SPRINGER NATURE

springernature.com

Research Data

FIVE ESSENTIAL FACTORS FOR DATA SHARING

White paper

ADVANCING DISCOVERY

Contents

Authors

Mithu Lucraft; Grace Baynes; Katie Allin; Iain Hrynaszkiewicz; Varsha Khodiyar

April 2019

Access white paper:
https://doi.org/10.6084/m9.figshare.7807949.v1

Foreword

It's been a busy year for the research data team at Springer Nature since we published *Practical Challenges for Researchers in Data Sharing* in March 2018. Personally, I have learned a lot through our continued research and conversations with funders, researchers and institutions.

We have also continued to develop the solutions we can offer to the research community to help make good data practice the norm. There is no learning like doing, and we continue to be heartened by enthusiasm from funders, foundations and institutions, and reality-checked by the early stages of allocated budgets for research data. We've seen that researchers may report that they see the value in sharing data, but in general they are not doing so with alacrity, even when support is offered. A reminder that we need to do more to make it easier for researchers to share, and more obvious why it is worth their time and effort.

Since our last report, we have followed up with research in China and Japan, to complement the picture of North America and Europe in *Practical Challenges*. We had a good response rate to both pieces of research, which has enabled us to draw insights and comparisons based on decent sample sizes. We also partnered again with Digital Science and figshare on the *State of Open Data Report 2018*. This report summarises all we have learnt through those further studies, and brings them together with *Practical Challenges* to provide a more global view.

While we continue to see researchers increasingly sharing data, the majority of the research community are not yet managing or sharing data in ways that make it findable, accessible or reusable. The utopia of findable, accessible, interoperable and reusable (FAIR) data is still some way off. A recent report for the European Commission[1] puts the minimum cost to the EU of poor data practice at €10.2 billion per year, emphasizing how much is at stake.

Our learnings over the past twelve months have cemented for me how much needs to be done to make good research data practice as routine and commonplace as the publication of research articles and monographs. Our thinking on this has crystallised into the "five essential factors" we set out in this report:

1. **Clear policy**: from funders, institutions, journals/publishers, and research communities themselves.
2. **Better credit**: to make data sharing worth a researcher's time.
3. **Explicit funding**: for data management and data sharing, as well as data publishing.
4. **Practical help**: for organising data, finding appropriate repositories, and provision of faster, easier routes to share data.
5. **Training and education**: to answer common questions from researchers on data sharing and to help build skills and knowledge.

Grace Baynes
VP Research Data &
New Product Development
Springer Nature

1 Directorate-General for Research and Innovation (European Commission), and PwC EU Services (2019) Cost of not having FAIR research data – Study. DOI: 10.2777/02999 https://publications.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1

As we note in the report, none of these essential factors can be solved by one stakeholder alone. Funders, institutions, publishers, and the wider research community, including researchers themselves, all have a role to play.

These five essential factors chime with recommended actions set out by experts on this topic. In the UK the Open Research Data Task Force (ORDTF) recently published a set of recommendations in their final report to the UK Government: Realising the Potential.[2] To build capacity for open research data, the ORDTF make recommendations on funding, services, policy, incentives and leadership – demonstrating the growing, collective awareness of these issues to be addressed in making comprehensive data sharing a reality.

Springer Nature's report adds to the debate with evidence from surveys of over 11,000 researchers worldwide. Our findings also make clear that these essential factors are global. While national and discipline-specific foci may vary, these recommendations apply to researchers in every country and region we have looked at.

Our work to better understand how to make a difference continues. In 2019, we are focusing on better understanding what "credit" means to researchers, and moving beyond positive attitudes to decision-making and behaviours. What makes a researcher decide to do the work to manage or share their research data in ways that are FAIR? Or to publish a data article to make datasets easier to find, understand and use? We hope by understanding what is making researchers take action, we can better help more researchers to do so. I'm curious about what we will learn, and we look forward to continuing to share our findings to help accelerate progress to make research data as open as possible.

Grace Baynes
VP Research Data & New Product Development
Springer Nature

2 UK Open Research Data Task Force. (2019). Realising the potential: final report of the Open Research Data Task Force. Retrieved from https://www.gov.uk/government/publications/open-research-data-task-force-final-report

Progress in data sharing: challenges and opportunities

Encouraging evidence of data sharing and increasing awareness of data best practice

Research data are the building blocks of discovery. Open access to research data makes research more reliable and efficient, helping to speed the pace of discovery, and delivering increased value by enabling reuse and reducing duplication. With global concerns about reproducibility and research integrity, the case for good data practice is even stronger. Previous research has shown that as much as 50% of preclinical research done in the US, at a cost of USD 56.4 billion a year, cannot be reproduced.[3] A *Nature* survey further found that 70% of over 1,500 respondents had tried and failed to replicate the work of others. More shocking was that 50% of respondents were unable to replicate their own work.[4]

Data sharing has been found to offer many benefits to researchers. Data archiving can double the publication output of research projects,[5] and has been associated with an increase in the citation impact of research papers by as much as 50%.[6] There is also evidence of public and societal benefits through data sharing. For example, an independent report found that the European Bioinformatics Institute, an intergovernmental organisation that delivers freely available molecular data and services to scientists around the world, returns GBP 1 billion in annual efficiency savings to researchers worldwide.[7] The minimum cost to the EU of poor data practice is €10.2 billion per year, according to a new report for the European Commission.[8]

Springer Nature has not only invested in data publishing options alongside our journal and book publishing, but we have started to develop tools and services that support good data management, recognising the benefits these bring to research at large. We believe researchers require increased education and support on good data management, and faster, easier routes to sharing data. In both cases, we are collaborating with researchers, institutes, funders, repositories and other research data infrastructure and service providers to make data sharing the new normal.

Throughout 2018, we conducted, participated in and published a number of projects that investigated researcher attitudes and behaviours towards data sharing and data management. These included:

- ***Practical Challenges for Researchers in Data Sharing***[9]: a white paper summarizing the findings of a Springer Nature survey with 7,000 researchers globally. Published March 2018.
- ***The State of Open Data Report 2018***[10]: a Digital Science project undertaking longitudinal research on data sharing, now in its third year. Published October 2018.
- ***Challenges and Opportunities for Data Sharing in China***[11]: a survey on data sharing best practice and behaviours with researchers in China. Published February 2019.
- ***Research Data Sharing in Japan***: a survey of researchers in Japan on data sharing best practice and behaviours.[12] Forthcoming, 2019.

