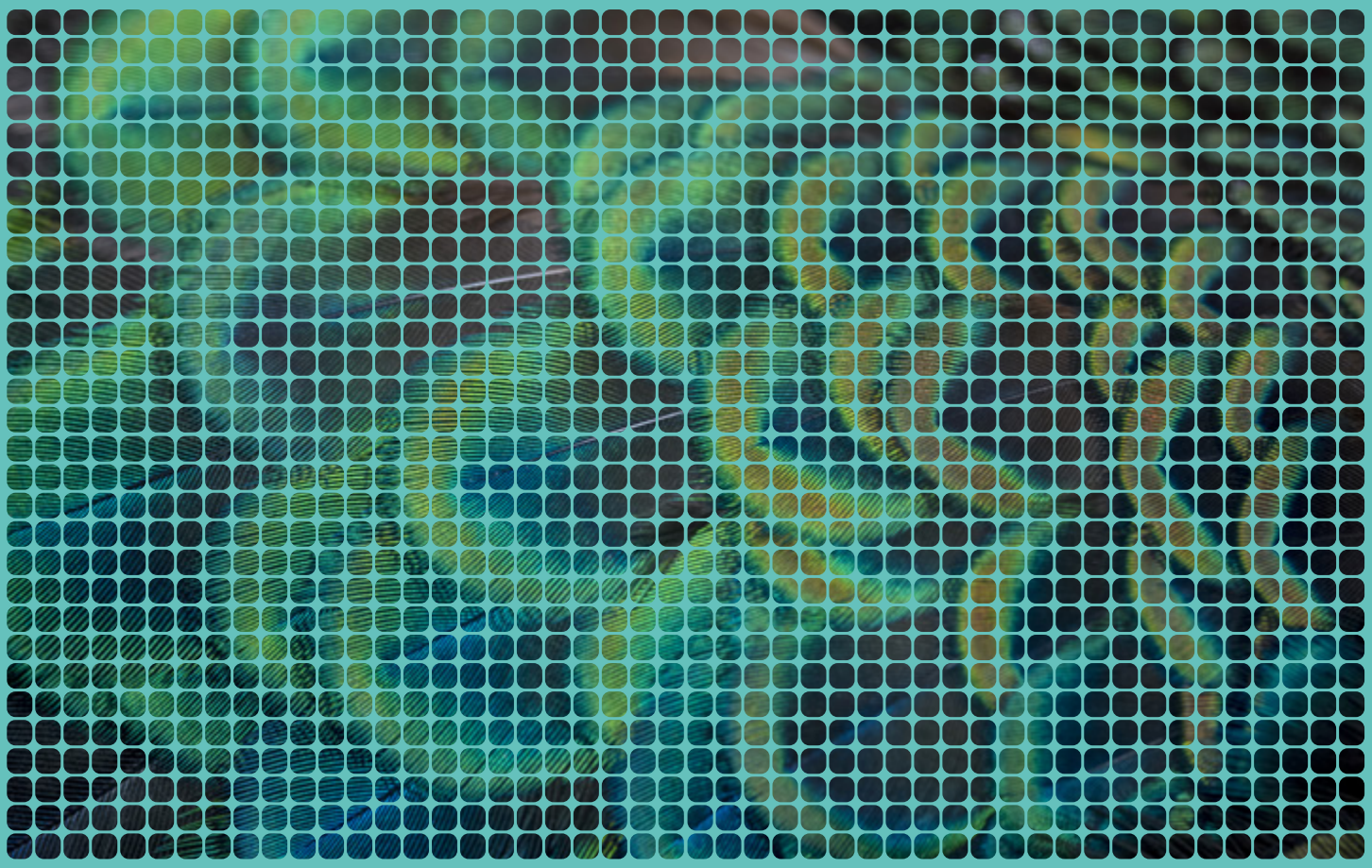


Digital Science Report

The State of Open Data 2019

A selection of analyses and articles about open data, curated by Figshare

Foreword by Dr Paul Ayris

OCTOBER 2019

DIGITAL science

figshare

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

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, Anywhere Access, BioRAFT, CC Technology, Dimensions, Figshare, Gigantum, GRID, IFI Claims, Labguru, Overleaf, ReadCube, Ripeta, Symplectic, TetraScience, Transcriptic and Writefull. We believe that together, we can help researchers make a difference.
Visit www.digital-science.com

Acknowledgements

Figshare and **Digital Science** are extremely grateful to **Springer Nature**, who have been our partners in scoping this year's survey, and provided survey design, hosting and global distribution.

