

## Digital Science Report

# The State of Open Data 2020

The longest-running longitudinal survey and analysis on open data

Foreword by Dr Leslie McIntosh, CEO of Ripeta and Executive Director, Emeritus - Research Data Alliance US

December 2020

## About Figshare

**Figshare** is a repository where users can make all of their research outputs available in a citable, shareable and discoverable manner. Figshare's aim is to become the place where all academics make their research openly available. It provides a secure cloud based storage space for research outputs and encourages its users to manage their research in a more organized manner, so that it can be easily made open to comply with funder mandates. Openly available research outputs will mean that academia can truly reproduce and build on top of the research of others. Visit [www.figshare.com](http://www.figshare.com)

## About Digital Science

**Digital Science** is a technology company working to make research more efficient. We invest in, nurture and support innovative businesses and technologies that make all parts of the research process more open and effective. Our portfolio includes admired brands including Altmetric, CC Technology, Dimensions, Figshare, Gigantum, GRID, IFI Claims, Overleaf, ReadCube, Ripeta, Symplectic and Writefull. We believe that together, we can help researchers make a difference. Visit [www.digital-science.com](http://www.digital-science.com)

## Acknowledgements

**Figshare** and **Digital Science** are extremely grateful to **Springer Nature**, who have been our partner in scoping the survey and have provided survey design, hosting and global distribution. We would also like to thank all of the contributors for their articles included in this report.

This report has been published by Digital Science which is part of the Holtzbrinck Publishing Group, a global media company dedicated to science and education.

Digital Science, 6 Briset Street, London, EC1M 5NR, UK. [info@digital-science.com](mailto:info@digital-science.com)

Figshare, 6 Briset Street, London, EC1M 5NR, UK. [info@figshare.com](mailto:info@figshare.com)

This work is licensed under the Creative Commons Attribution 4.0 International License.



# Contents

1.	<b>Foreword</b>	2
	Dr Leslie McIntosh, CEO of Ripeta and Executive Director, Emeritus - Research Data Alliance US	
2.	<b>If a Researcher Would Meet a Librarian</b>	4
	Mariëtte van Selm, Information Specialist (University Library), University of Amsterdam	
3.	<b>How Repositories Can Help Drive Positive Change in Data Sharing</b>	7
	Kathleen Shearer, Executive Director of COAR, Merce Crosas, University Research Data Management Officer at HUIT and Chief Data Science and Technology Officer at IQSS, Brian Nosek, Co-Founder and Executive Director of the Center for Open Science and Mark Hahnel, CEO and Founder of Figshare	
4.	<b>What is the State of Open Data in 2020? It All Starts With a Good Plan</b>	17
	Alan Hyndman, Marketing Director at Figshare and Grey Goody, Data Analyst at Springer Nature	
5.	<b>Research Practices in the wake of Covid-19</b>	22
	Grace Baynes, VP of Research Data and New Product Development at Springer Nature, and Mark Hahnel, CEO and Founder of Figshare	
6.	<b>Contributor Biographies</b>	26

"Covid-19 has illuminated the needs and capabilities in making science open and accessible and perhaps surprisingly, in doing so, suggested that science truly can be accelerated."

# Foreword

**Dr Leslie McIntosh, CEO of Ripeta and Executive Director, Emeritus - Research Data Alliance US**

The 2020 State of Open Data report provides an interesting lens to view how far open research has come, and to look at opportunities for improvement in data sharing. Every time we push science forward, we should also both reflect the past and predict the future benefits and challenges of our actions. 2020 has offered tests and trials like no other in my lifetime. This time has also shone a light into the gaps in our thinking as well as in our progress towards open research. This fifth edition of the State of Open Data report reveals the current thinking on open science from a global pool of over 4,500 respondents. We have the opportunity to simultaneously reflect on how this movement has transformed into practice as well as thoughts to consider on moving forward in this space.

Over the past five years, the science ecosystem of researchers, librarians, publishers, institutions, funders, and others have embraced improving data sharing. Policies requiring more research transparency and data sharing have emerged alongside communities aiming to improve scientific scholarship. Yet, have the policies and open data conversations affected practice?

Let us look at the effect of requiring data availability statements in publications – a mechanism to ostensibly make it easier to find data and encourage research data sharing. However, when one looks within a publication, there is frequently a statement such as ‘Data available upon request’, or to be either more polite or more cynical – ‘Data available upon reasonable request’. This may be why a majority of survey respondents want stronger mandates for data sharing, with enforcement. While data sharing policies have not fully translated into practice, the conversations and atmosphere of sharing research has changed.

Covid-19 has illuminated the needs and capabilities in making science open and accessible and perhaps surprisingly, in doing so, suggested that science truly can be accelerated. The scientific community, now indoctrinated in open science, has embraced the principles of sharing their research openly. Disparate communities – academic, governmental, commercial – have cooperated to create an array of solutions needed to combat Covid-19. Researchers, repositories, funders, and more have independently and jointly made Covid-19 related scientific research freely and openly available with necessary protections for private data.

The Research Data Alliance (RDA) convened over 200 global volunteers to report<sup>1</sup> on the needs of data sharing culminating in a report released in June 2020 to guide data sharing during a pandemic. Open Science and data sharing practices cited in the RDA report range from adhering to the FAIR (Findable, Accessible, Interoperable and Reusable) principles to documenting methodologies, incentivising early publications and expediting review processes, implementing legal frameworks for cross-jurisdictional data sharing balanced with ethical and privacy considerations. These principles and recommendations are salient to open science and open data at any time.

As we look to the future, what will be needed to intentionally improve open data? Three interrelated topics - trust, misuse, and equity - must continue to shape future conversations to enrich and protect our Open Data ecosystem.

As discussed in the previous and current State of Open Data reports, the issue of trust and data misuse weigh significantly in the minds of many and in particular of researchers. Researchers have concerns in having data misused and of others finding errors in their data. Yet there are other types of trust and data misuse that have been in the spotlight this year. The algorithmic biases embedded in the architecture of this digital age continuously surface at an alarming rate. We know that the algorithms have been developed from biased data, but how much has this affected or will affect open data?

Misuse of open data and open science practices can be intentional or unintentional, but its presence is undeniable. For example, GitHub was originally conceived for open code sharing and exchange but is now 'misused' to store scientific work one might argue belongs in repositories. Without formalities of metadata tagging and organisational structure, these open data are accessible but not necessarily findable or interoperable. Yet, the data are more available than sitting on a local computer.

On a nefarious front, as politics and science have collided, misinformation efforts have infected the scientific structure as they have in politics. Questionable scientific 'research' has been uploaded on established research sharing platforms then highlighted in the news as a source of legitimate scientific information. There has also been a growing misuse of pre-published research when moving research into the public eye before the methods, results, and conclusions have been rigorously scrutinised. Thus, our natural science ecosystem has been misused outside the non-scientific community.

The last point pertains to equity in open science and open data. This movement must be open and equitable for all. As we move toward greater transparency and data sharing, let us also remember the expanded reuse, and impact of data. For example, while the FAIR principles suggest what should be done with data to enhance reproducibility, the CARE framework (Collective benefit, Authority to control, Responsibility, and Ethics) provides context for how data should be honoured as a rich resource and put into a broader context of use and understanding.<sup>2</sup> While written to address the rights and needs of the indigenous communities, the framework and principles outlined in Indigenous Data Sovereignty could be used to guide broader conversations around data equity.

## So what is our hypothesis for moving open data forward?

The mantra in this space has been to make data as open as possible and as closed as necessary. This has been an excellent beginning, but we need more than this. The need to understand and incorporate trust mechanisms should be implemented as open science efforts are built into the future. As we scale data and data sharing, let us also keep in mind how to trust the data, algorithms, and artefacts of data.

We need to employ checks. This means scrutinising the ways in which open data practices work and establishing mechanisms to verify both the research and the processes. Now more than ever, the research needs to be open, equitable, and verified.

"Trust, misuse, and equity must continue to shape future conversations to enrich and protect our Open Data ecosystem."

<sup>1</sup> RDA Covid-19 Working Group. Recommendations and Guidelines on data sharing. Research Data Alliance. 2020. DOI: <https://doi.org/10.15497/RDA00052>

<sup>2</sup> Kukutai, T., Carroll, S. R., & Walter, M. (2020). Indigenous data sovereignty. In D. Mamo (Ed.), *The Indigenous World 2020* (34th ed., pp. 654–662). Copenhagen, Denmark: IWGIA. <https://researchcommons.waikato.ac.nz/handle/10289/13633>

"A data management plan is just one of those hoops researchers need to jump through, and they'll jump, only because of the funding involved and fear of repercussions if they don't."

# If a Researcher Would Meet a Librarian...

**Mariëtte van Selm, Information Specialist (University Library),  
University of Amsterdam**

If there's one thing I have discovered in the eight years I've now been supporting researchers in research data management, it is that we have grown to expect researchers to be several personas in one. Of course, we expect them to be experts in their own discipline and know what discussions are going on, which new methods are being devised and which research topics are in (or out of) fashion.

But we also expect researchers to know a thing or two, preferably more, about information security: Where and how to store data in a way that permits using the data not only now but in the future as well, without the data being accessible to people who have nothing to do with the data, at least for as long as the data is confidential. If researchers know how to convert files to more open formats or how to use encryption, that's a plus.

Thirdly, researchers should, whatever discipline they're from, be legal experts and in that capacity know how to successfully navigate a sometimes very difficult landscape of intellectual property rights, licenses, privacy and maybe patenting as well. If they know their way around legalese, even better.

Nearing the end of their research project, researchers should be able to put themselves in other researchers' shoes: What search terms would another researcher, sometimes from a totally different discipline, use to find my data and what metadata should my data have to be found by the right researchers? Learning to answer questions like that takes a librarian three to four years of education, but we expect researchers to be able to without any training whatsoever.

It would moreover be nice if a researcher would be an archivist as well and understand what to do to digitally preserve the data they have produced or procured in their research, and to have the data stay accessible for at least ten years from now.

And finally, we'd very much like researchers to be clairvoyants, so they can judge what data will be relevant, why and for whom in ten years or more from now and therefore must be kept, and what data can be deleted without any serious consequences.

