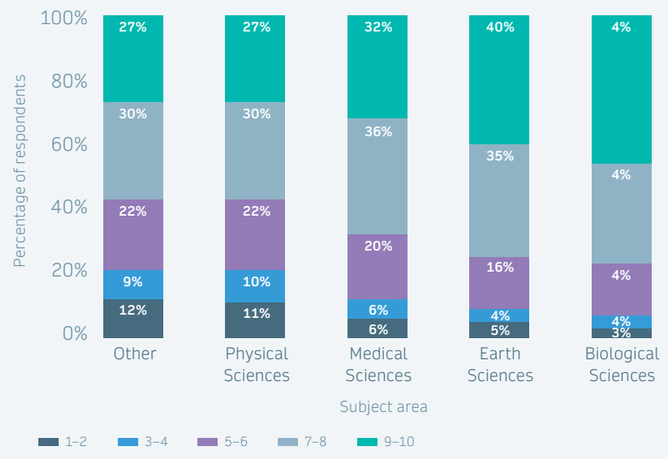
There are also differences between subjects in the way the data are shared when publishing research articles. In the Earth sciences and the biological sciences a greater proportion of datasets are shared as supplementary information files, whereas in the physical and medical sciences a greater proportion of data were shared in a repository.

There are a number of types of biological and medical data that have dedicated community repositories, including: Crystallographic data for small molecules; DNA and RNA sequences; DNA and RNA sequencing data; Genetic polymorphisms; Linked

genotype and phenotype data; Macromolecular structure; Microarray data; and Protein sequences. However, while researchers in these areas share their data more often, they also quite often fail to deposit them in one of these dedicated community repositories. Of the 2,288 respondents in the medical and biological sciences who produced these particular types of data, 83% shared their data when submitting a manuscript, in comparison to 68% for the medical and biological sciences as a whole. However only 54% of those 2,288 researchers always deposited these data in dedicated community repositories.

The relatively high level of data sharing in the biological sciences, and conversely the low level in the physical sciences, is reflected in the perceived importance of data discoverability. On average the respondents in the biological sciences rated the importance of data being discoverable as 7.8, in comparison to 7.7 in the Earth sciences, 7.2 in the medical sciences, and 6.6 in the physical sciences. Figure 9 shows the distributions in the importance that data are discoverable in the different subject areas, grouped in ranges of 2. Whilst overall there is a general recognition that it is important that data are discoverable, it is nonetheless worth noting that it is not universal: 8% of respondents in the physical sciences gave the importance of discoverability a rating of 1 out of 10.

Figure 7: The importance that data are discoverable in different subject areas (1 is the least important) (n=7,626)



Subject area differences in the challenges in sharing data

There was much commonality across subject areas in the challenges of sharing data. The main difference was the order of concerns cited: copyright and licensing was raised most often in the medical sciences, and lack of time was a bigger issue than knowing which repository to use in the physical sciences and Earth sciences. Licensing concerns in the medical sciences are to be expected due to laws protecting patient rights. The small proportion of researchers who see knowing which repository to use as a problem in the physical sciences may reflect that there are a smaller, defined number of recognised repositories in this field than in other subject areas. Less complexity in choice of repositories may contribute to a wider understanding in physical sciences of where researchers can deposit data.[xix]

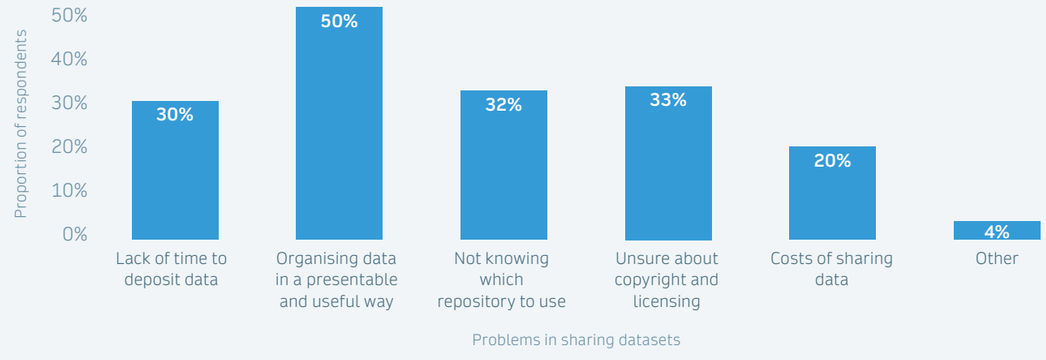
Challenges in organising data varied from 57% in the physical sciences to 40% in the medical sciences, copyright and licensing ranged from 44% in the medical sciences to 31% in the physical sciences, and not knowing which repository to use ranged from 37% in the medical sciences to 27% in the physical sciences. These variations between physical and medical sciences imply that the barriers researchers face in data sharing are linked to the expectation of data usage in each area e.g. medical data are more strictly regulated by data protection rules and issues arising from data reuse and so there is a greater concern around copyright, licensing and access through repositories. Lack of awareness of community mandates/repositories is prevalent even in disciplines where data sharing is common and there are established mandates, notably the biological sciences.