Figshare and **Digital Science** are also extremely grateful to the contributors for their thought leadership pieces included in the 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
Figshare, 6 Briset Street, London, EC1M 5NR, UK. info@figshare.com

DOI: https://doi.org/10.6084/m9.figshare.9980783
ISBN: 978-1-9993177-2-0

SPRINGER NATURE

Contents

"Research data is the new currency in the research landscape"

Foreword

Dr Paul Ayris, FRHistS, Pro-Vice-Provost,
University College London Library Services

I am honoured to introduce this year's *State of Open Data* report. Research data is the new currency in the research landscape. This data, the building blocks on which publications are based, can now be made available for sharing and re-use as open data alongside the publication which references it. As the LERU Open Science Roadmap[1] makes clear, embracing open science requires a culture change in the way research is undertaken, shared, published, evaluated, rewarded and curated. This change in the production and dissemination of research outputs represents a fundamental movement in the research landscape and the *State of Open Data* report is an important milestone in measuring progress along this road.

It has been a really important year for research data management in my own institution, University College London (UCL) in the UK. In June this year, we launched our Research Data Repository (RDR)[2] using Figshare as the underlying infrastructure following a competitive tender. UCL already had repositories for personal and sensitive data, and a storage service for data produced in the course of funded project work. What was needed was a repository for the long-term curation of data, and this is what RDR provides. The motivation was partly compliance with funder requirements, however UCL was keen to ensure that, where possible, research data should be as open as possible as this is good research practice. RDR is also the repository UCL hopes can interact with the European Open Science Cloud (EOSC).[3] The EOSC has been a slow starter in terms of developing rules of engagement for universities to adopt. That needs to change if Europe is to develop its position as a world leader in research data management.

Fundamental to good research data management is the concept of FAIR (Findable, Accessible, Interoperable, and Reusable) data.[4] It is a challenge to introduce FAIR principles at individual researcher level in universities. There is a need for co-ordinated skills development to train researchers in what is needed to deliver FAIR data and, indeed, in adopting open data as the norm. Is there a need for a new profession of data curators who can take on this role for research groups? This was the recommendation of the first High Level Expert Group on the EOSC, of which I was privileged to be a member. Their report [5] stressed the need for hundreds of thousands of data experts to be trained by 2020, and for each Member State to have at least one certified institute to support the introduction of data management across disciplines. 2020 will soon be upon us, but this ambitious goal has not yet been reached. The costs of such developments and the culture change needed to embed such practice at university level mean that it will not be delivered quickly.

[1] LERU: *https://www.leru.org/publications/open-science-and-its-role-in-universities-a-roadmap-for-cultural-change*; last accessed 10 September 2019.

[2] UCL: *https://rdr.ucl.ac.uk/*; last accessed 10 September 2019.

[3] EOSC: *https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud*; last accessed 10 September 2019.

[4] FAIR: *https://www.go-fair.org/fair-principles/*; last accessed 10 September 2019. To turn FAIR data into reality, see *https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf*; last accessed 10 September 2019.

[5] EOSC: *https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf*, p. 16; last accessed 10 September 2019.

Nonetheless, the opportunities presented by open data are enormous. The accumulated cost savings for the Member States in 2020 are forecast to equal 1.7 billion euros.[6] The study which has produced this figure, *Creating Value through Open Data*, also looked at a number of case studies and found, for example, that applying open data in traffic can save 629 million hours of unnecessary waiting time on the road in the EU. Open data also has the potential of saving 1,425 lives a year (i.e. 5.5% of the European road fatalities).[7]

One of the key requirements of the change of culture needed to deliver open and FAIR data is a change in the university reward and incentive system. Current practice is focused on publications and, in many cases, the impact factor of the journals in which articles are published. There is little room for research data in this model. Professor Bernard Rentier and a Working Group of the European Commission have recently presented a report entitled *Evaluation of Research Careers fully acknowledging Open Science Practices*.[8] This report identifies 23 rounded criteria for reward, of which datasets is one. No university in Europe has yet introduced this complete matrix, but UCL has already modified its academic promotions framework to acknowledge openness as a criterion for reward.[9]

Open data is a key component of open science, but cultural change needs to happen for open science to become the norm in research practice. The research community has started this journey and, with regular reports on The State of Open Data, it is possible to measure the pace of this fundamental transition.