3 Freedman, L.P.; Cockburn, I.M.; Simcoe, T.S. (2015) The Economics of Reproducibility in Preclinical Research. PLoS Biology 13(6): e1002165. https://doi.org/10.1371/journal.pbio.1002165

4 Baker, Monya (2016) 1,500 scientists lift the lid on reproducibility. Nature News. https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970#/correction1

5 Pienta, A. M.; Alter, G. C.; Lyle, J. A. (2010) The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data. University of Michigan. https://deepblue.lib.umich.edu/handle/2027.42/78307

6 Dorch, B., Drachen, T., & Ellegaard, O. (2015) The data sharing advantage in astrophysics. Proceedings of the International Astronomical Union, 11(A29A), 172-175. doi:10.1017/S1743921316002696

7 Beagrie, N. and Houghton, J. (2016) The Value and Impact of the European Bioinformatics Institute. Beagrie.com https://beagrie.com/static/resource/EBI-impact-report.pdf

8 See 1.

9 Stuart, D. Baynes, G.; Hrynaszkiewicz, I.; Allin, K.; Penny, D.; Lucraft, M.; et al. (2018): Whitepaper: Practical Challenges for researchers in data sharing. figshare. https://doi.org/10.6084/m9.figshare.5975011.v1

10 Digital Science; Hahnel, Mark; Fane, Briony; Treadway, Jon; Baynes, Grace; Wilkinson, Ross; et al. (2018): The State of Open Data Report 2018. figshare. https://doi.org/10.6084/m9.figshare.7195058.v2

11 Lucraft, M., Allin, K., Baynes, G. & Sakellaropoulou, R. Challenges and Opportunities for Data Sharing in China. (2019). https://doi.org/10.6084/m9.figshare.7718441.v1

12 Springer Nature (Forthcoming): Research data: challenges and opportunities for Japanese researchers.

These studies all offer encouraging evidence of data sharing and increasing awareness of data best practice:

Increased sharing of data
As a longitudinal survey, the *State of Open Data Report* shows steady growth in the number of researchers sharing their data, up consistently year on year to 64% in 2018. Our *Practical Challenges* report shows similar evidence of data sharing, with 63% generally submitting research data files at the point of publishing a research article, either as supplementary information or in a repository (or both).

The majority of researchers see data sharing as important
In *Practical Challenges*, when asked about the importance of making their data discoverable, researchers gave an average rating of 7.3 out of 10, with the most popular rating being the maximum rating, 10 out of 10 (25%). In our follow up regional surveys with researchers in China and in Japan, we saw similar levels of agreement, with an average score of 8.0 for researchers in China and 7.2 in Japan.

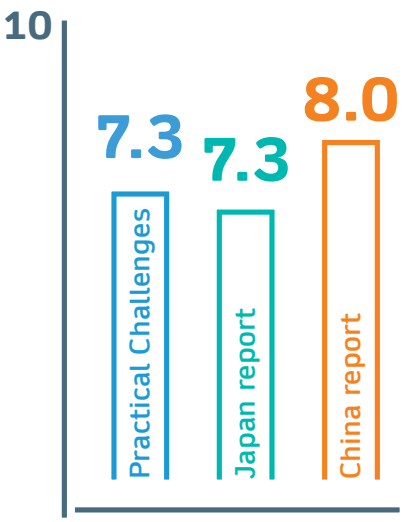
However, we also found low levels of good data practice in all of these studies:

Little evidence for data management plans (DMPs)
The State of Open Data Report found that only 41% of researchers globally create DMPs for more than half of their research. Similarly, our study on researchers in Japan found only 35% doing this. In our survey with researchers in China the proportion was higher at 58%, although this is still a low number of researchers undertaking best practice in data management overall.

64%

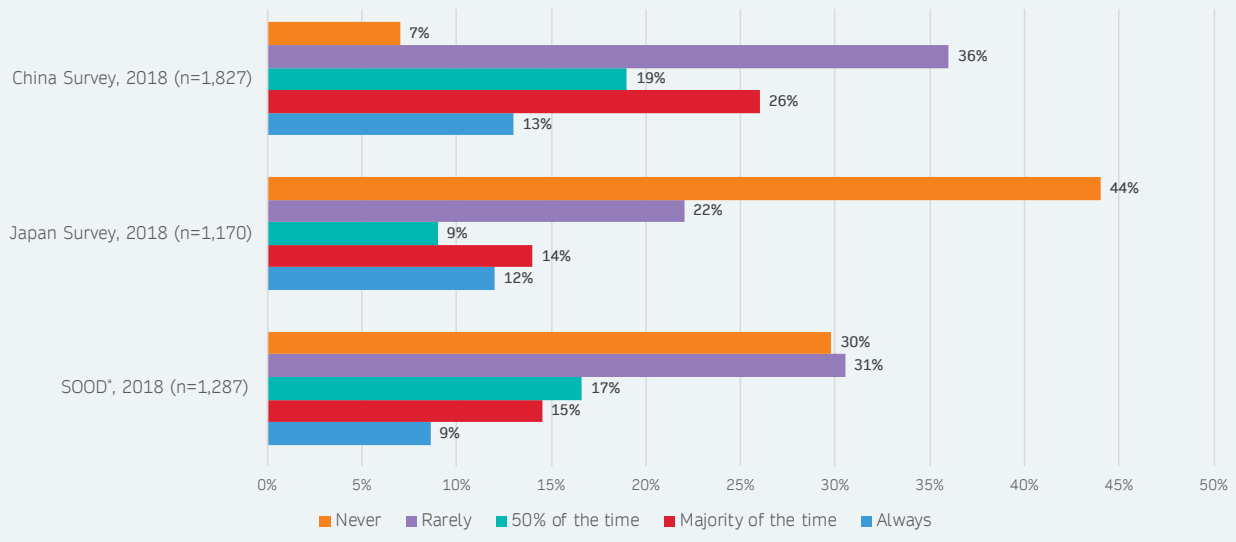
of researchers revealed they made their data openly available in 2018

How important is it to you that your data are easy for other researchers to find? (Average score out of 10)



(n=7,656) (n=1,062) (n=1,731)

How often do you create a data management plan for the research you carry out?



*State of Open Data Report

Data sharing methods are suboptimal

When asked in the *State of Open Data Report* where researchers publish their data, 35% of respondents had published their data as an appendix to a research article, with little change from 2017 (34%). We also continue to see the majority of private file sharing happening over email or by USB, as evidenced in our surveys with researchers in China and in Japan. Sharing data over email and stored on physical media that are not backed up are less secure and at greater risk of being lost than using data repositories and other securely archived storage media.

We see low levels of awareness of best practices amongst researchers, both across geographic and subject discipline boundaries. This is the case even in disciplines with established norms for data sharing, and with well-established repositories available to researchers. In understanding barriers to sharing, we see knowledge (what, how and where to share) and the need for faster, easier routes to sharing as key issues to address.

How do you share your data privately (person to person)?

49%
USB or flash drives

43%
email

30%
PC hard drive

China survey (n=1,441)