Figure 8: Problems in sharing datasets in different subject areas (n=7,719)

Biological sciences

Researchers in the biological sciences share data most often when submitting manuscripts, 75% of respondents sharing data in either a repository, as supplementary information, or both. Of the 2,640 respondents that stated that their data related to the biological sciences, 30% stated they shared in both repositories and as supplementary information, far higher than in other subject areas.

Figure 9: Problems in sharing datasets in the biological sciences (n=2,640)



Nonetheless, there are still 25% of respondents in the biological sciences that neither submit their data to repositories nor provide it as supplementary information. The biggest problem, reported by 50% of respondents, was 'Organising data in a presentable and useful way', although data sensitivity can also be a significant issue, accounting for 30% (26 out of 86) of the classified 'Other' comments: "Privacy issues related to health related data"; "Improper use of data (especially nest location) of sensitive species"; "Keeping subject data confidential".

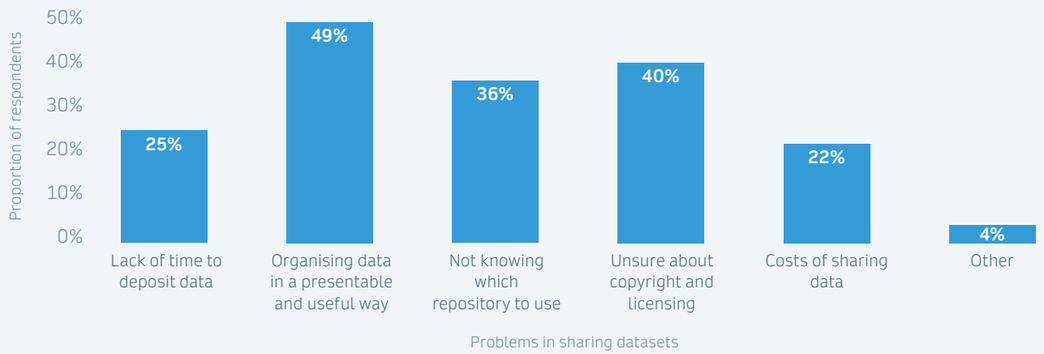
It is notable that although "Not knowing which repository to use" was only the third most popular problem identified (and overall selected by the fewest percentage of respondents compared with other subject groups), 32% is still a high percentage of researchers with a lack of knowledge of repositories for an area of research where data sharing is common and there are established mandates in place.[xx]

Earth sciences

Of the 365 respondents in the Earth sciences, 32% neither share their data as supplementary information nor in a repository when submitting a research manuscript. The most popular way to share data is as supplementary information (28%), sharing in repositories accounted for 25% and only 16% share both as supplementary information and in a repository. Earth sciences researchers rank discoverability of data highly (7.7 out of 10 on average), the second highest in the survey after biological sciences (7.8).

The most often reported barrier to sharing data in the Earth sciences was 'Organising of data in a presentable and useful way', reported by 56% of respondents, followed by 'Unsure about copyright and licensing' (35%), 'Lack of time' (31%), 'Not knowing which repository to use' (30%), and 'Costs of sharing data' (16%). Lack of time was raised more often in the Earth sciences than in any other subject area. Issues surrounding intellectual property rights also accounted for four of the eleven comments in the 'Other' classification: "data donated under constraints and restrictions", "Normally copyright because I don't own the data".

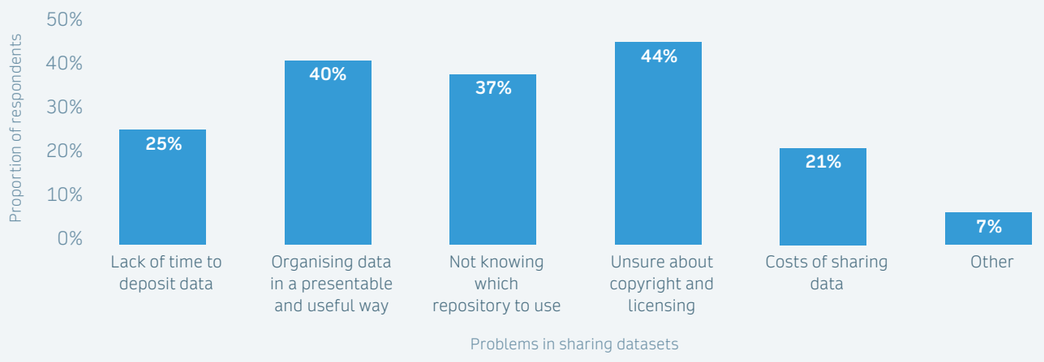
Figure 10: Problems in sharing data in the Earth sciences (n=365)



Medical sciences

Of the 2,683 respondents that reported that their data were in the medical sciences, 39% shared data neither as supplementary information nor in a repository when submitting a manuscript. This is a far higher percentage than the 25% in the biological sciences. The importance ascribed to the discoverability of data was also lower in the medical sciences (7.2) than both the biological sciences (7.8) and the Earth sciences (7.7).

Figure 11: Problems in sharing data in the medical sciences (n=2,683)



Of the four subject areas, the problems facing researchers in the medical sciences differ the most from other areas. The most often cited issue was being unsure about copyright and licensing (44%), followed by organising data (40%), not knowing which repository to use (37%), lack of time (25%) and costs of sharing data (21%).