"The State of Open Data report is an important milestone in measuring progress along this road."

[6] European Data Portal: *https://www.europeandataportal.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf*, p. 11; last accessed 10 September 2019.

[7] *Ibid.*, p. 12.

[8] European Commission: *https://ec.europa.eu/research/openscience/pdf/os_rewards_wgreport_final.pdf*; last accessed 10 September 2019.

[9] UCL: *https://www.ucl.ac.uk/human-resources/sites/human-resources/files/ucl-130418.pdf*; last accessed 10 September 2019.

"This year's Nobel Laureates have made their data openly available"

Power Struggles and Rewards for Academic Data

Mark Hahnel, CEO and Founder, Figshare

Some of the biggest academic headlines of the last 20 years have one thing in common - the data is always open. From the Human Genome Project, to gravitational wave detection; even this year's Nobel Laureates have made their data openly available. Why is this not the case for all research? Is there a rule that determines which research has a moral obligation to make the data available? One that covers both impact and subject?

"How do we move to a culture of appropriate impact for datasets?"

When we look at those discoveries, we see that the biggest impact for researchers is in the papers they write. For the Human Genome Project, there has undoubtedly been more work done on top of the actual data than on top of the paper. The paper states how the research was carried out. The data lives as a separate entity, that is not cited as the paper is. This is largely down to the culture of gaining credibility as a researcher by publication, the core currency of research. There is no doubt that the 256 authors have advanced their careers on their Nature publication and subsequent >15,000 citations it has generated, however the datasets probably don't feature on their CVs. For the human genome researchers, it may be because data publishing practices and infrastructure were in their infancy when their research was first published. So how do we move to a culture of appropriate impact for datasets, now that the infrastructure is available for all?

"It is somewhat surprising to find that researchers still see data as supplemental to what is seen as the core research output - the paper"

In a year that has so far seen the US Government sign the Open, Public, Electronic and Necessary (OPEN) Government Data Act and an explosion in links to datasets as separate independent research outputs (as a quick Dimensions search for datasets published in Figshare, Zenodo and Dryad demonstrates), it is somewhat surprising to find that researchers still see data as supplemental to what is seen as the core research output - the paper. In a world of fake news, interpretation of information can be seen as a weak point in factfullness. A researcher has many motivating factors to ensure that the interpretation of their data is as impactful as possible. This highlights how essential it is that the data - the real facts - are made openly available for others to reproduce the interpretations. It also suggests that the re-use measurement should be prioritized at the dataset level, and not the paper.

The UK Research Excellence Framework (REF) recognizes research to be more than research outputs, defining research as *"work of direct relevance to the needs of commerce, industry, culture, society, and to the public and voluntary sectors; scholarship; the invention and generation of ideas, images, performances, artefacts including design, where these lead to new or substantially improved insights; and the use of existing knowledge in experimental development to produce new or substantially improved materials, devices, products and processes, including design and construction"*.[1]

[1] *https://www.ref.ac.uk/guidance/*

The International Committee of Medical Journal Editors recommends that authorship be based on the following four criteria:

- Substantial contributions to the conception or design of the work, or the acquisition, analysis, or interpretation of data for the work, AND;
- Drafting the work or revising it critically for important intellectual content, AND;
- Final approval of the version to be published, AND;
- Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

"It seems that researchers will hold taxpayer funded research data to ransom"

However, in this year's State of Open Data survey, we see that co-authorship on a paper in exchange for sharing data use is not only on the wishlist, but also something that people already do. It seems that researchers will hold taxpayer funded research data to ransom. Of those who have shared their data with their peers, 35.1% reported to have received co-authorship on a paper in exchange.