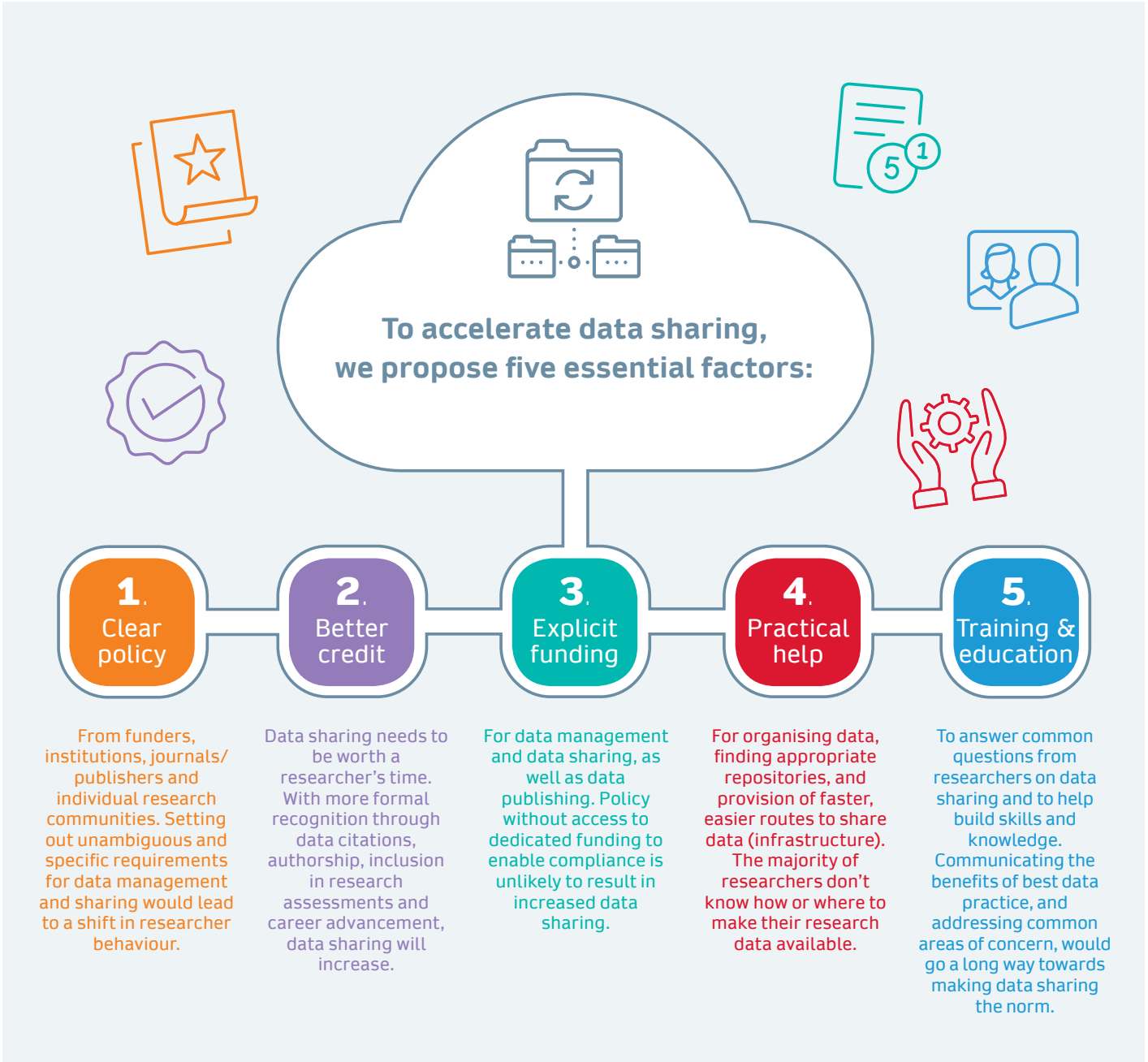
65%
email

41%
USB or flash drives

39%
file sharing services

Japan survey (n=905)

Five essential factors for data sharing



None of these essential factors can be solved by one stakeholder alone: we must act together, and we must act now, to encourage data sharing across discipline and geographic boundaries. Support from all stakeholders – funders, institutions, publishers, and the wider research community – could make all the difference.

This report sets out our recent findings in more detail, alongside examples of best practice and evidence of practical solutions.

1. Clear policy: from funders, institutions, journals/publishers, and research communities

1. Clear policy

More communication is needed on funder requirements to increase awareness and uptake

1a. Funders

Adoption of policy

Recognising the value of data sharing has led to the adoption of funder policies internationally, notably in the US and Europe, and particularly in the UK. Increasingly, these policies mandate good research data management by individual researchers (the use of data management plans and sharing data), whilst also acknowledging the need for global collaboration on infrastructure and best practice.

Springer Nature tracks funder policies on data to provide researchers with advice on compliance. In 2018, the policies of more than 50 global funders required data sharing. This includes:

- The European Commission's Horizon Europe proposal, which will mandate open access to research data as well as publications.
- In the UK, UKRI and Wellcome are notable early movers in encouraging and requiring data sharing.
- In 2018, China's Ministry of Science and Technology introduced their "Notice of the General Office of the State Council on the Measures for Managing the Printing and Distributing of Scientific Data", also effectively requiring that data is shared at a national level[13].

State of Open Data Report:
What circumstances would motivate you to share your data? (n=1,359) (multiple select)

Answer	%	Count
Increased impact and visibility of my research	62%	841
Public benefit	59%	802
Transparency and re-use	48%	652
Getting proper credit for sharing data	46%	621
Journal/publisher requirement	44%	599
Trust the person requesting my data	41%	561
Institution/organisation requirement	38%	522
It was made easy and simple to do so	36%	485
Funder requirement	33%	453
Freedom of information request	26%	352
Other (please specify)	5%	63
I would never share my data	1%	17
Total	100%	1,359

13 Enago Academy (2018) China Open Science and Open Data Mandate Released. https://www.enago.com/academy/china-open-science-open-data-manadate-released/

Impact of policies

Globally, our findings suggest there is little relationship between data sharing mandates and data sharing behaviours. In *Practical Challenges*, we found self-reported levels of sharing below the global average of 63% by respondents in the UK (58%) and US (55%), suggesting funder mandates may not be a key motivator for sharing data. This is supported by responses to the *State of Open Data Report* where researchers' motivations for data sharing placed funder requirements 9th (selected by 33%) out of 12 options provided. This however does contradict the findings of other studies,[14] and is in conflict with what we have seen in the growth of open access publishing which has been driven in large part by funders issuing clear and specific mandates.

Awareness of requirements

A shared issue across all regions was awareness of requirements:

- For researchers in China, 19% answered that they did not know what their funder requirement for DMPs was, and 23% did not know what their requirements were for data sharing.
- In our survey of researchers in Japan, 34% stated they did not know their funder requirement for DMPs, and 23% did not know their requirements for data sharing.
- Even where researchers thought they knew their funder requirements, this was often incorrect. For example in Japan, the Japan Science and Technology Agency (JST) requires data sharing, however in our survey with researchers in Japan, only 11% of respondents identifying JST as their funder correctly identified this requirement. 66% incorrectly identified JST's requirements thinking there were none, or that JST only encourages sharing of data.

More communication about these funder requirements is needed to increase awareness and uptake amongst researchers.

1b. Research communities

There appear to be stronger connections between funder mandates and data sharing behaviours by subject discipline. We found wide variation in data sharing behaviours in our *Practical Challenges* report ranging from 75% in the biological sciences who shared their data at the point of publication to 59% in the physical sciences. These subject variations follow funder mandates, with more requirements and greater availability of repositories in the biological and medical sciences compared to the physical sciences.

Practical Challenges The depositing of data in different subject areas (n=7,664)	Medical Sciences	Biological Sciences	Physical Sciences	Earth Sciences	Other Sciences
Neither	39%	25%	41%	32%	54%
Supplementary	20%	28%	18%	28%	13%
Repository	23%	16%	27%	25%	23%
Both	18%	30%	13%	16%	10%

Discipline-specific challenges to data sharing

Data sharing in the physical sciences may also be lower due to other practical considerations, such as the size of data involved. For example in *Practical Challenges*, "organising data" was the most commonly identified problem for researchers in this discipline, mentioned by 57% of respondents. Open comments

14 Schmidt, B.; Gemeinholzer, B.; Treloar, A. (2016) Open Data in Global Environmental Research: The Belmont Forum's Open Data Survey. PLoS ONE 11(1): e0146695. https://doi.org/10.1371/journal.pone.0146695

highlighted “employer restrictions” and “too big to share” as a technical issues. In high energy physics, data may not be publicly shareable as they are too large, and generated at large central facilities. Therefore it may not be practical for individuals to share data, as researchers may not have access to the aggregated or raw data themselves, or the technical means to share such large datasets.