Experts in their own field and in information security, legal expert, librarian, archivist and clairvoyant – all that, to be accomplished without getting any extra time for the tasks involved, of course.

And time may very well be an important issue here. Researchers already need to write a grant proposal that stands out from a deluge of other proposals for the same research budget. They must respond satisfactorily to a review from an ethics review board sometimes, either at their own

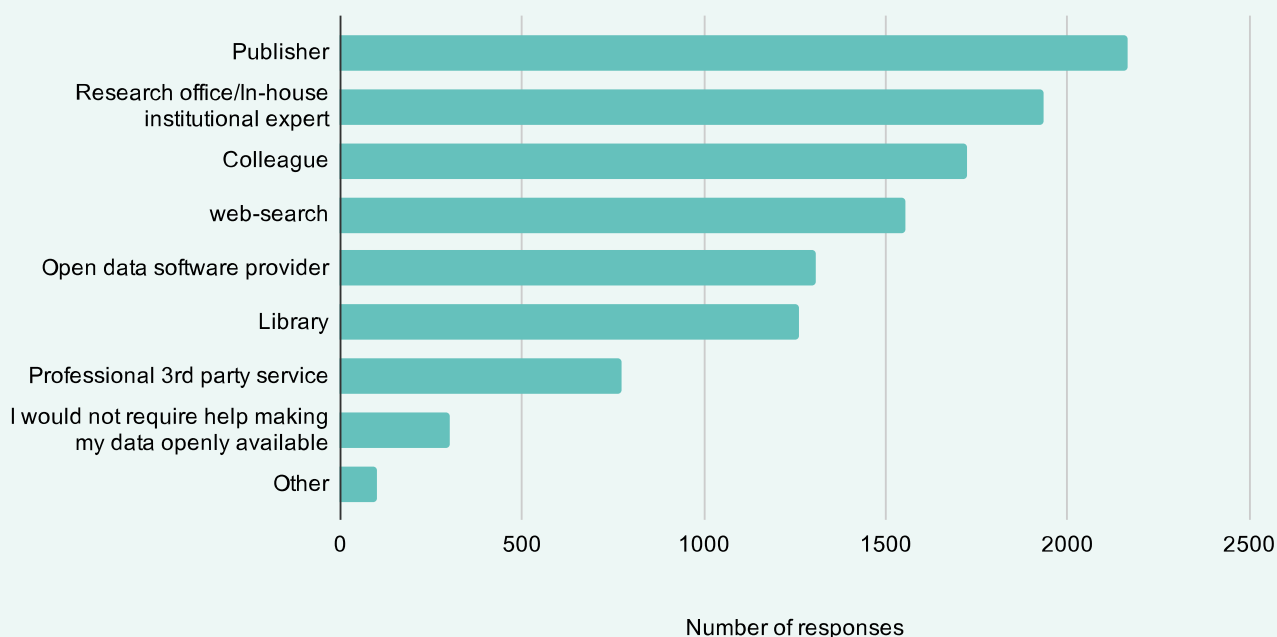
institution or at the funder or even both. They must figure out how to do most with a budget that's always too small. All while they are chomping at the bit to start their research, and don't want to be bothered with questions about research data management. I shouldn't want to support writing a data management plan by explaining what is required and why, I should just deliver a text they don't need to understand but can copy-paste. A data management plan is just one of those hoops researchers need to jump through, and they'll jump, only because of the funding involved and fear of repercussions if they don't.

"Libraries are – or should be – the spider in the web of all data support."

However, things are improving. In data management plans now, the Dutch Research Council (NWO) ask for the name of the research support staff member and the date they consulted on the plans, which are required after funding is awarded. Most researchers still don't want to take an hour to sit down with me – if they do, I usually stay two or three hours, at their explicit request, because they come up with a host of issues they've never allowed themselves the time to search support for – but slowly but surely, I am getting a foot in the door and can clear up any misunderstandings a researcher might have.

One of those misunderstandings is clear from previous editions of this report: researchers do not consider the library as the first port of call when it comes to getting help with data. And they don't have to, provided they have colleagues around who are able and willing to provide them with the help they need. And I'm not saying the library is the in-house data expert, since a lot of libraries are not.

If you required help in making data from your most recent research paper openly available, which sources would you rely upon?



But libraries are – or should be – the spider in the web of all data support. A lot of that support they can provide themselves. They know what metadata, metadata standards and ontologies are, and can explain how to use them. They can ask the right questions when it comes to preservation and selection of data: how much time and money went into collecting the data and is the data time-sensitive? They know how to search the internet and therefore should be able to point researchers in the direction of the right data repository for that researcher and that data.

Librarians know enough about copyright to be able to make Creative Commons licences easier to understand, tell a researcher what ‘NC’ does and doesn’t permit, for example. And they should know who to ask if legal or information security expertise is needed in situations that require more specific knowledge, for instance in the case of data gathering by scraping social media. The most important thing of all is that they can translate information into language a researcher understands and know how to discern between ‘need to know’ and ‘nice to know’ in conversations with time-pressed researchers.

Yes, researchers who want to become a little more self-sufficient in handling their data will have to learn some things. But no researcher has to know or be everything. Every conversation between a researcher and a data librarian will make the both of them a little wiser: the librarian learns more about the research that is going on at their institution, the researcher learns what to take into account and why. And all it costs is a little time.



# How Repositories Can Help Drive Positive Change in Data Sharing

Kathleen Shearer, Executive Director of COAR, Merce Crosas, University Research Data Management Officer at HUIT and Chief Data Science and Technology Officer at IQSS, Mark Hahnel, CEO and Founder of Figshare and Brian Nosek, Co-Founder and Executive Director of the Center for Open Science

## 1. What in your opinion are the biggest improvements that need to happen in the open data space in the next five years?

### Kathleen Shearer

We need to address three important issues concurrently: The first is capacity building, to increase support for data management and to develop more expertise within the research community. For data to be understood and reused by others, they must be managed properly, ideally from their inception. But this can be extremely resource intensive and require significant knowledge, knowledge that many researchers currently do not have. Bringing data management experts into the research team, as well as improving institutional data management services to help researchers, will go a long way to improving the current situation. In particular, I think research libraries have a much bigger role to play in supporting data management. The other two issues, which I will address in more detail in some of the following questions, are the lack of incentives and gaps in infrastructure. Researchers don't have a strong incentive to share their data at the moment, and we need to reform research assessment systems to better reflect the principles of open science and data sharing. The third issue is the lack of infrastructure for open data. More investments need to be made in data repositories, storage and preservation and other types of data sharing infrastructures.

### Mercè Crosas

There are four main aspects of the Open Data space that I believe need to improve in the next five years: 1) First, we should recognise that currently "open" data do not always include all the data necessary for research and scientific discovery. While Open Data mandates are essential and should be encouraged and supported whenever possible, we also need to acknowledge that much of the needed research to solve critical societal problems require proprietary data from companies or sensitive data collected from individuals. Thus the solution requires not only mandates but also improved means to make use of such data while still protecting privacy or stewardship. 2) As Open Data, or data in general, become larger and more complex, we should connect Open Data access with computational, analytical, and visualisation tools and locate these close to the data, eliminating the need to download data to get insights. 3) As we rush to make larger quantities of Open Data available, we cannot sacrifice quality and rigor. We need to provide credible ways to

"The solution is for the policy and infrastructure landscape to support the entire data lifecycle – upfront planning and curation along the way will mitigate publication bias, align with researcher incentives, and improve the quality and efficiency of the shared data."

decipher the quality of the data by asking, "does a dataset have adequate information to be understood and used by those who did not create it? Are the data complete, or only representative, or otherwise biased?" 4) Although we have made definite improvements that support data citation in repositories and journals (with persistent identifiers, proper attribution, and standardised citation metadata), it has not yet resulted in giving credit to data authors for sharing their data and recognising that credit in a way that becomes an incentive to share more data.

#### **Mark Hahnel**

Indicators of best practice would be a huge step up in professionalisation of the data publishing space. Over the past decade, we have seen the explosive growth of researchers making datasets available. At Figshare, we saw the positive effect of human intervention for metadata improvements on NIH (National Institutes of Health) funded datasets. What we don't have, however is standards to filter content on. There is a growing amount of indicators to filter on, such as FAIR (Findable, Accessible, Interoperable and Reusable) data stamps or citations demonstrating reuse. Improvements and standardisation across data infrastructure here will move us from "show me data relevant to my research" to "show me data that is reusable, that is relevant to my research".

#### **Brian Nosek**

Open data infrastructure exists and policies are shifting toward incentivising or requiring preservation for open or controlled access sharing. That work isn't done, but it is on the right track. During the next five years, many more researchers will be sharing data, but they may not be sharing data well. There is a learning curve to effective data sharing, and there are weaknesses in the policy and infrastructure landscape for supporting effective data sharing. The biggest weakness is that the policy and infrastructure focus is on implementation of data sharing tasks retrospectively. Most policy and tools support data sharing "upon publication." The problem is that this (1) misses data that is never published, particularly negative results due to publication bias, (2) comes after the key incentive is resolved (paper acceptance) essentially guaranteeing low motivation to do the work, and (3) is highly inefficient to retrospectively gather and prepare data for sharing. The solution is for the policy and infrastructure landscape to support the entire data lifecycle - upfront planning and curation along the way will mitigate publication bias, align with researcher incentives, and improve the quality and efficiency of the shared data.

## **2. Who can have the biggest effect on driving social change within academia when it comes to open data?**

#### **Kathleen Shearer**

I think it is still the funders who can have the biggest influence on data sharing practices. To date, research funders have been the most effective in advancing data sharing through the introduction of policy requirements and I expect more open science policies will be adopted in the coming years. Ideally, open data will be advanced collectively through a coordinated approach involving multiple stakeholders so that

policies are aligned with incentives and infrastructure development. Research funders adopt policies that require data sharing. Institutions adopt incentives that recognise and reward data sharing practices, and offer local services to support data management. Governments fund the development and ongoing operations of platforms and infrastructures. When all these stakeholders work together, for example at the national level, this is when we really start to see traction around open data. Research communities also have a role to play by helping to change the norms around data management from data ownership and control, to open data and data sharing. I like to use the metaphor of a three-legged stool, with policies, infrastructures and culture each representing one leg. All three legs of the stool must be a similar length, otherwise the stool will not be stable enough to sit upon.