We asked the question, 'If the reuse of your data in a subsequent paper resulted in you being credited as co-author, how much would this motivate you to make your data openly available to others?' Only 8% said it would not affect their decision making. Therefore it seems that there is still an accepted notion in academia that data can be held to ransom. In the massively competitive tenure process, data is power.

"The academic hierarchy cannot rely on Journal Impact Factor alone. We need broad measures of impact for broad measures of outputs"

We are under no illusion that the paper will remain the core published output for academics going forward, though the push for a more diverse set of metrics than the journal impact factor (JIF) should also come with rewarding a more diverse set of research outputs. We know about many researchers who have succeeded in gaining huge impact for their non-traditional research outputs, from the Statistic of the Year to heavily cited software packages. So who is responsible for driving this message to researchers? Or who should be held accountable to fix it? It is this author's opinion that the responsibility now lies with grant funders and institutional hiring committees. The academic hierarchy cannot rely on Journal Impact Factor alone. We need broad measures of impact for broad measures of outputs.

One encouraging final thought is that co-authorship is not the biggest rewarding factor for academics sharing their research data. For the fourth year in a row (ie. since the State of Open Data survey and report was initiated), citations are seen as the holy grail in terms of reward. Once again, we see that there are more citations to datasets by Open Access Week in 2019 than there were in 2018.[2] The rewards are growing, the incentives are growing, and the mandates are growing. It is an exciting time to start investigating what re-use of data looks like, and investigating the hypothesis that better described data leads to more re-use, which in turn leads to more rewards for academics, and ultimately more efficient research for humanity. Who does that curation is an open question, and one that I hope we will have a much clearer answer to by Open Access week 2020.

"It is an exciting time to start investigating the hypothesis that better described data leads to more re-use, which in turn leads to more rewards for academics, and ultimately more efficient research for humanity"

[2] *https://app.dimensions.ai/discover/publication?search_text%3D10.6084%2520OR%252010.5061%2520OR%252010.5281%26search_type%3Dkws%26full_search%3Dtrue&sa=D&ust=1571131935649000&usg=AFQjCNE6KKqOPBecLh5DhHb1Vh7wBALakg*

"The biggest barrier to research data sharing and reuse seems to be a matter of trust"

Building Trust to Break Down Barriers

Iain Hrynaszkiewicz, Publisher, Open Research, PLOS

The biggest barrier to research data sharing and reuse seems to be a matter of trust, and in particular trust in what others may do with researchers' data if it is made openly available. The 2019 State of Open Data survey revealed that over 2,000 respondents had concerns about misuse of their research data.

Concerns about data misuse represent a multitude of issues; fears that errors could be found in their work, or that the data could be misinterpreted or research participant privacy be compromised. Researchers might also be concerned that their data will be reused for purposes they did not intend, such as commercial exploitation, or for misleading or inappropriate secondary analyses.[1]

The 2019 survey provides insights from one of the largest pools of respondents ever, but this particular barrier - concerns about misuse of data - should not come as a surprise. It is supported by several previous surveys which, when combined, total many thousands of researcher opinions.[1,2,3,4] With such compelling evidence that these researcher concerns exist, there is an ever-more important need to explore why they persist.

Practical problems, including where, how and when to share research data, have received much attention.[5,6] However, concerns about misuse, alongside the potential to lose publication opportunities ("fear of being scooped"), or going against the more conservative data sharing culture in their own field of research often feature amongst researchers' top concerns.

Features of technological solutions, such as the ability to share data privately in repositories before publication, and policy-level support for researchers' reasonable first use of their data[8] go some way to addressing this issue of trust. However, trust is more a matter of culture than technology. With repositories being used by around a quarter of researchers[9,10] investing in people rather than infrastructure may be a more pressing issue to change research culture, as Dr Marta Teperek, who coordinates one of the largest institutional data stewardship programmes at TU Delft in the Netherlands, has concluded.[11] Investing in skills and training for individual researchers is one possible solution, although Professor Barend Mons has argued that data stewardship, like computer programming, is too specialist a task to expect all researchers to undertake. He recommends that universities provide one data steward for every 20 researchers.[12] Like most difficult problems, there is rarely a single solution. Experimentation and collaboration are essential if we are to enable researchers to build greater trust in the power of data reuse to advance science - in the same way that the unbridled reuse of content helped grow the world wide web so rapidly.