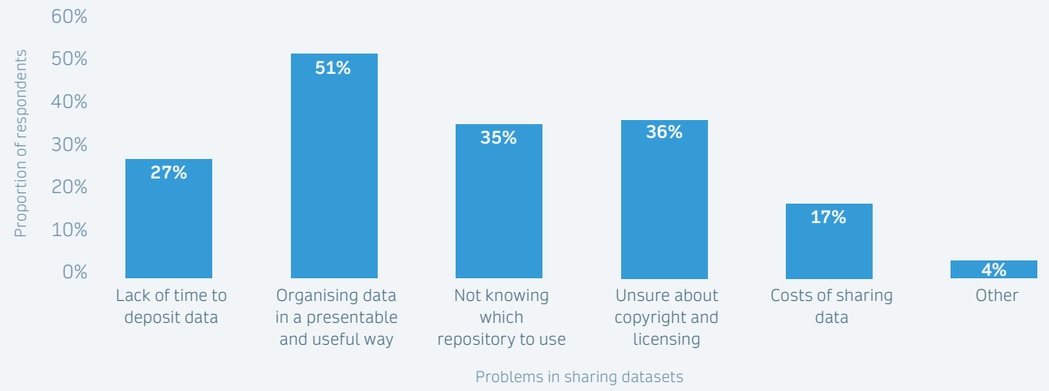
Medical sciences were the only subject area for which organising data was not the most often cited issue. More often than any other subject area, respondents in the medical sciences selected that they were "unsure about copyright and licensing", did not know which repository to use, and were concerned about the costs of sharing data. They selected the problems of organising data and a lack of time less often than any other subject area. Although organising data in a presentable and useful way is identified as an issue by 40% of respondents in the medical sciences, this is far less than the 57%

that selected it in the physical sciences, 56% in the Earth sciences, and 50% in the biological sciences. This supports the findings of an earlier survey which found that the most common data sharing concerns of clinical researchers were related to "appropriate data use, investigator or funder interests, and protection of research subjects".[xxi] A large number of 'Other' problems stated by medical researchers relate to data sensitivity (117 out of 172): "Ethical consideration, particularly participant's confidentiality", "Ethics and anonymisation of patient data", "Privacy and ethics issues with clinical data". Anonymisation of clinical research data for sharing can be time consuming, and costly[xxii] which might account for cost being a more important factor for medical researchers.

Physical sciences

Of the four subject areas, data sharing alongside publication was least prevalent within the physical sciences. Of the 487 respondents, 41% stated they neither shared data as supplementary information nor in repositories when submitting a research manuscript. Such a proportion is even higher than in the medical sciences, for all its concerns about intellectual property rights and the sensitivity of information.

Figure 12: Problems in sharing data in the physical sciences (n=487)



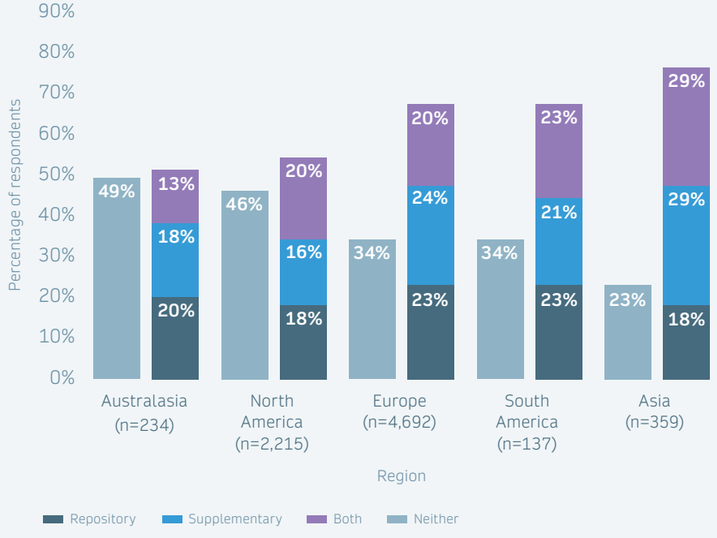
The low prevalence of data sharing in the physical sciences may be partly applicable to the lower importance ascribed to the discoverability of data (6.6 out of 10), but there are also more practical barriers. 'Organising data' was the most often identified problem for data sharing in the physical sciences, mentioned by 57% of respondents. The second most mentioned issue was being unsure about copyright and licensing (31%), followed by lack of time (27%), not knowing which repository to use (27%), and the cost of sharing data (15%). The 19 classified 'Other' comments were spread across all categories, with four related to organisational policies ("Employer restrictions"), four to technical issues ("Too big to share") and four to data issues ("raw data are of no use for people outside our collaboration."). In high energy physics, data may not be publicly share-able as they are too large – generated at large central facilities. While at CERN, for example, there is a strong data sharing policy and repository,[xxiii] this is not common across this field. In these cases it is not practical to share data even on request, as researchers might visit facilities for short periods to carry out research projects, and may not have access to the aggregated or raw data themselves.