### **Mercè Crosas**

Three main stakeholders come immediately to mind: 1) the leadership of academic institutions must support and incentivise open data and, in general, require sharing of all research outputs (code, workflows, data) along with scholarly publications, 2) journals' data policies must require data sharing with article publication, and 3) funding agencies must mandate data management and sharing plans.

In the last years, we have seen that policies that mandate data sharing and archiving have improved research data access. Furthermore, the number of such data policies has increased (see, for example, Vines et al., 2013 [doi.org/10.1096/fj.12-218164](https://doi.org/10.1096/fj.12-218164) or Crosas et al., 2018 [10.31235/osf.io/9h7ay](https://doi.org/10.31235/osf.io/9h7ay)). There will only be a substantial social change, however, when the mandates have become almost irrelevant because making data accessible has become the norm; only then will others always be able to verify and build upon prior research. Some scientific disciplines have made much more progress than others in this area (see, for example, in political science Key, 2016 [doi.org/10.1017/S1049096516000184](https://doi.org/10.1017/S1049096516000184), AGU's position statement on earth and space science data, and astronomy's standard practices).

Finally, public trusted repositories also play a critical role in enabling Open Data. But the technology necessary to publish Open Data already exists in the current repositories, so availability is not what constrains a more uniform sharing of research data. Nevertheless, we can and should do more to improve the technology to continue facilitating social change, as addressed in the next question.

### **Mark Hahnel**

Show me the incentives and I'll show you the outcome. We now have years worth of data demonstrating the cause and effect of researchers sharing data. Previous State of Open Data surveys have highlighted researchers responding to publisher and funder policies/mandates. This is a global effect and we should continue to see the effects of researchers jumping through hoops to further their careers. To put this in context, researchers must publish papers and win grants to be successful. These two actions are closely linked and will continue to be the driving incentives behind researcher behavior. As more funder and publisher data-policies

"When all the stakeholders work together, for example at the national level, this is when we really start to see traction around open data."

"Show me the incentives and I'll show you the outcome."

"Many researchers still do not have a suitable repository available to them; and this is especially true for researchers in the global south."

are released and as they carry more weight, the more data publishing will become a critical yet normal step in the career of a researcher.

#### **Brian Nosek**

The primary challenge for culture change in science is solving the coordination problem. Science is highly decentralised with many incentives and policy makers among the funders, publishers, and institutions, and a highly active but siloed research community. There is no single agent that can shift the default to open. The only way to achieve true culture change is to activate each agent's sense of ownership over the policies and behavior that they control, and to facilitate the communication and collaboration across agents to shift the norms, incentives, and policies in concert. Grassroots communities play an essential role by giving voice to the desire for change and demonstrating with innovators and early adopters that it is possible, even desirable. Stakeholder communities play an essential role by demonstrating with policy, training, and other actions that they are contributing to change.

### **3. We have subject specific repositories and a wealth of generalist repositories, what infrastructure is missing today and needs building?**

#### **Kathleen Shearer**

The infrastructure for data management is improving and expanding, but there are still many gaps. Re3data, the directory of research data repositories, currently lists close to 2,500 repositories, but the vast majority are based in Europe and North America. 2,500 data repositories sounds like a lot, but many researchers still do not have a suitable repository available to them, and this is especially true for researchers in the global south. Many repositories are restricted to collecting data from a certain domain only, or from only a specific institution or region. Additionally, many repositories are not capable of collecting large and/or more complex data sets, due to the technology they are using or the resources available to them. The optimal scenario is one where there is a healthy mix of sustainably funded domain, institutional, national and regional repositories that can support data sharing across disciplines and regions. While it is unrealistic to think that each institution will be able to maintain a repository, there should be local platforms available to most researchers so they don't have to send their data outside their own country. Consortial repository models, which are becoming more common, can help to address some of the current gaps because they allow institutions to pool resources, enabling them to offer repository services to their affiliated researchers, sharing skills and expertise across organisations, while also lowering institutional costs.

#### **Mercè Crosas**

Repositories must provide responsible methods to share sensitive and private data, as well as integrate better with computational resources and research tools for data analysis, exploration, etc. These are precisely two areas that we are working on with the Dataverse software platform.

Sharing sensitive data does not need to mean allowing others to access or download the raw data. Data can be made more open by publishing any metadata that can be publicly shared (metadata details might vary depending on data sensitivity) so that the dataset is findable through a public repository. Nonetheless, the data are kept restricted in a data enclave or similar secure remote storage and can only be accessed by authorised data users with approved data use agreements. We are working with the OpenDP community (<http://opendp.io>) to build tools that provide differential private statistics from a sensitive dataset. These differentially private releases, in which a minimum amount of noise is added to preserve the individuals' privacy, allow one to learn about the data without ever accessing the raw data. It is a way to make the data as open as possible, enabling others to derive insights, thereby removing the false tension that data must be fully open or fully closed. Differential privacy tools are not the only solution to privacy-preserving analysis - there is a significant opportunity for contribution by any group that wants to join the challenge.

To improve interoperability between data repositories, cloud computation, and software tools, we need datasets with standard, machine-readable metadata to allow external tools to act on the data. This is partly the vision of a data commons - connecting data repositories with active research and facilitating analysis and computing while tracking data transformations. Agreement on the standards is critical to ensure data commons will talk to each other and not result in data and computing platform silos. We are working in this area to standardise metadata that describes a data package or container. In turn, these efforts can help with data quality because they will facilitate tracking data transformations and documenting the data during the active research.

In both areas - providing tools for privacy-preserving analysis - and integrating with computational and research systems - using open-source software and standards will benefit the outcome. It will allow for more scrutiny about the tool's quality, more credibility with the transparency of underlying algorithms, and more accessibility to all. A call for open-source is aligned with the recently (October 21st, 2020) approved Open Source Strategy by the European Commission ([https://ec.europa.eu/info/news/european-commission-adopts-new-open-source-software-strategy-2020-2023-2020-oct-20\\_en](https://ec.europa.eu/info/news/european-commission-adopts-new-open-source-software-strategy-2020-2023-2020-oct-20_en)).

### **Mark Hahnel**

I have a few ideas of gaps in the infrastructure:

Thematic repositories - Datasets should be published in a subject specific repository if available. However, for the majority of data generated by researchers, there is no subject specific repository and they revert to generalist repositories. By having an intermediate layer of thematic repositories, with more stringent metadata requirements and ideally human curation, the quality of outputs will increase.

Credit mechanisms for data publishing - We have multiple ways to measure the impact of researchers for publishing papers, from the

"There are many repositories. But the real long-term benefits of data sharing will be realized when discovery, integration, and transfer across repositories is easy and efficient."

"The importance of data sharing is on the radar like never before and this is an opportunity to advance the case for open data, as well as further develop the workflows, policies and infrastructure that can be generalized beyond Covid-19."

impact factor to the H-index. What we do not have is an abundance of ways to measure the quality of datasets and metadata, or ways to reward researchers for their commitments to open data and research transparency.

API checker – Movement of information where possible is essential for researchers and in the future machines to query and combine datasets. Many legacy repositories claim to be open, but are not audited to an adequate level to ensure that the tech stack does what it is supposed to do. This is an acute problem for repository APIs.

#### **Brian Nosek**

Connectivity. There are many repositories. Their individual effectiveness, user experience, and support of FAIR sharing can improve, but the real long-term benefits of data sharing will be realised when discovery, integration, and transfer across repositories is easy and efficient. The National Academy of Sciences issued a report this year about repositories defining how they support three data states across the data lifecycle. An effective infrastructure ecosystem would make it trivial for researchers to manage their data during active acquisition and analysis, and then transfer that data into domain-specific repositories to integrate with similar data and wider discovery, and then transfer the valuable data into long-term storage for preservation. Already, repositories exist to do all of these things, but there is not yet a coordinated solution.

#### **4. Where do you see the quick and long term wins when it comes to open data?**

##### **Kathleen Shearer**

For me, the quickest win is the data management plan (DMP). DMPs oblige researchers to think about how they will manage and share data before the research project begins, and this can go a long way to improving the quality of data and metadata, eventually making it easier for the data to be shared. DMP tools are freely available on the internet, so there are very few barriers to their use and they are relatively non-controversial so they can be included in policy requirements without too much push back. Another quick win is the very visible and relevant use case of Covid-19. The importance of data sharing is on the radar like never before and this is an opportunity to advance the case for open data, as well as further develop the workflows, policies and infrastructure that can be generalised beyond Covid-19 after the pandemic has passed. Longer term wins will come when we start to see the impact of data sharing across many disciplines and countries. We need more concrete examples of how data sharing has contributed to research innovations, saving money, and/or improved society across a range of disciplines. This will justify further investments in time and resources for open data. In addition, in the longer term, we need to look beyond data sharing, towards the interoperability of data across disciplines. This is a huge task, but could also result in huge breakthroughs.

### **Mercè Crosas**

I think that we already have had a quick win – see all the open data available through domain-specific or generalist repositories that did not exist even ten years ago and are being downloaded daily. A few innovators and early adopters have driven these efforts to make data more open to improving research and scientific discovery. Whenever there is an emergency, more join us. For example, we've seen incredible efforts from the Open Covid-19 Data Curation Group that in a short time created a repository to share epidemiological data to model infections and risk factors (Schwab & Held 2020, Xu, Kraemer, & the

Open Covid-19 Data Curation Group, 2020), as well as many other groups that acted similarly.

How about long term wins? Automate the exploration and use of open data to accelerate the scientific discovery process, integrating data with advanced automated workflows that interact with researchers and assist with recommendations for new data collection, and implement user-friendly data science tools available to all with access to vast amounts of the world's open data.

### **Mark Hahnel**

Short term wins are already evident. [McGillivray et al](#) found an association between articles that include statements that link to data in a repository and up to 25.36% ( $\pm 1.07\%$ ) higher citation impact on average. This means that if you publish your datasets associated with a paper, you have the potential to get citations to your dataset PLUS on average you get more citations to said paper - compounding citations, if you will.