Medical researchers further identified concerns about data sensitivity and misuse (representing 66% of open comments by medical researchers in *Practical Challenges*) and concerns about protecting research participants, consistent with other surveys of clinical researchers.

Taking these considerations into account, it is unlikely that there can be blanket policies applied across all disciplines for data sharing.

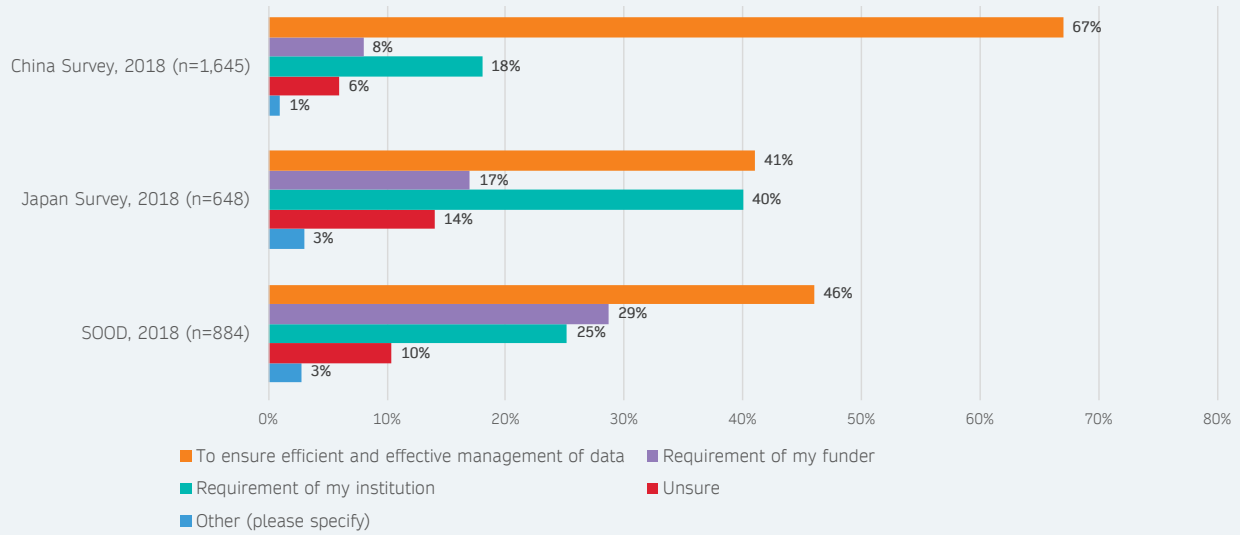
Awareness of requirements

Lack of awareness is also evident as a challenge within subject disciplines. The *Practical Challenges* report found that even in subject communities with established norms for data sharing, with both funder mandates and available community repositories, there was low awareness of data sharing requirements. Of the 2,288 respondents from the medical and biological sciences who produced specific data outputs such as nucleotide sequences (data types which have long had community mandates for sharing and dedicated repositories for these data), only 54% answered that they "always deposit" their sequence data to these repositories. Making it easy for researchers to find out where to share data is clearly important.

1c. Institutions

Globally, researchers place institutional requirements slightly higher than funder requirements as a motivator for data sharing (*State of Open Data Report*), but this is still low at 38% of respondents. For DMP creation, institutional requirements were also less important in influencing researcher behaviour. For researchers in our Japan survey, only 40% cited institutional requirements as the reason for creating a DMP, and this was even lower for researchers in China at only 18%.

Why have you created data management plan(s) in the past? Please select all that apply.



Alongside the introduction of policy is the need to help researchers understand and comply with funder requirements. This may come in the form of training, establishing local research data management solutions, or the provision of infrastructure. One example of institutional support is at TU Delft, where to embed good data practice they have created "data stewards" in every faculty, providing training, additional funding for data management and publication, as well as a data repository via DANS11.[15] Partnering with data initiatives, repositories, and other useful parties, including publishers, will help reduce potential duplication of effort and ensure sustainability.

1d. Journals and publishers

Adoption of policy

Springer Nature and many other publishers now have journal data policies that require or recommend data availability statements and data sharing. In 2016, we introduced a set of standardised research data policies that can be easily adopted by journals and understood by authors.[16] Providing a set of policies allows journals and editors to adopt the policy that most closely follows the data sharing norms of their discipline, and provides a pragmatic series of 'steps' for journals wanting to encourage and facilitate increased data sharing in the communities they serve. We aim to have the most comprehensive and inclusive research data policy of any large publisher, and we have released our set of policies under a CC BY license to enable others to follow our lead.

- More than 1,500 Springer Nature journals have now adopted a standardised data policy.
- The practice of providing journals with a set of research data policies has since been adopted and implemented by many other major publishers.
- The Research Data Alliance (RDA) data policy standardisation and implementation interest group,[17] co-chaired by Iain Hrynaszkiewicz at Springer Nature, is helping to define common frameworks for research data policies, and is developing appropriate support and guidance for researchers in complying with these policies.

Impact of policies

The impact of journal policies on data sharing is higher still than both funder and institutional requirements, based on reporting in both the *State of Open Data Report* and our surveys with researchers in Japan and China. In the *State of Open Data Report*, journal requirement was identified as the 5th highest reason a researcher would be motivated to share their data (44%). For researchers in China, lack of journal requirement was the top reason why researchers had not shared data (35%).

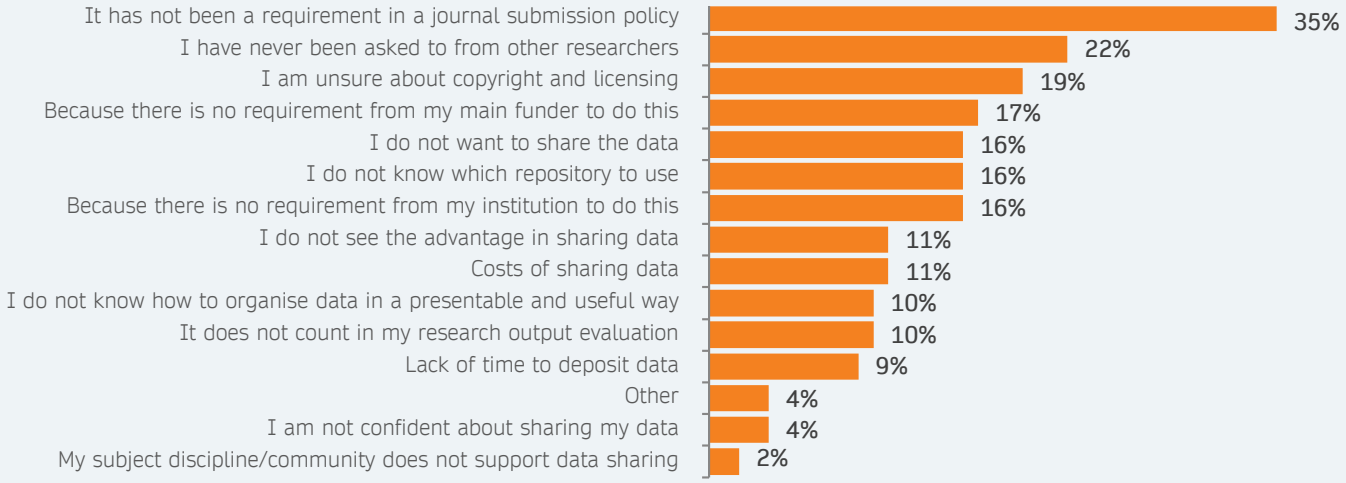
15 Data, Scientific; Teperek, Marta (2019): On a (cultural) journey towards FAIR data by Marta Teperek. figshare. Fileset. https://doi.org/10.6084/m9.figshare.7560221.v1