[1] Wiley Open Science Researcher Survey 2016 [Internet]. [cited 15 Nov 2018]. Available: *https://figshare.com/articles/Wiley_Open_Science_Researcher_Survey_2016/4748332/2*

[2] Science D, Hahnel M, Fane B, Treadway J, Baynes G, Wilkinson R, et al. The State of Open Data Report 2018. 2018

[3] Allagnat L, Allin K, Baynes G, Hrynaszkiewicz I, Lucraft M. Challenges and Opportunities for Data Sharing in Japan. Figshare. 2019; doi:10.6084/m9.figshare.7999451.v1

[4] Lucraft M, Allin K, Baynes G, Sakellaropoulou R. Challenges and Opportunities for Data Sharing in China. Figshare. 2019; doi:10.6084/m9.figshare.7326605.v1

[5] Lucraft M, Baynes G, Allin K, Hrynaszkiewicz I, Khodiyar V. Five Essential Factors for Data Sharing. Figshare. 2019; doi:10.6084/m9.figshare.7807949.v2

[6] Stuart D, Baynes G, Hrynaszkiewicz I, Allin K, Penny D, Lucraft M, et al. Whitepaper: Practical challenges for researchers in data sharing [Internet]. 2018. Available: *https://figshare.com/articles/Whitepaper_Practical_challenges_for_researchers_in_data_sharing/5975011*

[7] Houtkoop BL, Chambers C, Macleod M, Bishop DVM, Nichols TE, Wagenmakers E-J. Data sharing in psychology: A survey on barriers and preconditions. Advances in Methods and Practices in Psychological Science. 2018;1: 251524591775188. doi:10.1177/2515245917751886

[8] Concordat on Open Research Data [Internet]. 2016. Available: *http://www.rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf/*

[9] Science D, Hahnel M, Treadway J, Fane B, Kiley R, Peters D, et al. The State of Open Data Report 2017. 2017;

[10] Open Data: the researcher perspective - survey and case studies [Internet]. 4 Apr 2017 [cited 15 Nov 2018]. Available: *https://data.mendeley.com/datasets/bwrnfb4bvh/1*

Scholarly publishers have made good progress in the first step towards changing research culture to support data sharing by raising awareness through the implementation of journal research data policies in the last five years.

At PLOS we have invested significantly in people and processes to support a strong journal data sharing policy since 2014.[13] We are seeing a steady increase year-on-year in the proportion of PLOS authors who use a data repository.[14] Although less costly for publishers, journal policies that only encourage data sharing have much lower levels of compliance.[15]

With so many papers subject to a journal data sharing policy now published it is possible to explore actual data sharing behaviours, and the benefits of sharing, at a large scale. A study of more than half a million PLOS and BMC papers found that researchers who stored their data in a repository were associated with an average 25% increase in citations to their research papers.[16] This is important given that the 2019 survey suggests that a full citation to research papers remains the strongest incentive for researchers to share their data.

While more publishers must invest further in data sharing support for researchers, we should be encouraged by collaborations between publishers and other stakeholders to enable data sharing to be more effective and rewarding. In 2019 the STM Association is collaborating with a Research Data Alliance initiative to implement consistent journal data sharing policies.[17] There are indications from some large publishers that they will focus on strengthening journal policies, signalling the importance of a greater commitment than only encouraging data sharing, or "data available on request". As publishers, we can help to build trust with researchers by being open ourselves: open with our content; our data; our policies; and open with our own data sharing insights.

There are undoubtedly costs associated with implementation of strong policies and solutions for data sharing but, to paraphrase Dr Jean-Claude Burgelman at September 2019's CODATA Beijing conference, open science is ultimately an investment, not a cost.[18] Open science, and indeed open research, is an investment in creating more reliable and reusable knowledge for the future. While data sharing culture and fears over misuse of data persist in today's research environment, with so many organizations invested in finding innovative solutions to the problems that prevent optimal data sharing, we can remain optimistic that these barriers will ultimately be broken down in the future.