Long term, it is all about moving research further faster. By removing blockers to access to information, increasing use of data will bring a paradigm shift to the nature of science, known commonly as [the Fourth Paradigm](#).

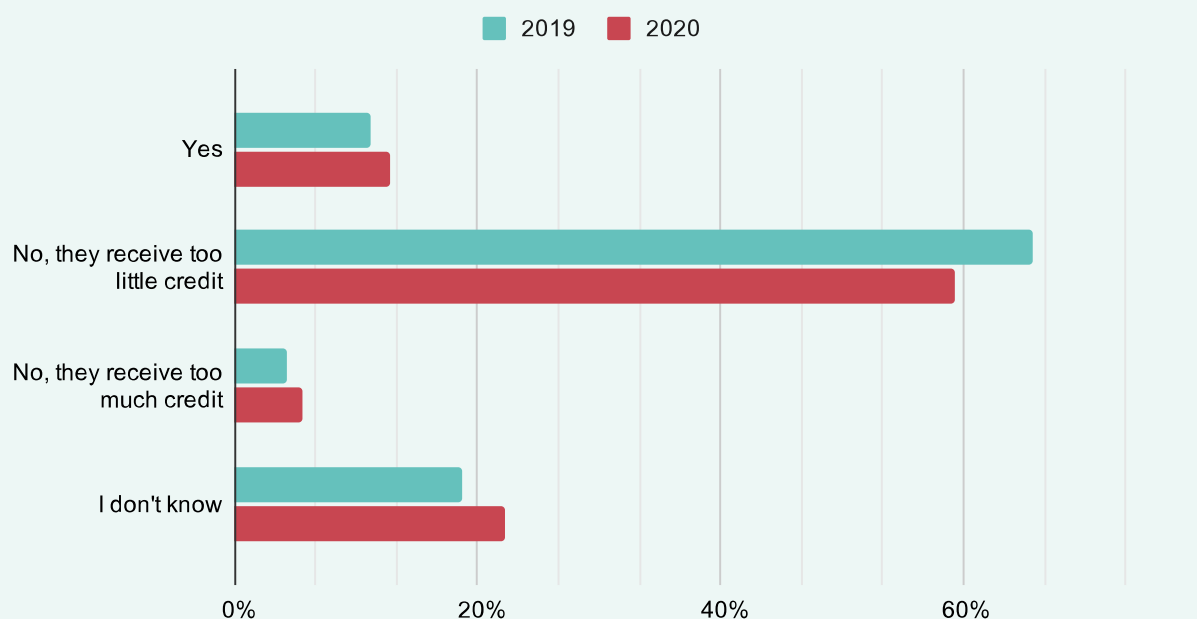
### **Brian Nosek**

The quick wins are continuing the transformation of policies by funders, journals, and institutions to make open the default. The policies are developed and tested, we just need the stakeholders to adopt them. Making the progressive actors more visible and rewarding them will help induce normative pressure on the slower actors to improve their policy frameworks.

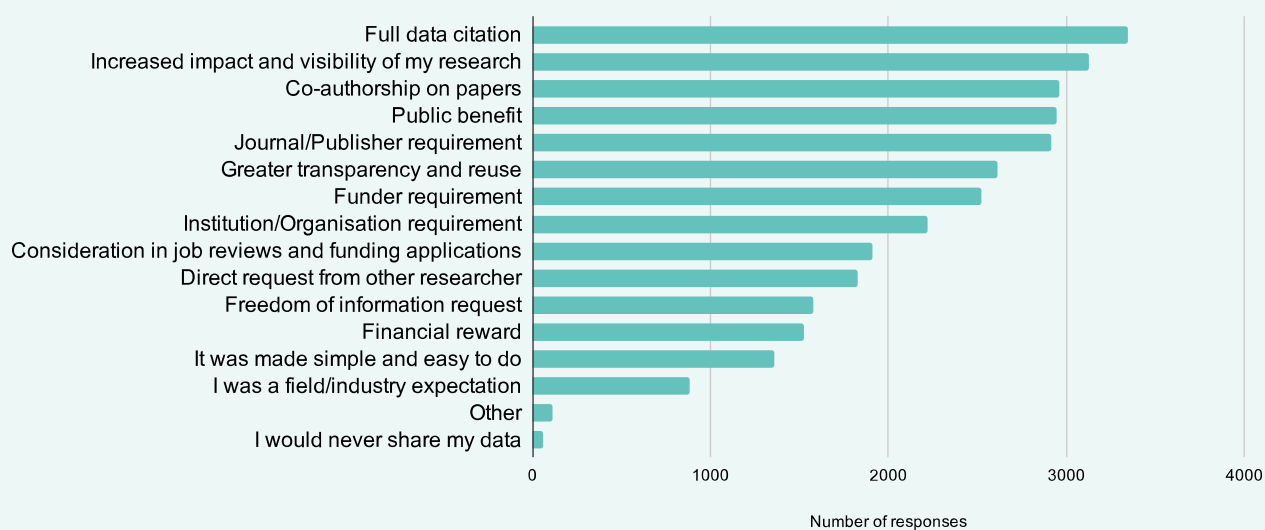
The long-term wins are all about moving the community from "doing the behavior" to "doing the behavior well." The policymakers, infrastructure providers, and metascientists that have been monitoring open data already have a lot of insight about how data sharing falls short of its potential. We need continuous collaboration between researchers evaluating data quality and infrastructure providers creating workflows and solutions. It would be wonderful to set some benchmarks for data quality that the community can challenge itself to reach over the next five years.

"It would be wonderful to set some benchmarks for data quality that the community can challenge itself to reach over the next five years."

## Do you think researchers currently get sufficient credit for sharing data?



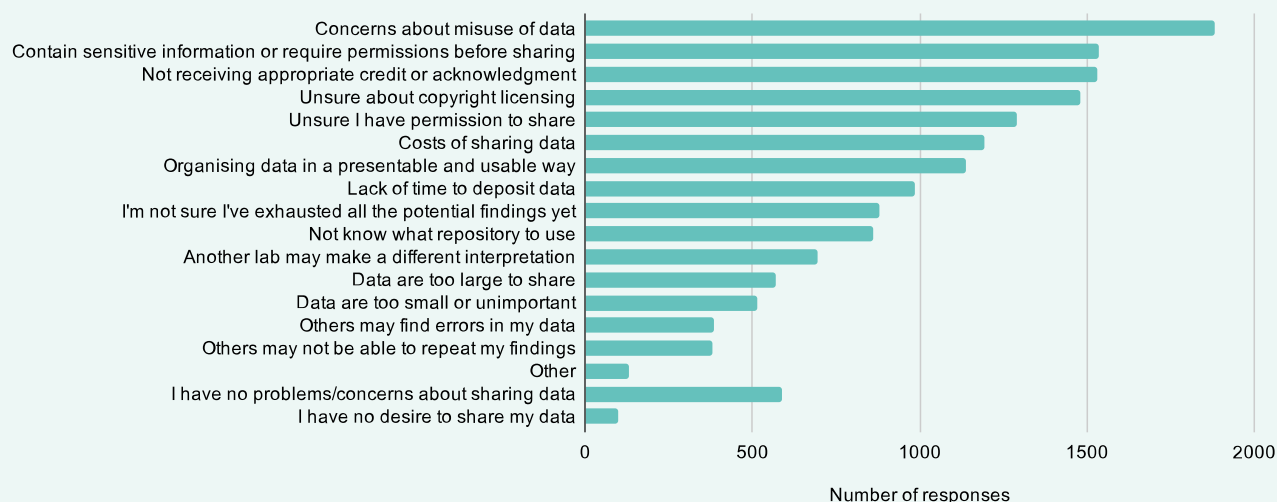
## What circumstances would motivate you to share your data?



<sup>1</sup><https://sfdora.org/read/>



## What problems/concerns, if any, do you have with sharing datasets?



### 5. What credit mechanisms need to be in place for data to be as recognised as paper publications?

#### Kathleen Shearer

True open science requires a paradigm shift, which is, in some ways, in direct conflict with the competitive system of science we have today that ranks researchers according to metrics, such as number of citations, publishing venue, or amounts of funding. Open science, on the other hand, is about encouraging researchers to collaborate, to share with each other and the public, and to be transparent for the greater good of research and society. While rewards for researchers who share their data can help to increase open data practices, it is also very important to acknowledge that simply integrating a narrow range of data sharing metrics into current research assessment systems (such as the number of data sets shared or number of data citations) could actually hinder our overall progress towards open science, because research data will continue to be perceived as a commodity. I agree with the San Francisco Declaration on Research Assessment (DORA) recommendation, “for the purposes of research assessment, (we should) consider the value and impact of all research outputs (including datasets and software) in addition to research publications, and consider a broad range of impact measures including qualitative indicators of research impact, such as influence on policy and practice.”<sup>1</sup> Data sets are becoming easier to cite because many are now accompanied by a DOI, but I think we also need other mechanisms that correlate more directly with what we are trying to achieve through open science: greater collaboration, better quality, and increased impact of research on society.

"The days when only the scholarly article was the entire output of research are long gone, but this is not yet uniformly or institutionally recognised."

"To solve the coordination problem, everyone has to play their part. Be the change!"

#### **Mercè Crosas**

We already have some of the foundations to recognise data publication: most generalist repositories now generate a data citation with a Digital Object Identifier (DOI) associated with each dataset, create standard citation metadata that is usually sent to DataCite, and DataCite works with CrossRef DOIs to connect data with literature (other options are also available besides DataCite DOIs). The *Make Data Count* project is helping standardise data usage and citations metrics across repositories. So what is missing? In general, academic institutions do not yet recognise that data citations (or software citations, for this matter) can be as relevant as article citations; they may not 'count' in such important activities as tenure reviews. Journals do not generally include data citations in their bibliography section. The days when only the scholarly article was the entire output of research are long gone, but this is not yet uniformly or institutionally recognised. We all spend a lot of time collecting and cleaning valuable datasets, generating well-curated information packages, and building software. These outputs are critical parts of the research and are often needed to understand the scientific claims and outcomes. Referencing a PLOS article from 2014, which I had the honor to co-author with a few open data champions, we need to value more the "care and feeding of scientific data".