16 Springer Nature Data Policies: https://www.springernature.com/gp/authors/research-data-policy/data-policy-types/12327096

17 Research Data Alliance data policy standardisation and implementation interest group: https://www.rd-alliance.org/groups/data-policy-standardisation-and-implementation-ig

China report

Why have you not shared data generated by your research? (n=108)



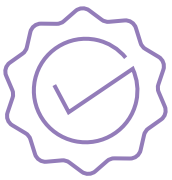
However, even where strong journal policies have been implemented, there remain low levels of sharing. Analysis of data sharing in *The BMJ* found that rates of sharing were low despite a strong data sharing policy, with one possible explanation being that the wording of the policy left room for individual interpretation.[18] As noted with awareness of funder requirements, for journals too, researchers need greater support on what is expected of them, and guidance on how to comply with funder and journal policies.

- Springer Nature launched a free Research Data Support Helpdesk in 2016 that helps authors and editors find out how best to manage and share their data, and provides guidance on understanding funder and institutional mandates. The Research Data Support Helpdesk helps to ensure research data are appropriately archived and, where possible, made widely accessible.
- We have also developed data availability reporting for Springer Nature publications, in order to help institutions and funders track compliance with their data sharing policies.

In addition to journal policy requirements, publishers can also support good data practice through publication of data articles. These provide an established credit mechanism – a citable publication – while making datasets easier to find, access and reuse. We facilitate data publishing in open access journals *Scientific Data* and *BMC Research Notes*, pilot projects on data citation and data badges.

18 Rowhani-Farid, A. and Barnett, A.G. (2016), Has open data arrived at the British Medical Journal (BMJ)? An observational study: BMJ Open: https://bmjopen.bmj.com/content/6/10/e011784

2. Better credit: incentives, including in research assessment



2. Better credit

58% of researchers don't think they get sufficient credit for sharing research data

Beyond policy and requirements, the need to incentivise researchers with good credit mechanisms is of significant importance. In all of the research we conducted during 2018, increased impact, progress in their field, and public benefit were highlighted as the most significant motivators for researchers in sharing their data. This was as high as 62% for "increased visibility and impact of my research" in the *State of Open Data Report*. For researchers in our China survey, 46% selected "to progress research in my field", and in our Japan survey this was 50%. However, in our view, we believe researchers would begin to share data more routinely, and more openly, if they received proper credit for doing so.

- In The *State of Open Data Report*, 58% of respondents said "No" to the question "Do you think researchers currently get sufficient credit for sharing research data?" Only 9% answered yes to this question.
- In the *State of Open Data Report*, "getting proper credit" placed as the fourth highest motivation for sharing (46%).

Measuring usage and citations

In open text comments from the *State of Open Data Report*, researchers point to data citations, co-authorship for articles using the data, and credit in research assessment as some of the possible routes to increasing credit in data sharing. Whilst routine inclusion of datasets in research assessments and CV evaluation are most likely years away, immediate steps in the right direction include measuring the usage and citation of datasets, including providing DOIs or other unique permanent identifiers for datasets. There has been significant progress in this area, such as the GO FAIR metrics group,[19] the FAIRdat project from DANS[20] and MakeDataCount.[21] figshare and other repositories include download and citation statistics, and alternative metrics for datasets.

State of Open Data Report:
What credit mechanisms do you think would encourage more researchers to share their data? – Coded (Base n = 623; total n = 1,874; 1,251 missing)

Answer	%
Citation	30%
Co-authorship	18%
Acknowledgement	13%
Financial/discounts	7%
Counts towards tenure/grants	7%
Cultural	5%
Mandates	5%
Visibility/transparency on use	5%
Limit misuse/security	3%
Data index/dedicated system	3%
Making it easier/education	2%
Other	8%
Unsure	15%
None	2%

19 Go Fair: https://www.go-fair.org/

20 FAIRdat: https://dans.knaw.nl/nl

21 Make Data Count: https://makedatacount.org/

Data citations

Encouraging and enabling data citations is also critical. *The State of Open Data Report* reported 55% of respondents would value a data citation as much as an article citation (although this is a decline from the 2017 and 2016 reports which reported 67% and 68% as valuing these two types of citations equally). There has been progress in this area, such as DataCite[22] and the FORCE Data Citation Roadmap.[23] Publishers are also increasingly providing links to datasets on articles, and including dataset citations in article metadata.

State of Open Data Report: Which of the following statements do you most agree with? (n=1,152)		
Answer	**%**	**Count**
I value a data citation the same amount as I value a citation to an article	55%	633
I value a data citation less than I value a citation to an article	30%	346
I value a data citation more than I value a citation to an article	9%	107
I do not value data citations at all	6%	66
Total	100%	1152

Acknowledgement through data re-use

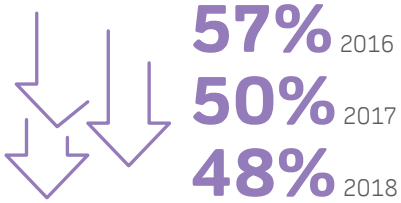
Evidence of the value in data re-use may also further incentivise researchers to share their data.

- In the *State of Open Data Report*, 79% of respondents said they would be willing to reuse open data in the future.
- A further 60% noted that re-use that resulted in credit as a co-author in a subsequent paper would motivate them "quite a lot" or "a lot" to share their data.
- Despite these indications, the *State of Open Data Report* shows a continued decline in the number of respondents who say they have reused open data in their research. In 2018 this was 48% of respondents, compared with 50% in 2017 and 57% in 2016.

79%

of respondents said they would be willing to reuse open data in the future

Continued decline in the number of respondents who say they have reused open data in their research



Data articles

As previously noted, data articles provide an established credit mechanism, although to date uptake of publishing data articles continues to be low. In both the *State of Open Data Report* and our surveys with researchers in China and Japan, we saw much higher reporting of publication in data journals than we believe to be the reality, looking at data journal publication volumes. In the *State of Open Data Report*, 18% of respondents reported publishing data in a data journal. The same percentage was reported by researchers in our Japan survey, and was even higher in our survey with researchers in China at 33%. This suggests there is some misunderstanding amongst researchers about what constitutes a data article, and there is more that can be done to make it easier for researchers to write and publish this kind of article.

We see the issue of credit for good practice in data management and data sharing as fundamental to accelerating progress, and shifting the needle towards making data sharing the norm. Springer Nature will be actively undertaking further analysis of researcher motivations for data sharing in 2019, explicitly reviewing credit mechanisms further and identifying ways to make data sharing worth a researcher's time and energy.

22 Data Cite: https://datacite.org/

23 Cousijn, H.; Kenall, A.; Ganley, E.; Harrison, M.; Kernohan, D.; Lemberger, T.; Murphy, F.; Polischuk, P.; Taylor, S.; Martone, M. & Clark, T. (2018) A data citation roadmap for scientific publishers. Scientific Data 5, 180259. https://doi.org/10.1038/sdata.2018.259

3. Explicit funding: for data management and data sharing, as well as data publishing

Whilst funders are increasingly developing policy for good data management and data sharing, this must be coupled with dedicated funding and clear guidance about using grants.