3. Regional Differences in Data Sharing Challenges and Norms

The large scale and the international nature of the survey allow for some exploration of regional differences in the data, although Africa has been excluded from the regional analysis due to the small sample size (n=65). As can be seen from Figure 13, Australasia, North America and Europe show lower levels of data sharing compared to Asia and South America. 51% of respondents in Australasia stated that they shared data as supplementary files or in a repository, whereas in Asia it was 77%. Similarly, 33% of respondents in Australasia shared data in a repository, in comparison to 47% in Asia.

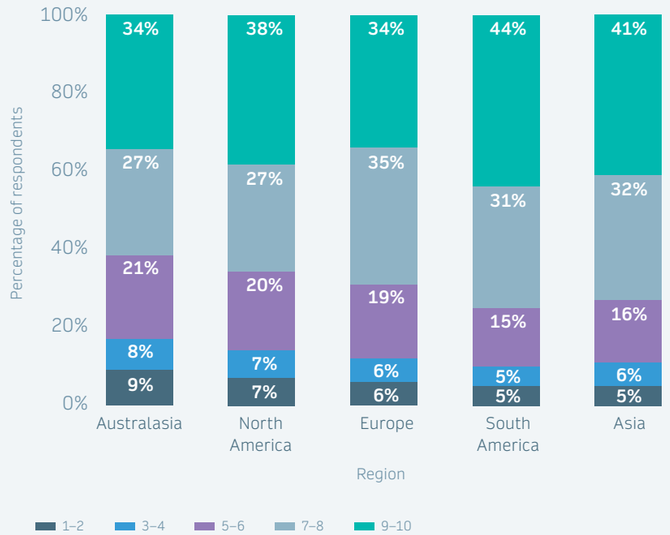
As previously noted, care should be taken with these findings as sample sizes in Asia and South America were small (n = 359 and n = 137, respectively) in comparison to the population and known levels of research in these regions, and the sample sizes from North America (n = 2,215) and Europe (n = 4,692). Respondents from Asia and South America may represent a self-selecting, data-interested group.

Figure 13: The depositing of data in different regions (n=7,632)



Australasia, North America and Europe are also the regions that rate the importance of discoverability least highly: Australasia, 6.9 out of 10; North America, 7.2; Europe, 7.3; Asia, 7.6; and South America, 7.7. Figure 14 shows distribution by region of the importance that data are discoverable, grouped in ranges of two.

Figure 14: The importance of discoverability in different regions (1 is the least important) (n=7,591)



There is also a lot of variation between countries within the same region, although outside of Australasia, North America and Europe, country-level analysis is limited by the number of respondents.

There were 17 countries in the survey with more than 100 respondents: Canada, United States, Australia and 14 European countries.

The European countries vary by as much as 20% in terms of the proportion of respondents stating that they share data through a repository, as supplementary information files, or both when submitting a manuscript. Respondents from Australia, United States and Canada report sharing data even less.

Table 1: Percentages of respondents sharing data through a repository, as supplementary information files, or both, in countries with >100 respondents

Country	Percentage of respondents
Poland	76%
Germany	75%
Switzerland	69%
Greece	69%
Italy	68%
Spain	66%
France	65%
Netherlands	64%
Norway	64%
Sweden	61%
Denmark	60%
Belgium	59%
United Kingdom	58%
Portugal	56%
Australia	55%
United States	55%
Canada	50%

There is a need for further research, especially amongst large producers of research such as China, Japan, and India, where sample sizes were too small for analysis by country.

Subject influences on regional behaviour

There are large differences in the subject speciality of respondents from the different regions. Noticeable subject differences between the regions can be seen in Figure 15, most evident within the medical sciences, contributing 18% of respondents in South America and 44% of respondents in Australasia.

Applying an equal weighting to different subjects in each region (Figure 16), a similar distribution occurs as Figure 13 with Australasia, North America and Europe sharing less. This does not necessarily mean there is an underlying cultural difference between the regions, however, the sample sizes from Asia and South America are relatively small and there may be a self-selection bias toward those in favour of data sharing. This is an area for further analysis of the results of this survey, and for future research.

Figure 15: The relationship between regions and sciences (n=7,686)