#### **Mark Hahnel**

Credit mechanisms need to be applied at every level. At the top, organisations such as the Research Excellence Framework (REF), the UK shared funder policy, and comparable measurements in other countries need to treat all scholarly outputs equally. Following on from this, there is no reason why metrics such as the H-Index do not include Datasets – something that could easily be done by Google Scholar and Google Dataset Search. Finally, we need to be focusing more credit for transparent research and good academic technique. Publishers ensure that all papers have ethics statements or waivers, extrapolating out the editorial checklist would further enhance the need for complete papers – including all files needed to reproduce the findings.

#### **Brian Nosek**

Data citation as a regular reference in papers. Funders explicitly asking for evidence of open behaviors in grant proposals and explicitly identifying data, code, materials, and papers as "like kinds" in asking for evidence of impact. Institutions explicitly asking for evidence of open behaviors in job application and promotion materials. Societies developing awards for open behaviors. For example, many have a "best paper" or "high impact paper" award. It would be easy to add similar ones for data, code, and other research content. Again, there is no one solution. To solve the coordination problem, everyone has to play their part. Be the change!

# It All Starts With a Good Plan

Alan Hyndman & Greg Goodey

"Planning is bringing the future into the present so that you can do something about it now."

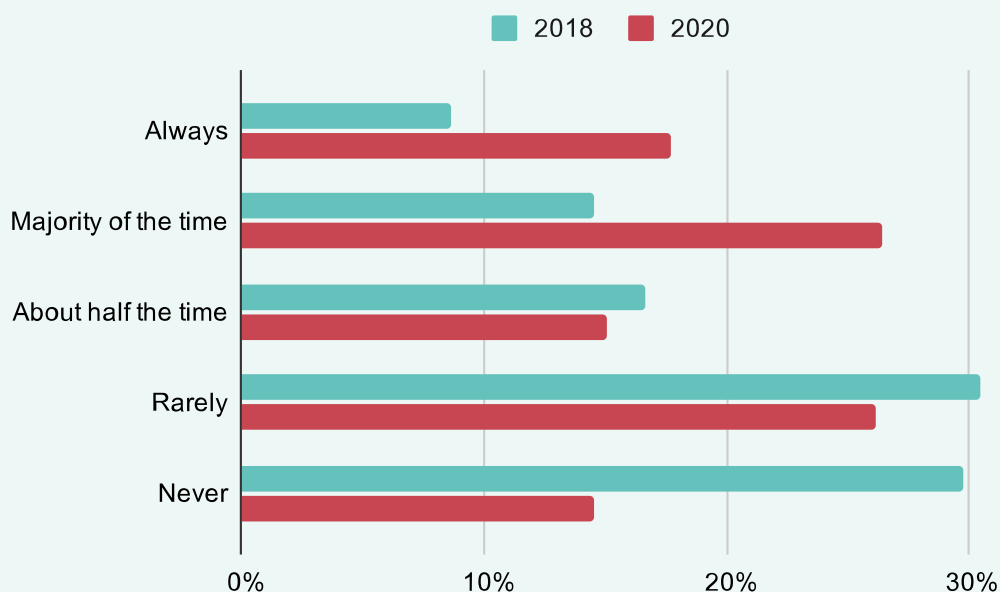
For the 2020 State of Open Data Survey, we decided to redesign some of the sections to better reflect the current issues facing researchers. This resulted in the removal of some questions where we have seen consistent growth or repeat trends year-on-year.

In order to keep the survey as concise as possible, we tried to operate a 'one-question-in, one-question-out' method, removing some questions that were not providing actionable insights and shifting others to being asked biannually. For example, in the first State of Open Data survey in 2016 we were surprised to find that 78% of researchers value a data citation as much as a paper citation; this finding was then validated over consecutive years. Removing some questions allowed us space to ask more pressing questions, such as those on Covid-19.

## Data Management Plans

This year we included an expanded section on data management plans (DMPs). Increasing numbers of funders around the world are mandating researchers to submit a DMP with their grant application, including the most recent draft mandate from the NIH which states that all grants must have a DMP attached by January 2023. As American author Alan Lakein

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



"Data sharing policies are becoming more important to researchers, based on the results of this year's survey."

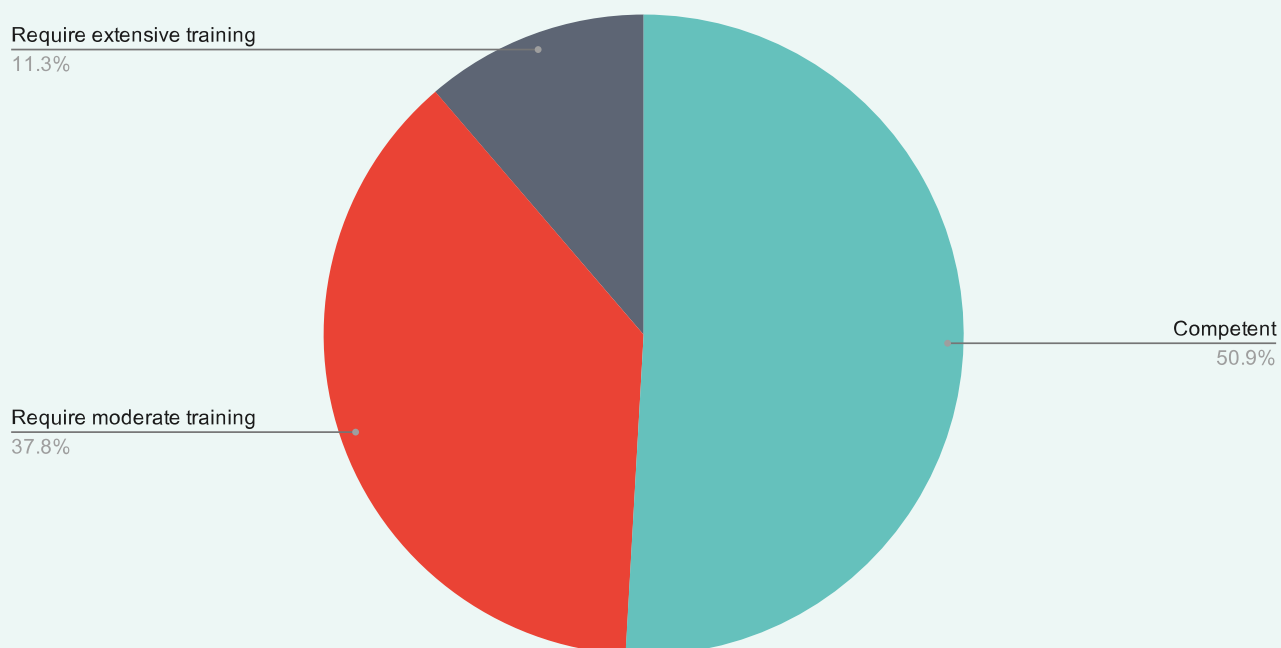
said *"Planning is bringing the future into the present so that you can do something about it now."* A data management plan can play a crucial role in researchers understanding how they will store and publish their data at the outset of a project and encourages best practices rather than it being an afterthought.

We have seen a significant increase in researchers creating DMPs, with the amount of researchers always making a plan doubling from 9% to 18% and the number of researchers never making a plan halving from 30% to 15%.

When looking at regional differences, researchers in Europe were significantly more likely to indicate that they rarely, if ever, create data management plans (51% indicating this) although the UK sits outside of this trend with 51% indicating they frequently, if not always, create a data management plan. This is due to more UK funders mandating DMPs and likely the associated support, provided by organisations like the [Digital Curation Centre with the DMP Online tool](#).

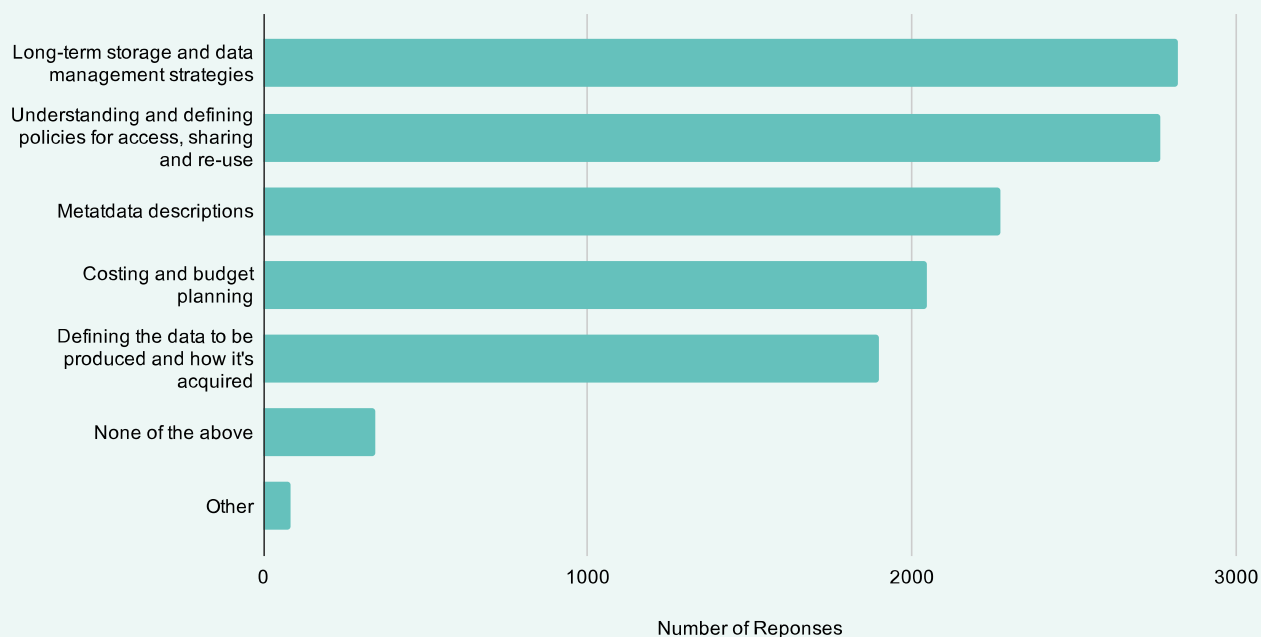
The proportion of respondents who indicated that they felt sufficiently competent to develop a practical data management plan (51%) and that that felt they would require further training (49%) was relatively evenly split.