As yet, few funders explicitly make funding available for data management, storage and curation. The 2016 European Commission High Level Expert Group on the European Open Science Cloud "estimate that on average about 5% of total research expenditure should be spent on properly managing and 'stewarding' data".[24] There are examples of moves in this direction, such as the European Open Science Cloud, NIH Data Commons pilot, and initiatives from Wellcome and UKRI/JISC. In 2018, Springer Nature began a pilot with Wellcome to make data deposition services available to researchers with Wellcome funding at no cost to the individual.

There is evidence for support from researchers for a data sharing mandate at a national level.

- The *State of Open Data Report* found 63% of respondents in favour of such a mandate, although this figure has decreased since 2016 when 78% were in favour.
- This decrease coincides with an increase in the percentage of researchers uncertain about where funding to make their data openly available will come from: 27% of respondents (329 researchers) said that they did not know how they would meet the costs of making their research data openly available.
- Responses on how researchers would meet the costs of making data open show 37% likely to use their own funds for data sharing; 39% who would use money specifically for this purpose from a funder; and 41% who would use funds identified in their grant.

Funding may be particularly relevant in regions where cost is identified as a particular barrier to data sharing. In *Practical Challenges*, cost was perceived to be a larger challenge to researchers in South America and Asia, (where up to 25% of researchers cited it as a barrier) compared to researchers from Australasia, Europe and North America (where as few as 17% cited it).

As we have already noted, there is low awareness of funder policy and this includes the availability of funding in many cases. Today almost a third of funders with data mandates explicitly support the use of grant money for research data management (allocation of 1-5% of grant), some also for data publications, according to information manually tracked at Springer Nature on funder and institutional policies globally.

3. Explicit funding

Few funders explicitly make funding available for data management

27%

of researchers do not know how they would meet the costs of making their research data openly available

How researchers would meet the costs of making data open

37%

likely to use their own funds for data sharing

39%

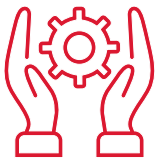
would use money specifically for this purpose from a funder

41%

would use funds identified in their grant

24 European Commission High Level Expert Group on the European Open Science Cloud (2016): Realising the European Open Science Cloud. https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf#view=fit&pagemode=none

4. Practical help: organising data, finding appropriate repositories and faster, easier routes for sharing



4. Practical help

Researchers require faster, easier routes to optimal ways of sharing data. Ideally data management would be embedded within both research and publishing workflows; from the start of a project with a data management plan that is put into practice, to end of the project with data being appropriately shared. The solutions outlined here require collaboration between researchers, institutions, funders, publishers, repositories, and other research data infrastructure providers.

The solutions outlined here require collaboration between researchers, institutions, funders, publishers, repositories, and other research data infrastructure providers

Infrastructure

The Human Genome Project (HGP) was ground breaking for research, fundamentally changing attitudes to data sharing and research practice as a whole.

- Whilst the project was ongoing, all generated data were validated, and then immediately made publicly accessible for all to use.
- All data (the sequence of the entire human genome) remain open to anyone who wants to use it.
- The data sharing standard set by the HGP for sequence data remains in place to date, all sequence data must be shared via INSDC repositories.[25]
- The project received a total of USD $3.8 billion investment from the US government from 1988 to 2003. A report by the Battelle Memorial Institute estimated that the project generated more than USD $796 billion in economic impact.[26]

The success of HGP in setting persistent standards for sequence data sharing highlights the importance of tools and infrastructure for data sharing. The provision of repositories for HGP data, which have then gone on to be used by all researchers generating sequence data, provided a technical solution to support the cultural shift in this community for data sharing. This cultural shift to using repositories was further supported by the publishing community, making it a requirement to share sequence data in an INSDC repository in order to publish articles on such work.

Data shared via a repository are more likely to be findable, and shared in an accessible, reusable format. As such, repositories – whether generalist, subject specific, or within an institution – are essential for data sharing to happen effectively.

More effort is needed globally to convince researchers of the benefits of sharing data via repositories, and to make it easier for them to do so. In *Practical Challenges*, we specifically focused on researcher behaviour at the point of publication, recognising this to be the time researchers have completed data collection and analysis, and are ready to share their data.

- We found a slightly lower proportion of researchers share their data via a repository (41%) than as supplementary information (42%) at the point of submitting a publication.

41%

researchers share their data via a repository

42%

as supplementary information at the point of submitting a publication

25 International Nucleotide Sequence Database Collaboration. See https://www.ncbi.nlm.nih.gov/genbank/collab/

26 Batelle Memorial Institute (2011): $3.8B Investment in Human Genome Project Drove $796B in Economic Impact Creating 310,000 Jobs and Launching the Genomic Revolution. https://ec.europa.eu/futurium/en/content/38b-investment-human-genome-project-drove-796b-economic-impact-creating-310000-jobs-and

- For researchers in China and Japan, the proportion sharing data via supplementary information was even higher, with 55% of respondents in the Japan survey and 75% of respondents in the China survey submitting files as supplementary information at the point of publication, compared with 31% in the Japan survey and 43% in the China survey depositing files in a repository.
- Whilst sharing data as supplementary information is better than not sharing at all, it remains suboptimal as data are not discoverable independently of the article in which they are shared. Data shared via supplementary information are also unlikely to be made available in a format that would enable easy data reuse.

Springer Nature actively encourages the use of community repositories:

- We offer free guidance via our Research Data Helpdesk. The majority of enquiries to our Helpdesk service in 2018 related to appropriate repositories and depositing data (26%).
- Alongside this, our Recommended Repositories list helps researchers identify the best option for them to deposit their work. We currently list over 100 repositories across the biological, medical, physical and social sciences. These are pre-evaluated specialist and generalist repositories that have been assessed for implementation of best practice on data access, preservation and stability. The repository list is shared for use by others under a CC BY license, and has been used by other publishers and journals to provide guidance on repositories for their own authors.
- Our optional Research Data Support service, which provides curation and hosting of data in figshare, carries out checks to ensure data that have a community specific repository are deposited there, rather than in figshare.

Much research data do not have an obvious community repository, evidenced by general repositories, such as figshare and dryad, being the most well used repositories by authors publishing in *Scientific Data* and other Springer Nature journals. General data repositories – which also include Zenodo, Dataverse and the Open Science Framework – as well as institutional data repositories are important mechanisms for sharing many types of research data.

Organisations like the Digital Curation Centre (DCC) also provide expert advice and practical help for research data management, as do research offices, scholarly communications teams and libraries at many institutions worldwide.