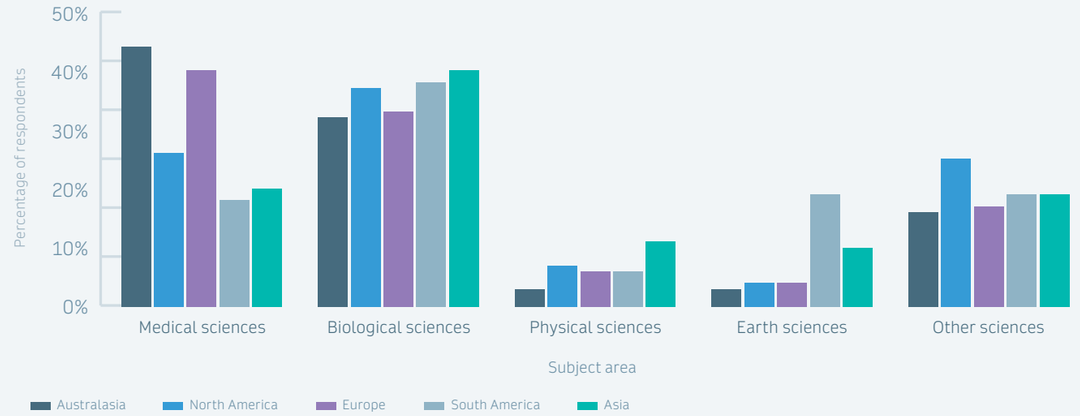
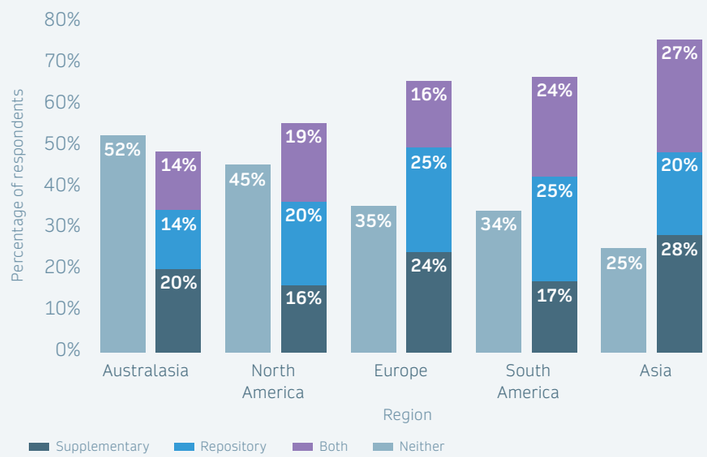


Figure 16: The depositing of data in different regions – adjusted for subject specialism (n=7,600)



Regional differences in the challenges in data sharing

There are noticeable differences between regions in terms of perceived barriers. Although 'Organising data in a presentable and useful way' was mentioned most often by respondents, the proportion that mentioned it ranged from 53% in South America to 43% in Australasia.

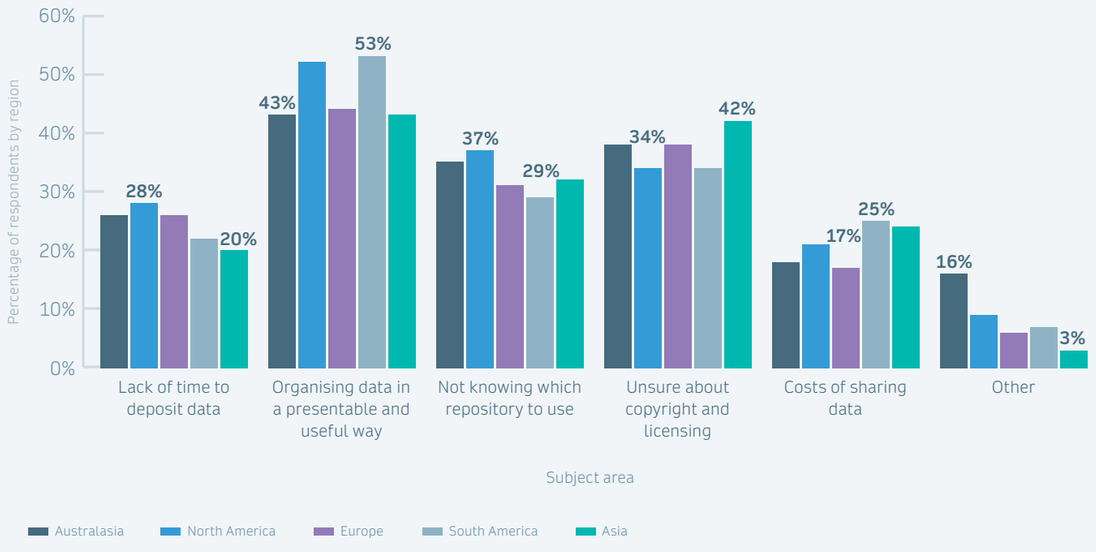
How much time is a factor compared to budget varies between Australasia, North America and Europe on the one hand and South America and Asia on the other. Percentage differences are not hugely different, but the results are nonetheless interesting. 'Lack of time to deposit data' was more highly rated by respondents in Australasia, North America and Europe than in South America and Asia:

- Australasia, 26%
- North America, 28%
- Europe, 26%
- South America, 22%
- Asia, 20%

'Costs of sharing data' were identified by a higher proportion of respondents from South America and Asia:

- Australasia, 18%
- North America, 21%
- Europe, 17%
- South America, 25%
- Asia, 24%

Figure 17: Reasons for not sharing data by region (n=7,654) – including % of highest and lowest responding regions



4. Discussion and Springer Nature Perspectives

The results of this survey show good reasons to be optimistic about the status quo, and the future, of data sharing: 63% of respondents stated that when submitting a manuscript they shared data files in a repository, as supplementary files, or both. Whilst in response to the question '*How important is it to you that your data are discoverable?*' the average rating was 7.3 out of 10. There is much work to be done in reducing the 37% who don't share their data, and this needs to be sensitive to the specific needs and challenges of differing fields and career stages of researchers.