### To what extent do you feel capable of developing a practical data management plan if you were required to do so?



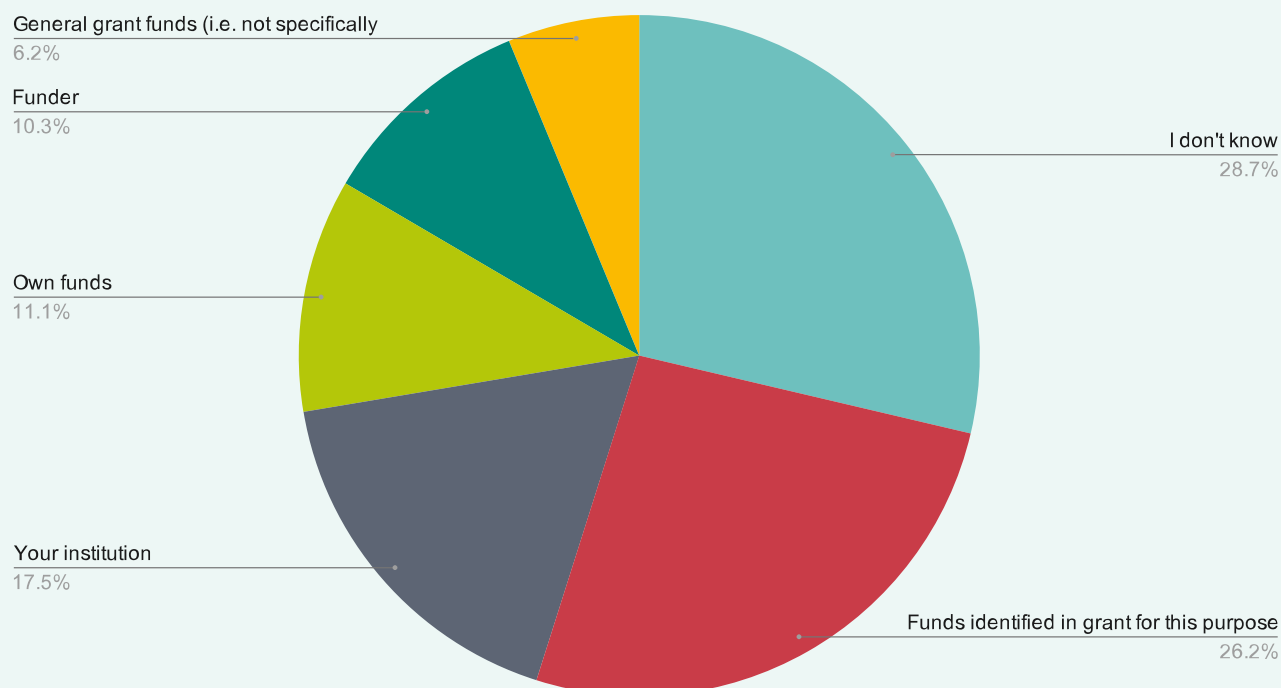
Two aspects of data management planning that the majority of respondents felt that they would benefit from further skills training in were: 1) long-term storage and data management strategies (57%) and 2) understanding and defining policies for access, sharing and reuse (56%). Costing and budget planning was of particular interest to early career researchers (48%).

In which, if any, aspects of data management planning do you feel you would benefit from further skills training?



Yet there is still a gap in researchers' understanding of how their open data efforts will be financed with a majority of researchers (29%) answering "I don't know" to the question "Do you know who would meet the costs of making your research data open access?"

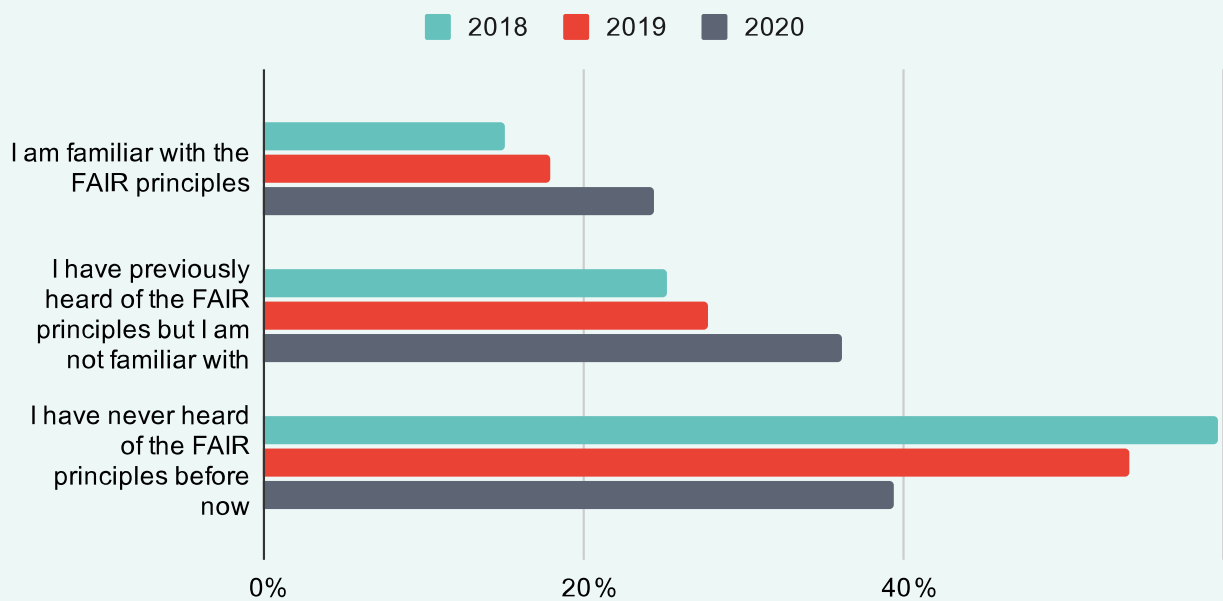
'Do you know who would meet the costs of making your research data open access (e.g. resource for curation)?



## Open Data Initiatives

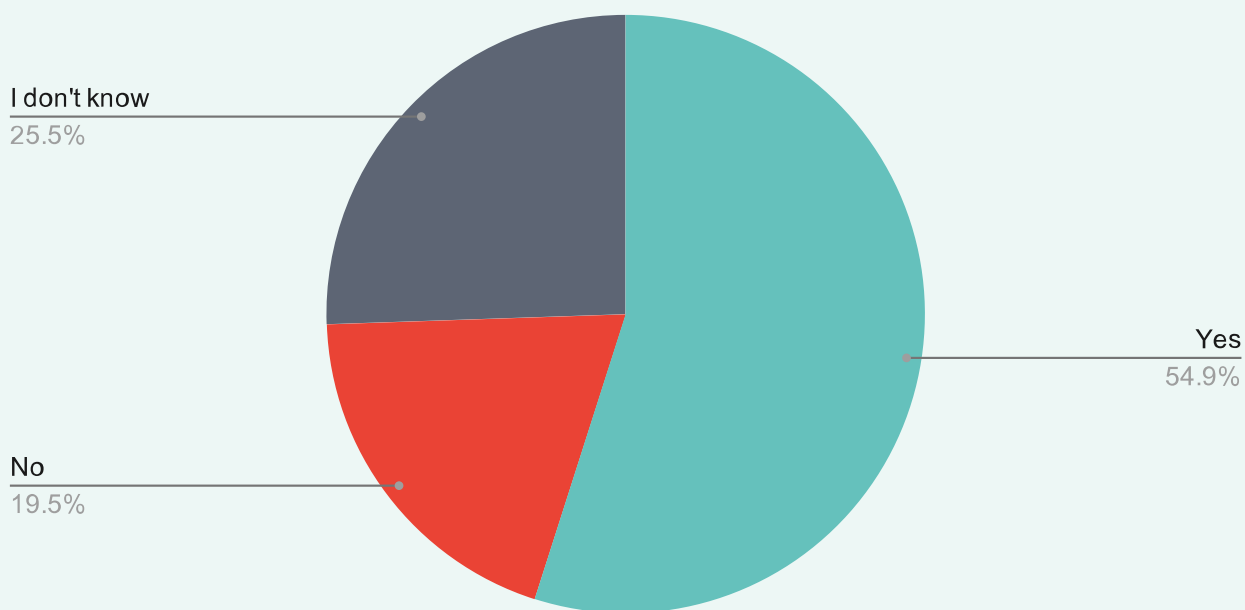
The FAIR principles are the foundation of good data management and it's encouraging to see the understanding and adoption of them becoming more widespread. When we first asked the question in 2018, 60% of respondents had never heard of the FAIR principles; this year that number was down to 39% and the overall familiarity has increased from 15% to 24%.