Support with data deposition and curation

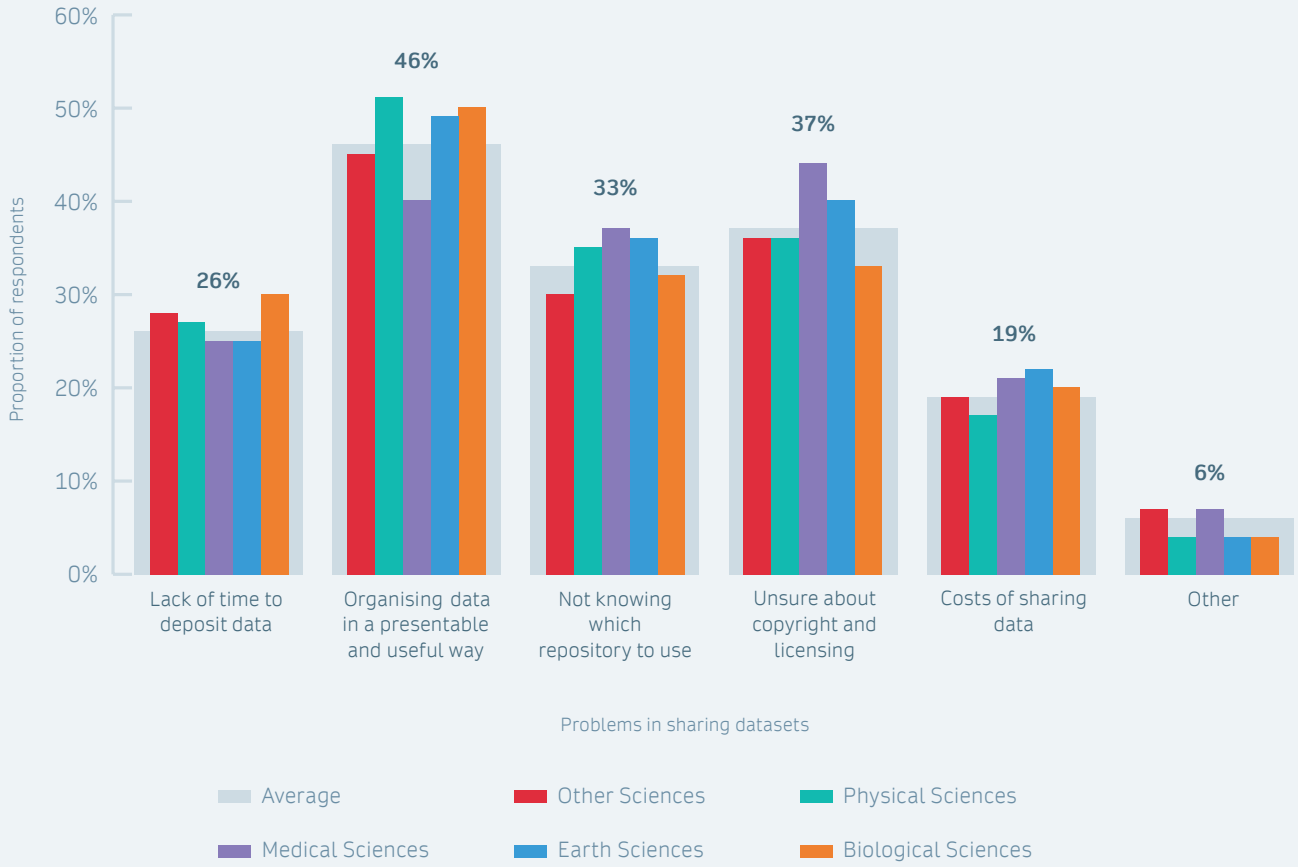
In *Practical Challenges*, "organising data in a presentable and useful way" was selected by nearly half of respondents (46%) as a challenge they faced in sharing their data. The same challenge was the 5th highest identified in the *State of Open Data Report*, selected by 36% of respondents. For *Practical Challenges* there were notable differences by subject area: challenges in organising data were highest for researchers in the physical sciences (57%), compared to 40% in the medical sciences.

Lack of time to deposit data was another challenge identified by *Practical Challenges*, selected by 26% of respondents. Differences by career stage were evident, with more senior researchers likely to be hampered by time to deposit data, compared with early career researchers (selected by 29% of the most senior researchers, and by 23% by early career researchers).

26%

We offer free guidance via our Research Data Helpdesk. The majority of enquiries to our Helpdesk service in 2018 related to appropriate repositories and depositing data

Practical Challenges:
Problems in sharing datasets in different subject areas (n=7,719)



Practical Challenges highlighted the need to support researchers with information on how the process of data sharing works: "whether to use metadata schemes", "selecting the right level of detail at which to share data sets", "where and how to share" and "not sure where to put the data" all feature in open text comments. This is why helping researchers to deposit, describe and share their data, using good metadata, remains a priority for Springer Nature.

- Through optional services such as Research Data Support, we are enabling researchers to make their data findable, accessible and reusable.
- Our new research data training, run for institutions as part of the Nature Research Academies programme, further supports researchers with the knowledge and skills to conduct good data management and practice.
- Through data publishing (including in *Scientific Data*, and *BMC Research Notes*), we aim to make scientific data more reusable, discoverable, interpretable, and citable.

5. Answers to common questions: training and education

5. Training & education

We have emphasised throughout this report the need for collaboration across the research community in building familiarity with data management best practice, and the benefits of data sharing. For there to be low awareness even within subject communities with established norms for data sharing shows that there is a long way to go here.

Alongside this, we support Digital Science's focus on building awareness of FAIR principles. This was highlighted within the *State of Open Data Report* as a particular area of concern, with 60% of respondents having never heard of these principles or their relevance in enhancing the reusability of academic data.

There is a particular need for education amongst early career researchers. In *Practical Challenges*, 40% of early career researchers cite not knowing where to share data, versus 30% of senior researchers. Similarly 43% of early career researchers cite uncertainty about copyright and licensing, versus 33% of the most senior researchers.

Access to support

In the *State of Open Data Report*, 57% of respondents identified their peers as the best people to help review, curate and prepare their data for making it publicly available, followed by publishers (52%) and libraries (42%). Similarly for DMPs, in our survey with researchers in China, the research office or other researchers came out most prominently as best placed to help researchers achieve the aims of their DMP (48% for the research office, 37% others in their research group, and 36% their supervisor).

However there does not appear at this time to be adequate provision of training for the majority of researchers.

- In the *State of Open Data Report*, 65% felt there was not sufficient training, support and advice in regard to data management.
- For researchers in China, for around half of respondents there was no support available for data management from their institution, department or funder (47% said there was no funder support, 44% cited no institutional support).
- In our survey with researchers in Japan, only 34% agreed or strongly agreed that they were "confident in creating a good DMP", and only 13% agreed or strongly agreed that there is enough information and help available to create DMPs.

More practical support is needed if we are to see any real increase in data sharing and data best practice. As noted elsewhere in this report, there are many initiatives that can be built on and extended: whether that be within the community, such as DCC; via publisher services such as Springer Nature's research data training, Helpdesk, Research Data Support or educational events such as Better Science Through Better Data; or via institutional initiatives such as highlighted from TU Delft.

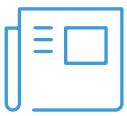
The following represents some of the most commonly raised areas for further education and training.

65%

of researchers feel there is not sufficient training, support and advice in regard to data management

Who is best placed to support data management?

57%
peers



52%
publishers

42%
libraries

More practical support is needed to increase data sharing and data best practice

Copyright

In both *Practical Challenges* and the *State of Open Data Report*, knowledge of copyright and licensing was the second highest concern with sharing datasets (37% and 38% respectively). The *State of Open Data Report* also found that 65% of repondents did not know what license they had previously published under. This was a clear area researchers wanted more information, with 61% citing this as an area they need help with.

- There may be particular regional requirements for training on this topic also, recognising data protection laws vary by geography, such as China's cyber security law.[27]
- Differences by discipline were identified in *Practical Challenges*, with medical science researchers reporting that copyright and licensing issues were their biggest challenge (the only discipline for whom "organising data in a presentable and useful way" was not ranked highest). Medical science researchers were also most likely to report concerns about data sensitivity and misuse, and concerns about protecting research participants, consistent with other surveys of clinical researchers.
- *Practical Challenges* found further variation by career stage, with more early career researchers unsure about copyright or licensing compared with senior researchers.