"Open science is ultimately an investment, not a cost"

[11] The main obstacles to better research data management and sharing are cultural. But change is in our hands | Impact of Social Sciences [Internet]. [cited 24 Sep 2019]. Available: https://blogs.lse.ac.uk/impactofsocialsciences/2018/11/14/the-main-obstacles-to-better-research-data-management-and-sharing-are-cultural-but-change-is-in-our-hands/

[12] Popkin G. Data sharing and how it can benefit your scientific career. Nature. 2019;569: 445–447. doi:10.1038/d41586-019-01506-x

[13] PLOS' New Data Policy: Public Access to Data | EveryONE: The PLOS ONE blog [Internet]. [cited 28 Mar 2019]. Available: https://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/

[15] Federer LM, Belter CW, Joubert DJ, Livinski A, Lu Y-L, Snyders LN, et al. Data sharing in PLOS ONE: An analysis of Data Availability Statements. PLoS ONE. 2018;13: e0194768. doi:10.1371/journal.pone.0194768

[16] Vines TH, Andrew RL, Bock DG, Franklin MT, Gilbert KJ, Kane NC, et al. Mandated data archiving greatly improves access to research data. FASEB journal : official publication of the Federation of American Societies for Experimental Biology. 2013; fj.12-218164-. Available: *http://www.fasebj.org/content/early/2013/01/07/fj.12-218164*

[17] *Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B. The citation advantage of linking publications to research data. arXiv. 2019; https://arxiv.org/abs/1907.02565*

[18] *Cost-benefit analysis for FAIR research data - Publications Office of the EU [Internet]. [cited 24 Sep 2019]. Available: https://publications.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1*

"Researchers see an important role for funders in ensuring good citizenship among researchers"

What is the State of Open Data in 2019?

Briony Fane, Data Analyst, Digital Science

This is the fourth year that Figshare has run the State of Open Data survey. As in previous years, this has been a collaboration with Digital Science and Springer Nature. In 2016, when we first launched the report, we were pleased to receive more than 2,000 responses – a number which was broadly consistent in 2017 and which dipped slightly in 2018. With the current survey, we were simply staggered by the number of responses that came from across the globe, with more than 8,000 participants in more than 190 countries. Even if we saw no other changes than these higher-level response rates, one thing is clear; open data - including how we use it, produce it, licence it, and otherwise interact with it - is gaining in importance for the research community.

In our first report in 2016, we were struck that more than 78% of respondents valued a citation to a dataset as highly or more highly than they did a citation to a standard research paper. This appeared slightly at odds with the low number of respondents to the question (just 1,715 replies). At the time, this may have been due to lack of engagement with open data, lack of experience with open data, or indeed many other factors. However, it is highly notable that this figure is robust in all our surveys, with 2017, 2018 and this year's data agreeing well with the 2016 number.

Insights from this year's survey include:

- 67% of respondents think that funders should withhold funding from, or penalize in other ways, researchers who do not share their data if the funder has mandated that they do so (Figure 1).
- 69% of respondents think that funders should make the sharing of research data part of their requirements for awarding grants.
- while "open data" clearly has more recognition in the community (by virtue of the high response rate to our survey), "FAIR principles" are relatively unknown to the community with 52% of respondents who are frequent data-sharers never having heard of them.

Who took part this year?

In line with 2018, we asked less information about the demography of respondents and focused more on proxy measures that are more universal, which in turn allow for a greater "like for like" interpretation of the data. For example, rather than asking for career stage directly, we asked respondents to state the date of their first publication of a peer-