Within the headline figures there are both subject and regional differences. The results of this survey would seem to support Tenopir et al.'s earlier survey findings that data sharing around the world can be understood in part as a reflection of collectivist and individualist cultures,[xxiv] although it may equally be attributable to the importance of time and financial constraints in different regions. Time is seen as a bigger challenge in Australasia, North America and Europe than in South America and Asia, whilst cost is a greater concern in South America and Asia than in Australasia, North America and Europe. Between the different subjects it was found that those in the biological and Earth sciences are far more likely to report sharing data than their counterparts in the physical and medical sciences.

There are also differences in data sharing for individual researchers, depending on the stage of their career and specifics of their research. In the same way as there are regional differences in the relative importance of time and cost, there is a trade-off between time and knowledge at different career stages; reporting a 'lack of time to deposit data' rises with a researcher's seniority whereas a lack of knowledge about 'copyright and licensing' falls. A higher proportion of first stage researchers also state 'not knowing which repository to use' as a problem. The size of the data produced by researchers in the same country and subject will also vary considerably. Research that generated the smallest sized data files had the highest proportion of data that were neither shared as supplementary information nor deposited in a repository.

Recognition of the value of data sharing has led to the adoption of many funder policies and mandates designed to motivate data sharing. These have been broadly found previously to have a positive impact; funder policies are seen as an important motivator[xxv] and data sharing mandates an effective strategy[xxvi] in encouraging data sharing, although there is a need for standardisation and harmonisation[xxvii] of policy both from funders and journals, where policies are also being more widely adopted.[xxviii] At a global scale, however, this current survey has found there seems to be little relationship yet between data sharing mandates in North America and Europe and behaviours around sharing data alongside research publication as supplementary information files or deposition in repositories. Subject variations in data sharing actions follow funder mandates more closely. A number of funder mandates were introduced or strengthened in 2017, during the time this research was conducted, so we are yet to see their impacts in researcher behaviour.

A recent analysis of data sharing in *The BMJ* found that rates of sharing were low despite a strong data sharing policy, with one possible explanation being that the wording of the policy left room for individual interpretation.[xxix] Unless there is a more joined up approach to data sharing policies that take into consideration the range of issues associated with a particular subject, it seems likely that researchers will continue to lack clarity on what is expected of them, and how to comply with funder and journal or publisher policies.[xxx]

Data sharing policies are not enough, however. To increase the amount of data that are shared, there is a need for clearer routes to help researchers through the increasingly complex scholarly ecosystem. The three problems raised most often reflect this need for help rather than simple money and mandates: 'Organising data in a presentable and useful way' (46%), 'Unsure about copyright and licensing' (37%), and 'Not knowing which repository to use' (33%).

These are much bigger issues than individuals and organisations that adhere to a closed culture of not sharing. Relatively few comments for 'Other' problems included: "I don't share", "I have no desire to share datasets", "Don't believe data should be shared w/o my specific control." These are a small proportion overall, accounting for only 34 of the 429 'Other' problems given. A much more pressing issue, though not considered here, is lack of incentives and credit for researchers to share their data, in terms of academic achievement and career advancement.

The survey suggests two principal recommendations, which support recommendations in an earlier ethnographic study[xxxi]:

- Increased data management, support and education
- Faster, easier routes to optimal ways of sharing data

Increased data management, support and education

Awareness and attitudes to the benefits of data sharing were not directly addressed in this survey, as they have been investigated in previously published research.

The survey does support and highlight key challenges noted in previous research, including the uncertainty about copyright and licensing and the problem of not knowing which repository to use. These were particularly seen as problems for early career researchers.

The lack of awareness of what repositories to use even within subject communities where there are norms for data sharing are also apparent from these results, supporting the view that increased general education and advice on repositories and policies is needed.

No researcher should be able to say, as one respondent did: "I didn't know it was possible." But even those who do know that it is possible need help through the process: "Whether to use metadata schemes and if so, which", "Selecting the right level of detail at which to share datasets", "Where and how to share", "Not sure where to put the data", "Lack of guidelines from journals". As Cameron Neylon puts it succinctly in his recent blog on the topic: *"As a researcher concerned to develop better RDM [Research Data Management] practice, I need support to meet me where I am"*.[xxxii] Simple knowledge sharing about which repositories to use and associated copyright and licensing laws and regulations could also reduce significant problems to data sharing. These are potentially cost-effective measures that could be put into place quickly.

Faster, easier routes to optimal ways of sharing data

The problem of organising data, and the lack of time to do so, requires readily available ways to organise and share data, which are easily accessible and usable by researchers.

Data management needs to be integrated into research and publishing processes from the beginning. This starts at the project proposal stage, ensuring that there are sufficient resources for data management, and a data management plan is written and then put into practice. Such integration in research and publishing workflows is unlikely to be achieved by researchers alone. Researchers are time-poor and don't necessarily want to become data experts: change will rely on closer collaboration between researchers, institutes, funders, publishers, repositories and other research data infrastructure providers (such as the Research Data Alliance, the Digital Curation Centre, DataCite and other national and international bodies)[xxxiii].