Repositories

Not knowing which repository to use differs by subject area. In *Practical Challenges*, this was most noted as a challenge in the medical sciences (37%), compared to 27% in the physical sciences citing this as an issue. The smaller proportion in the physical sciences may reflect that there are a smaller number of recognised repositories in this field than in other subject areas.

Misuse of data

Unlike *Practical Challenges*, the problem most respondents highlighted in the *State of Open Data Report* was "concern about misuse of data" (38%). This was similarly evident in our regional surveys of researchers in China and Japan, where this was the highest rated concern at 48% in China, and 49% in Japan. This was not evident in *Practical Challenges*, where "fear of misuse and being scooped" was raised rarely, accounting for only 12% (46 of 385) of all "other" problems given. This is an area we need to begin to understand better: what do researchers consider "misuse" and what are they concerned about? Would better understanding about rights to share, copyright and licensing allay this, or are concerns deeper about data being selectively used by others, or used without citation or acknowledgement of the original author?

Sensitive data

Content analysis of free text responses to "other" problems in *Practical Challenges* revealed the challenges of managing sensitive data. This was particularly high amongst medical researchers (66% of "other" problems reported by medical researchers), who may be dealing with patient data and other data relating to human research participants. As respondents noted, the challenges include the complexity of anonymising the data, legal requirements for human data, the costs involved in anonymisation, and the time-consuming nature of doing so. Regionally there were other challenges around sensitive data not related to discipline. In our survey with researchers in China, a high number of respondents commented that data relating to unpublished findings (75%) and data relating to national security (70%) or personal data (70%) would not be shared publicly.

27 Bird, R. (2018) Where are we now with data protection law in China. http://knowledge.freshfields.com/m/Global/r/3824/where_are_we_now_with_data_protection_law_in_china

Cultural attitudes to sharing

During our research we identified some specific cultural differences in the sharing of data and attitudes towards sharing. For example, the *State of Open Data Report* found that globally 70% of researchers were sharing data both publicly and privately, whereas for researchers in Japan our findings were much lower at 59%. For researchers in China, we found a much higher importance is placed on sharing data with immediate colleagues and collaborators than more broadly, with most private sharing either directly with colleagues from their institution (61%) or with known peers (55%). For researchers in Japan, we also saw highest importance placed on sharing with peers they know (66%) and colleagues from their institution (41%). Significantly more researchers in China considered it important to share their data with other researchers in China (64%), with only 45% considering it important to make openly available for anyone to use.

Size of data

In *Practical Challenges* we found that the size of the dataset can have a direct impact on whether researchers shared their data or not. A high number of researchers in the study were generating small datasets. Of these, researchers that generated the smallest sized data (<20MB) were sharing the least data in either repositories or as supplementary information. In contrast, 70% of researchers with data above 50GB were sharing these data, predominantly via repositories. This highlights there is further education needed on the value of sharing small datasets.

Data Management Plans (DMPs)

Globally, 70% of researchers have made a DMP, according to the *State of Open Data Report*. However, less than 40% (39.70%) of these respondents were creating DMPs for more than half of their research. Regionally, even more researchers were creating DMPs: in our China survey, 93% of researchers said they had made a DMP before, however in 36% of cases this is only rarely. The overall percentage of researchers in Japan was also higher than the global average at 56%.

- Unfamiliarity is the biggest challenge for DMPs. In the *State of Open Data Report*, 48% of respondents had never heard of a DMP. This was 50% for researchers in China, and 46% in our Japan research, showing a similar low level of awareness in all three surveys.
- Those who have created a DMP before recognise it is good practice to do so. “Efficient and effective management of data” and “good practice when undertaking research” appeared as the top reasons for creating DMPs in all three surveys.
- Increasing familiarity will boost not only the number of researchers creating DMPs but also the likelihood that they will do so again in the future. 69% of researchers in China reported that they are “likely” or “extremely likely” to create a DMP in the next few years (this was also reported at 58% for researchers in Japan), with those who created them in the past more likely to do so again in the future. Increasing confidence and knowledge about how to create a good DMP will encourage this repeat behaviour.

For further information

All reports and their underlying data have been made openly available in the figshare repository. Where additional resources, such as free infographics, are available, these can also be found in the links below.

Methodology

All of the data referred to in this report has been collected via surveys run between 2017-2019. Where comparisons have been drawn we have been careful to use like for like questions. The methodology of running each survey has varied slightly as have those people who have been invited to take part, meaning that the demographic profile of the response bases to each survey do vary. Additionally, as with any research that is survey based, the respondents have been self-selected to an extent which can lead to certain bias in the data. To view more details about each survey, access the individual reports and raw data by using the links listed below.

Practical Challenges for Researchers in Data Sharing

White paper: https://doi.org/10.6084/m9.figshare.5975011
Full survey dataset: https://doi.org/10.6084/m9.figshare.5971387.v2
Infographic: https://doi.org/10.6084/m9.figshare.5996786.v4

The State of Open Data Report 2018

White paper: https://doi.org/10.6084/m9.figshare.7195058.v2
Full survey dataset: https://doi.org/10.6084/m9.figshare.7234985.v1
Interactive data: https://knowledge.figshare.com/articles/item/state-of-open-data-2018

Challenges and Opportunities for Data Sharing in China

White paper: https://doi.org/10.6084/m9.figshare.7718441.v1
Full survey dataset: https://doi.org/10.6084/m9.figshare.7321604.v1
Infographic: https://doi.org/10.6084/m9.figshare.7782761

Research Data Sharing in Japan

White paper: forthcoming, 2019
Full survey dataset: https://doi.org/10.6084/m9.figshare.6328952.v1
Infographic: https://doi.org/10.6084/m9.figshare.6609056.v1

Acknowledgements

We would like to thank the following individuals and organisations for their contributions to this white paper:

Digital Science and figshare (The State of Open Data)
Kazuhiro Hayashi, Senior Research Fellow, National Institute of Science and Technology Policy, Japan (Japan survey)
National Science Library, Chinese Academy of Science (China survey)
Dan Penny, Director, Market Intelligence, Springer Nature
Kelly Duoos, Marketing Manager, Springer Nature
Kerstin Weinheimer, Strategy and Market Intelligence Manager, Springer Nature
Roza Sakellaropoulou, Brand Engagement and Marketing Manager, Springer Nature

Springer Nature advances discovery by publishing robust and insightful research, supporting the development of new areas of knowledge and making ideas and information accessible around the world. We are committed to supporting researchers in sharing their research data. Find out how we could help you https://www.springernature.com/research-data

Join the research data community to tell us your views http://researchdata.springernature.com

Understanding past climates in Antarctica
The permafrost at Table Mountain in Antarctica is thought to be more than 6 million years old and holds valuable clues to what the environment was like in eras when temperatures and sea levels were higher than they are today. With recent studies showing that the environmental temperature in which an organism lives is encoded in their DNA, an international research team led from New Zealand set out to sample sediment cores to obtain ancient bacteria. These individual bacterial cells were individually removed from the cores, DNA extracted and are currently being sequenced to reconstruct their genomes in order to help to predict the temperatures under which the bacteria lived.

Picture: NatureJobs Scientists at Work competition