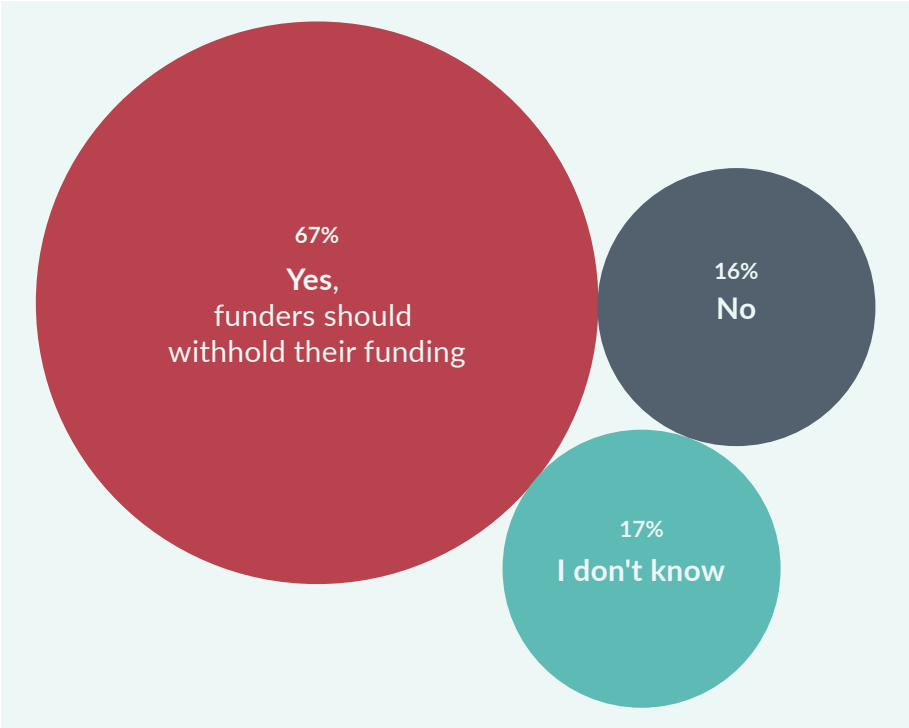


Figure 1: Should funders withhold funding from (or penalize in other ways) researchers who do not share their data if the funder mandated that they should

reviewed research article as well as their current tenure status, to allow us to estimate the stage of career of respondents.

Of those surveyed, a total of 9,500 took part. After removal of low completion rate surveys, a total of 8,423 participants were included in the analysis below.

This year's response cohort included:

- 38% professors
- 41% tenured (with a further 7% on tenure-track)
- 36% had first published a peer reviewed article during the current decade (2010's)

Funding finds

As previously mentioned, researchers see an important role for funders in ensuring good citizenship among researchers: 67% of respondents agreed that funders should withhold funding from, or otherwise penalize, researchers who do not share their data if the funder has mandated that they do so. This opinion was robust across geographies and tenure status. However, Professors were more likely to agree that funding should be withheld in this situation than other respondents. The same bias was also held by those in institutions that have no system of tenure.

A corroborating response confirmed that funders are perceived as important in changing attitudes to engagement with research data, as

"There is more pressure than ever before to make data available for sharing in a timely manner so that it can be reused"

69% of respondents thought that funders should make the sharing of research data part of their requirements for awarding grants. Professors and those on tenure-track were most likely to agree with this position. Although this observation was consistent across geographies, the strength of the agreement was slightly stronger in Africa, Europe and South America compared with other regions.

Licensing

Levels of uncertainty over the licenses used to make data openly available has decreased by 18% since the survey began in 2016. 66% of respondents were unsure about the license used to make their data openly available. This compares to 48% in 2019 and perhaps indicates that, even though there is still confusion over licenses, there is beginning to be less uncertainty (see Figure 2).