Another possible area of focus for publishers is around the use of supplementary information. Although sharing data as supplementary information is better than not sharing data at all, it may be considered a sub-optimal solution. Data deposited in a repository can still be linked to from the journal, but can also facilitate greater discoverability and accessibility by bringing together similar data in one place and providing data-specific metadata and persistent identifiers (such as digital object identifiers, DOIs) – enabling more powerful search functionality. In many cases, bi-directional linking between data and published work is also possible, joining up the research record for the benefit of readers and authors alike. Repositories often provide open licenses for deposited data, which can also be considered to help a more rapid transition to open data.

Future research

This survey is one of the largest of its kind, but has its limitations. There is a need for further in-depth studies to explore some of the subject, country, and regional differences and the relationships between them. Although the sample size of this survey (n=7,719) was larger than that of The State of Open Data Report 2017 (n=2,352), it was less balanced in terms of regional representation. The sample of this survey was heavily weighted towards respondents from North America and Europe, accounting for 90% of respondents in comparison to 52% of respondents in the earlier survey. Africa had to be excluded from the regional comparison, because of the small number of respondents (n=65). China (n=62), Japan (n=32), and India (n=99) didn't elicit sufficient responses for country-level analysis despite being some of the largest producers of research publications. Springer Nature will be conducting further in-depth analysis and research in specific countries and subject areas that will explore some of the noted differences in more depth. Research in China and Japan is planned for 2018.

There is also a need to explore the relationship between other factors that might be influencing data sharing behaviour in specific subjects and countries, such as data policies and mandates from journals and funders, and the wider data infrastructure.

Appendix

This data was collected between April and June 2017 by contacting registrants to nature.com, biomedcentral.com and springer.com. It went to ~249,000 recipients of which ~15,000 clicked through to the survey, resulting in 7,719 respondents from 126 different countries (see Figure 18). Africa was omitted from the regional analysis due to the small sample size, although the responses were included in the overall analysis and the subject analysis.

Of the 7,686 that stated the subject area of their data, 5,323 were in either the medical or biological sciences (see Figure 19). 'Other sciences' covers many disparate fields, including the social sciences, computer science, humanities and mathematics.

Figure 18: Number of respondents by region of the world

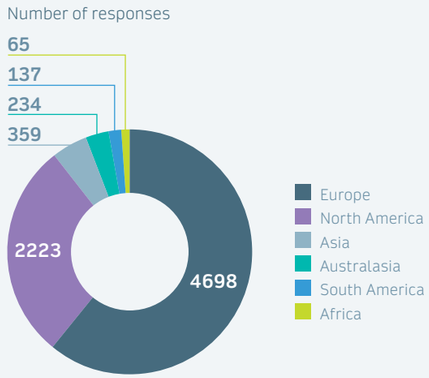
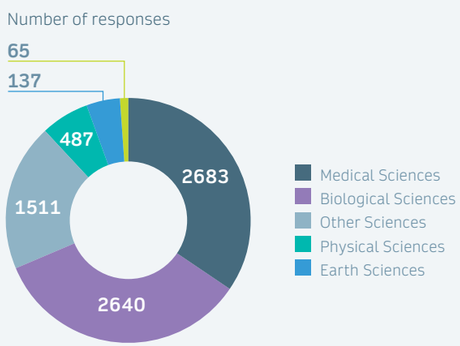


Figure 19: Number of respondents by subject



References

i. Digital Science (2017), The State of Open Data Report 2017, Digital Science, https://doi.org/10.6084/m9.figshare.5481187

ii. Tenopir, C. et al. (2015), Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide, *PLOS One*, https://doi.org/10.1371/journal.pone.0134826

iii. SHERPA Juliet, http://v2.sherpa.ac.uk/view/funder_by_data_req/requires.html

iv. Naughton, L. and Kernohan, D. (2016), Making sense of journal research policies, *Insights*, http://doi.org/10.1629/uksg.284

v. Tenopir, C. et al. (2015), Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide, *PLOS One,* https://doi.org/10.1371/journal.pone.0134826

vi. Digital Science (2017), The State of Open Data Report 2017, Digital Science, https://doi.org/10.6084/m9.figshare.5481187

vii. Tenopir, C. et al. (2015), Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide, *PLOS One,* https://doi.org/10.1371/journal.pone.0134826

viii. Springer Nature (2016), Over 600 Springer Nature journals commit to new data sharing policies, https://www.springernature.com/gp/group/media/press-releases/over-600-springer-nature-journals-commit-to-new-data-sharing-policies/11111248

ix. http://blogs.plos.org/everyone/2017/05/08/making-progress-toward-open-data/

x. Kratz JE & Strasser C (2015) Researcher Perspectives on Publication and Peer Review of Data. *PLOS ONE* https://doi.org/10.1371/journal.pone.0117619

xi. Kratz, J. & Strasser, C. (2015) Making data count. *Scientific Data* doi: 10.1038/sdata.2015.39

xii. Tenopir, C. et al. (2015), Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide, *PLOS One,* https://doi.org/10.1371/journal.pone.0134826

xiii. Schmidt, B. et al. (2016), Open Data in Global Environmental Research: The Belmont Forum's Open Data Survey, *PLOS One,* https://doi.org/10.1371/journal.pone.0146695