### How familiar are you with the FAIR principles in relation to open data?



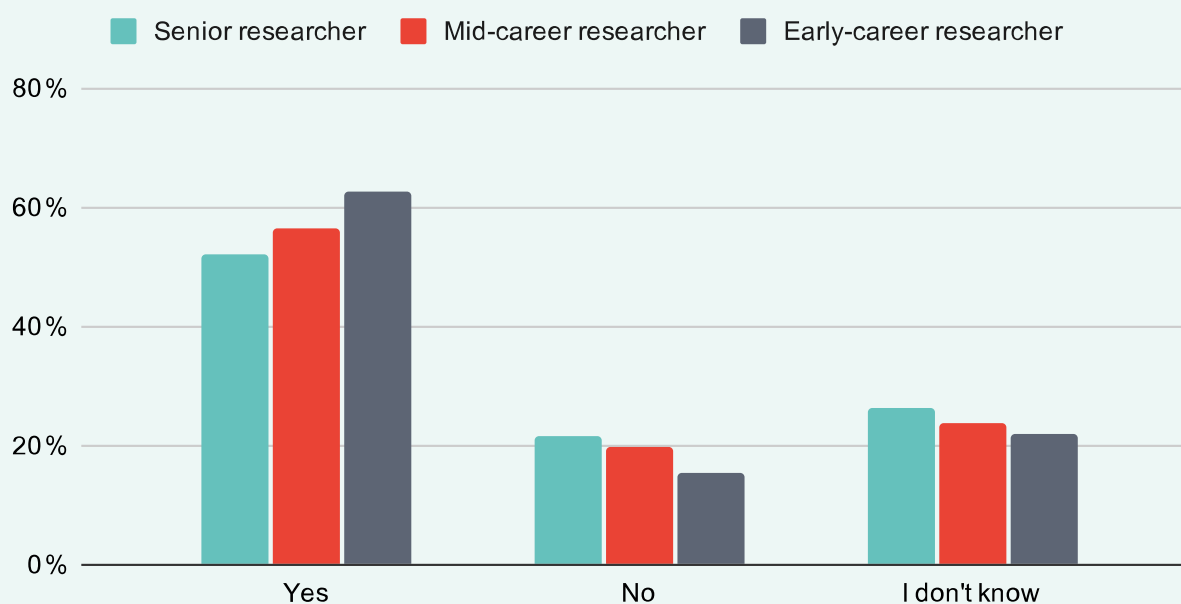
When asked whether researchers felt that sharing data should be a part of the requirements for awarding grants 55% agreed. This proportion was significantly higher for early-career researchers (62%).

## 'Should funders make the sharing of research data part of their requirements for awarding grants?



Over half of respondents indicated that they would be supportive to some extent of a national mandate for making research data openly available. Early-career researchers are more likely to strongly support this than their senior counterparts.

## Should funders make the sharing of research data part of their requirements for awarding grants? by Seniority



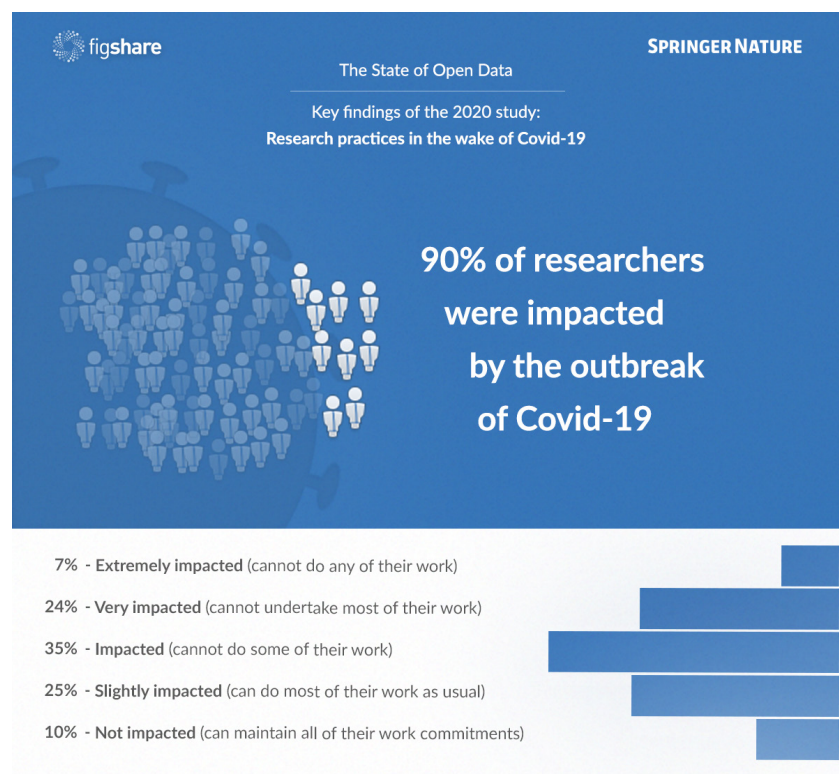
# Research Practices in the Wake of Covid-19

By Grace Baynes and Mark Hahnel

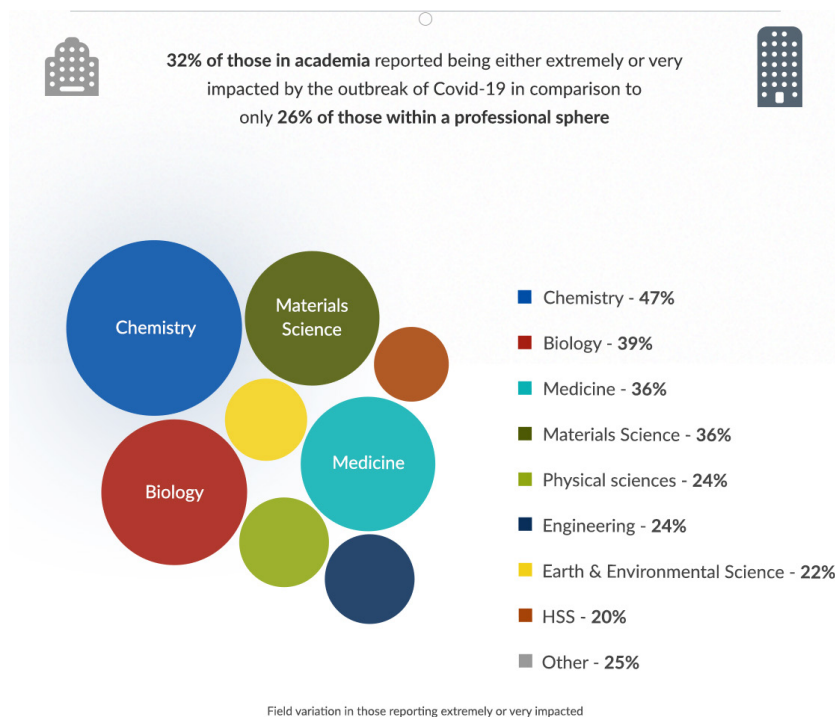
*First published on 7th August 2020 on the Springer Nature and Figshare blogs.*

The current climate has put a spotlight onto the value and importance of data sharing, curation and good data management for boosting the reproducibility and reliability of research. Its value has never been pulled more sharply into focus as you can see the real life impact of data sharing as we navigate this pandemic. After five years of collaborating on this annual survey, we can see increasingly positive attitudes and behaviours when it comes to data sharing. Yet, many researchers and those within the research community still face roadblocks – be this because of challenges in working practices, the lack of tools or services supporting them, or the wider misconception around the role, use and appropriate re-use of data – and this is a problem.

Since 2016 Figshare, Springer Nature and Digital Science have partnered on the State of Open Data report, based on a survey tracking researcher attitudes and behaviours towards open data sharing and research data management. The most recent survey launched in May this year, and with the global pandemic we took the opportunity to ask researchers how Covid-19 was impacting their ability to carry out research, and their views on reuse of data and collaboration. We wanted to get a better understanding of how researcher behaviour was being affected. When the



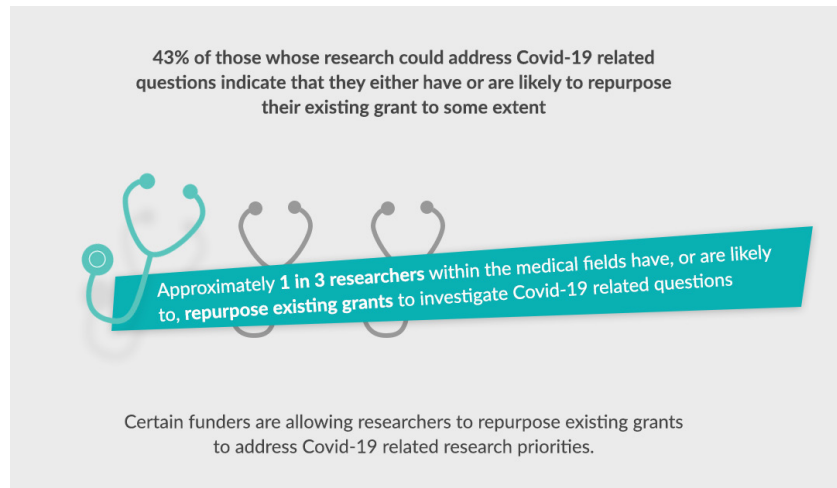




survey was conducted much of the world was under lockdown. We were aware of the time sensitivity of these insights so we released a snapshot of the data to the community as soon as we could, to allow stakeholders the time to analyse the data to help inform policy and actions going forward as we entered a new phase of the pandemic. The data below was from surveys completed between 24th May to 18th June, n=3,436.

### As a snapshot of the full report, key takeaways indicated that:

- Over a third (32%) of academic researchers reported that their research had been 'extremely' or 'very' impacted by the outbreak of Covid-19. This is higher than those working in professional settings (26%).
- The disciplines affected most by Covid-19 were those working in Chemistry (47%), Biology (39%), Medicine (36%) and Materials Science (36%). The lowest level of impact was reported in Humanities and Social Sciences (20%).
- 43% of those surveyed have already or are likely to repurpose their grant to some extent for Covid-19 research.
- Lockdown is seen by half of respondents as 'extremely' or 'somewhat' likely to result in re-use of open data provided by other labs, and 65% expect to reuse their own data.
- More than a third of researchers say they expect to see more collaboration as a result of Covid-19; for those in countries like Brazil and India where the impact of Covid-19 on research appears significant, around half expect collaboration to increase as a result.
- Those researching in Medicine and/or working in a clinical setting were more likely to state they expect to see collaborations increasing as a result of Covid-19, compared to the wider sample.



Covid-19 has demonstrated that the research community has the ability to react to a crisis, and quickly. We have seen an increase in the publication of preprints, expedited peer review and clinical trials, an increase in collaboration and data sharing, as well as funders allowing the diversion of funds to Covid-19 research. All of this together has demonstrated the incredibly responsive nature of our sector, under immense pressure, at a time when the use, re-use, access to and engagement with research has, and continues to be critical. In turn the practices and outreach conducted during this time have led us to a greater understanding of the disease which will hopefully result in better therapeutics and a successful vaccine.