xiv. Tenopir, C. et al. (2015), Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide, *PLOS One,* https://doi.org/10.1371/journal.pone.0134826

xv. https://researchdata.springernature.com/users/69154-springer-nature/posts/21387-research-data-support-helpdesk

xvi. Tenopir, C. et al. (2015), Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide, *PLOS One,* https://doi.org/10.1371/journal.pone.0134826

xvii. European Commission (n.d.), EURAXESS – Researchers in Motion, https://euraxess.ec.europa.eu/europe/career-development/training-researchers/research-profiles-descriptors

xviii. Digital Science (2017), The State of Open Data Report 2017, Digital Science, https://doi.org/10.6084/m9.figshare.5481187

xix. *Scientific Data* (2018), Recommended Data Repositories, *Scientific Data,* https://www.nature.com/sdata/policies/repositories

xx. https://www.springernature.com/gp/authors/research-data-policy/faqs/12327154; http://www.nature.com/authors/policies/availability.html#data

xxi. Rathi, V. (2012), Sharing of clinical trial data among trialists: a cross sectional survey, *BMJ*, www.bmj.com/content/345/bmj.e7570

xxii. http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002304; https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-017-2382-9

xxiii. hepdata.net

xxiv. Tenopir, C. et al. (2015), Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide, *PLOS One,* https://doi.org/10.1371/journal.pone.0134826

xxv. Schmidt, B. et al. (2016), Open Data in Global Environmental Research: The Belmont Forum's Open Data Survey, *PLOS One,* https://doi.org/10.1371/journal.pone.0146695

xxvi. Thelwall, M. and Kousha, K. (2017), Do journal data sharing mandates work? Life sciences evidence from Dryad, *Aslib Journal of Information Management,* https://doi.org/10.1108/AJIM-09-2016-0159

xxvii. Naughton, L. and Kernohan, D. (2016), Making sense of journal research data policies, *Insights,* http://doi.org/10.1629/uksg.284

xxviii. Vines, T.H. (2012) Mandated data archiving greatly improves access to research, *FASEB Journal,* https://doi.org/10.1096/fj.12-218164

xxix. Rowhani-Farid, A. and Barnett, A.G. (2016), Has open data arrived at the British Medical Journal (BMJ)? An observational study, *BMJ Open,* https://www.ncbi.nlm.nih.gov/pubmed/27737882

xxx. Hrynaszkiewicz, I. et al. (2017), Standardising and Harmonising Research Data Policy in Scholary Publishing, *International Journal of Digital Curation,* http://www.ijdc.net/article/view/12.1.65

xxxi. Jahnke, L.M. and Asher, A. (2012), The problem of data: Data management and curation practices among university researchers, Council on Library and Information Resources, https://www.clir.org/pubs/reports/pub154/problem-of-data

xxxii. Neylon, C. (2017), As a researcher…I'm a bit bloody fed up with data management, Science in the Open, http://cameronneylon.net/blog/as-a-researcher-im-a-bit-bloody-fed-up-with-data-management/

xxxiii. Baynes, G. (2017), Collaboration and Concerted Action are Key to Making Open Data a Reality, The State of Open Data Report 2017, https://figshare.com/articles/_/5481187

As a global publisher Springer Nature is dedicated to providing the best possible service to the whole research community. It helps authors to share their discoveries; enables researchers to find, access and understand the work of others; and supports librarians and institutions with innovations in technology and data.

Springer Nature is a leading publisher in improving the sharing, visibility and reusability of research data – offering researchers a variety of options to help them handle, share and promote their research data, as well as helping them to comply with funder and institutional data policies.

Research Data Support

To help authors and journals follow good practice in sharing and archiving research data, Springer Nature provides an optional data deposition and curation service – Research Data Support. This service enables greater data sharing, compliance with funder policies and enhances peer-reviewed publications.
http://go.nature.com/SNRDS

Research data training

Springer Nature's research data training program allows institutions to provide knowledge and insights in research data best practice to their researchers, building a custom curriculum based on the needs of the institution. Training is designed to benefit anyone who has an interest in research data best practice, including: researchers, policy makers, librarians, research data managers, open access teams, scholarly communication offices, and research offices.
http://go.nature.com/RDSInstitutions

Research Data Helpdesk

Springer Nature provides a free-to-use Research Data Helpdesk, which anyone can contact to get advice on research data – including: choosing a repository; how to write a data availability statement; help on licences for sharing data; advice on implementing a data policy; and much more.
http://go.nature.com/RDHelpdesk

Tracking glacier response to climate change
Currently about half of Greenland's ice losses are through the release of icebergs into the ocean, a process known as calving - rather than by surface melt. In order to investigate the complex response of Arctic glaciers to warming of the atmosphere and ocean, the CRIOS project (Calving Rates and Impact on Sea Level), recorded data from key locations including Kronebreen, a fast-flowing glacier in Svalbard, Arctic Norway. Data from the project is now being used to develop improved predictive glacier models as part of a NERC-funded project.

For more information, visit
springernature.com/openresearch

Follow facebook.com/SpringerNature
Follow twitter.com/SpringerNature

Picture: NatureJobs Scientists at Work competition