Lockdown has also notably resulted in greater intended re-use of data with over 60% of respondents likely to reuse their own data during lockdown (64%), and a similar percentage over the next 12-18 months (65%). This compares to 58% who report previously reusing their own data. We see similar levels of increase in expected reuse in others' data - 50% during a state of lockdown and 51% over the next 12-18 months. Forty-four percent (44%) of respondents report that they have previously reused others data. The inability of many researchers to gain access to their labs or carry out new research has fuelled a planned increase in the reuse of their own and others data.

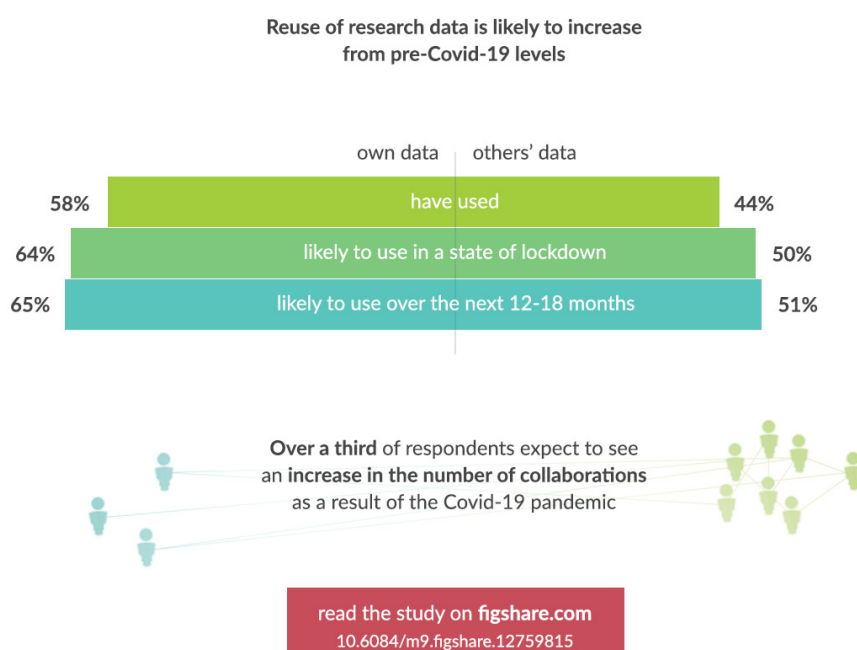
Early responses to our findings have highlighted concerns that the re-use of data (the same data for new publications) could fuel academic misconduct, while others believe academics have made the most of this time to analyse data they had not got around to yet. These concerns underscores the importance, and value, of releasing all data related to a publication so that its provenance is clear and the data can be scrutinised by reviewers and readers. If data underpinning a publication is published alongside the article there is less chance of researchers' salami slicing or writing contradictory papers from the same dataset. Published, citable datasets, are, as they should, being recognised as an equal research output to the paper.

What we have also seen emerge from the survey results, is a much greater focus on collaboration - collaboration across researchers, better collaboration to support the sustainable use of data, and a greater awareness from funders, research organisations and publishers in how to

enable sustainable re-use of data and the structures needed in which to do so. Open data, although faced with challenges, is an integral part of being able to advance the conversations and collaboration around research engagement. Appropriate re-use of data where resources are limited enables the vital research that is needed, pandemic or not, to continue to develop and maximise the return on investment in the original study. With the survey results indicating a heightened awareness into the role and needs of open data, we can take this as a really positive sign that attitudes, and practices, are changing and collaboration, across stakeholders, is taking place which in turn will enable real effective and sustainable change for the use of open data and the benefits of open research.

What we are seeing from this snapshot of data is the unwavering value of having access to data and the importance of rapid data sharing. Good and appropriate data management has, and will continue to enable researchers to reuse their own data where they are not able to conduct new experiments – which in an environment where many are still unable to be back in active research settings, is vital in enabling research, and collaboration, to continue to take place.

Whilst there is arguably still hesitation around ways in which data can be reused and shared appropriately and sustainably, throwing the spotlight on open data, its practices and its value in such a pandemic is an important conversation and one that is needed in order to continue to effect positive change through collaboration, awareness and innovation. We all have a role to play in this – supporting uptake through policy and credit, the better management of open research and data, and the development of tools and services to enable high quality research to be conducted, collaborated on and shared both through times of crisis and as we move back to the ‘new normal’ – whatever that may look like for the wider research community.



# Contributor Biographies

**Kathleen Shearer** has been the Executive Director of the Confederation of Open Access Repositories (COAR), an international association that brings together individual repositories and repository networks in order to build capacity, align policies and practices, and act as a global voice for the repository community, since 2013. She is based in Montreal, Canada and has been working in the area of open access, open science, scholarly communications, and research data management for over 15 years. She participates in numerous other organizations and activities around the world advancing open science. She is the author of numerous publications and has delivered many presentations at international events. Most recently, she was the lead author of the paper *Fostering Bibliodiversity in Scholarly Communications: A Call for Action*. Shearer is also a Research Associate with the Canadian Association of Research Libraries and has been instrumental in many of CARL's activities related to open science, including the launch of the Portage Initiative in Canada, a national research data management network.

 <https://orcid.org/0000-0001-8617-5781>

**Mercè Crosas** is the University Research Data Management Officer, with Harvard University Information Technology (HUIT), and Chief Data Science and Technology Officer at Harvard's Institute for Quantitative Social Science (IQSS). In her role at HUIT, Dr. Crosas provides leadership to mature Harvard's research data management and governance practices. She works in close collaboration with key constituencies in Research, Information Technology, and the Library to coordinate support for the data lifecycle and guide university policy, process, and procedures for research data. Dr. Crosas brings to this role a wealth of experience in data management architecture and international community data standards as well as the vision to make data more accessible for research while preserving privacy. She co-leads the Harvard Data Commons.

 <https://orcid.org/0000-0003-1304-1939>

**Brian Nosek** is co-Founder and Executive Director of the Center for Open Science (<http://cos.io/>) that operates the OSF (<http://osf.io/>) – a collaborative management service for registering studies and archiving and sharing research materials and data. COS is enabling open and reproducible research practices worldwide. Brian is also a Professor in the Department of Psychology at the University of Virginia. He received his Ph.D. from Yale University in 2002. He co-founded Project Implicit (<http://projectimplicit.net/>), a multi-university collaboration for research and education investigating implicit cognition--thoughts and feelings that occur outside of awareness or control. Brian investigates the gap between values and practices, such as when behavior is influenced by factors other than one's intentions and goals. Research applications of this interest include implicit bias, decision-making, attitudes, ideology, morality, innovation, barriers to change, open science, and reproducibility. In 2015, he was named one of Nature's 10 and to the Chronicle for Higher Education Influence list.

 <https://orcid.org/0000-0001-6797-5476>

**Mark Hahnel** is the CEO and founder of Figshare, which he created whilst completing his PhD in stem cell biology at Imperial College London. Figshare currently provides research data infrastructure for institutions, publishers and funders globally. He is passionate about open science and the potential it has to revolutionize the research community. For the last eight years, Mark has been leading the development of research data infrastructure, with the core aim of reusable and interoperable academic data. Mark sits on the board of DataCite and the advisory board for Directory of Open Access Journals (DOAJ). He was on the judging panel for the National Institutes of Health (NIH), Wellcome Trust Open Science prize and acted as an advisor for the Springer Nature master classes.

 <https://orcid.org/0000-0003-4741-0309>


**Mariëtte van Selm** is a historian, holding a PhD in Theology, who worked at several research institutes in the Netherlands before starting at the Library of the University of Amsterdam (UvA) and Amsterdam University of Applied Sciences (AUAS) ten years ago. She developed and now coordinates research data support at the Library, is manager of the UvA/AUAS Research Data Management Programme and is a member of the Advisory Board of the National Coordination Point Research Data Management (LCRDM) in the Netherlands. As of this year, she's also a member of the ResearchIT team at UvA/AUAS' IT department. She holds the questionable honour of being the first UvA employee to register an ORCID ID.

 <https://orcid.org/0000-0003-3711-4282>

**Dr. Leslie McIntosh**, PhD is the founder and CEO of Ripeta, a company formed to improve scientific research quality and reproducibility. Now part of Digital Science, the company leads efforts in rapidly assessing scientific research to make better science easier. She served as the inaugural executive director for the US region of the Research Data Alliance and is still very active with the RDA. She has experience leading diverse teams to develop and deliver meaningful data to improve scientific decisions. Dr. McIntosh is an accomplished biomedical informatician and data scientist as well as an internationally known consultant, speaker, and trainer who is passionate about mentoring the next generation of data scientists. She holds a Masters and PhD in Public Health with concentrations in Biostatistics and Epidemiology from Saint Louis University and a Certificate in Women's Leadership Forum from Washington University Olin's School of Business

 <https://orcid.org/0000-0002-3507-7468>

**Grace Baynes** is VP of Research Data and New Product Development at Springer Nature. She is responsible for Springer Nature's approach to research data, including advocacy for open data and good data practice; journal data policies; and data publishing including the journal Scientific Data. Her new product development responsibilities are currently focused on developing research data services and solutions for researchers, institutions and funding organizations, and establishing the new product development approach for researcher services. Grace has spent twenty years in publishing, sixteen of those working in open research, joining open access publisher BMC in 2003, and since then in roles at Nature Publishing Group and now Springer Nature.

 <http://orcid.org/0000-0002-4933-3186>

**Gregory Goodey** is a Data Analyst at Springer Nature. He is responsible for managing projects to collect and analyse information on customers, markets, products and communications within the STM market. He plays a major role in coordinating the annual State of Open Data survey where he manages development of the survey and the analysis of the outcomes. Greg completed his Ph.D. in Physiology at UCL and has since held research roles in a number of industries joining Springer Nature in 2016.

 <https://orcid.org/0000-0002-1541-6805>

**Alan Hyndman** is the Marketing Director at Figshare and has worked on the survey design and data analysis of the State of Open Data since its inception in 2016. He has an extensive background in technology marketing and has been responsible for the Figshare brand since 2012. He is passionate about open data and the potential it has to improve academic research, science communication and society at large.

 <https://orcid.org/0000-0002-1523-1499>

Part of **DIGITAL**science