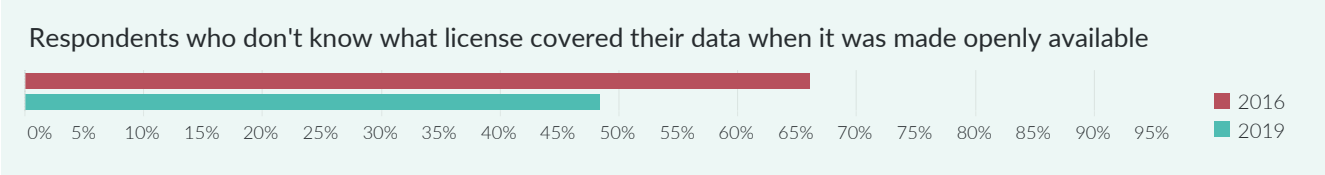
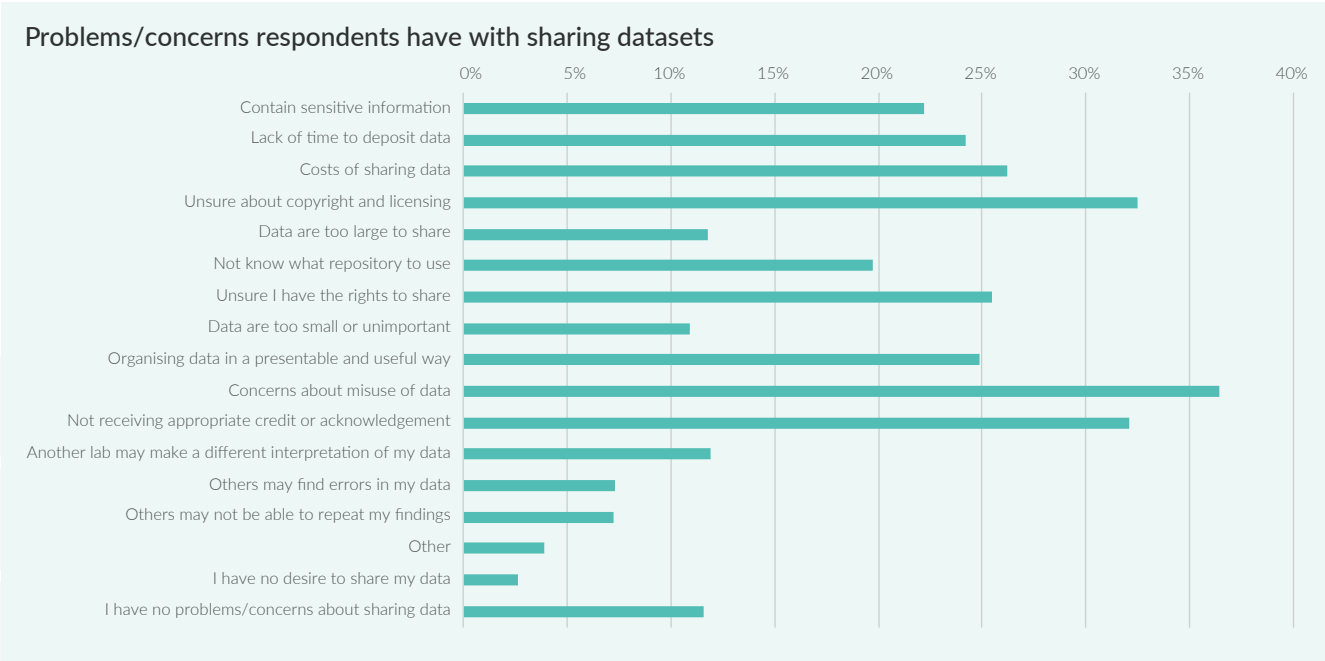


Figure 2: Respondents who do not know the license under which their data was made open

Concerns over sharing data

Sharing research data brings with it many concerns and challenges for researchers. There is more pressure than ever before to make data available for sharing in a timely manner so that it can be reused. The 2019 survey revealed that one of the most pressing concerns was the potential misuse of data. 36% of respondents expressed the concern that their data may be misused if it was shared. Other notable concerns researchers highlighted included the rights researchers themselves have to share their data, the costs involved in making their data available, and lack of knowledge about copyright and licensing. (see Figure 3)

Figure 3: Problems/concerns respondents have with sharing datasets



Is it FAIR?

While open data is clearly established as a topic that is now in the mainstream for researchers, FAIR principles are not so widely known. Indeed, most researchers have not encountered one of the many FAIR initiatives (see Table 1).

2019	I am familiar with the FAIR Principles	I have heard of, but am not familiar with FAIR	I have never heard of the FAIR principles
GO FAIR	5%	18%	77%
FAIRdat	5%	17%	78%
MakeDataCount	11%	24%	65%
DataCite	4%	14%	82%
FORCE	11%	20%	69%

Table 1: Level of awareness of FAIR initiatives

In spite of these many initiatives, 2019 has seen only a slight increase in familiarity with the FAIR principles compared to 2018.

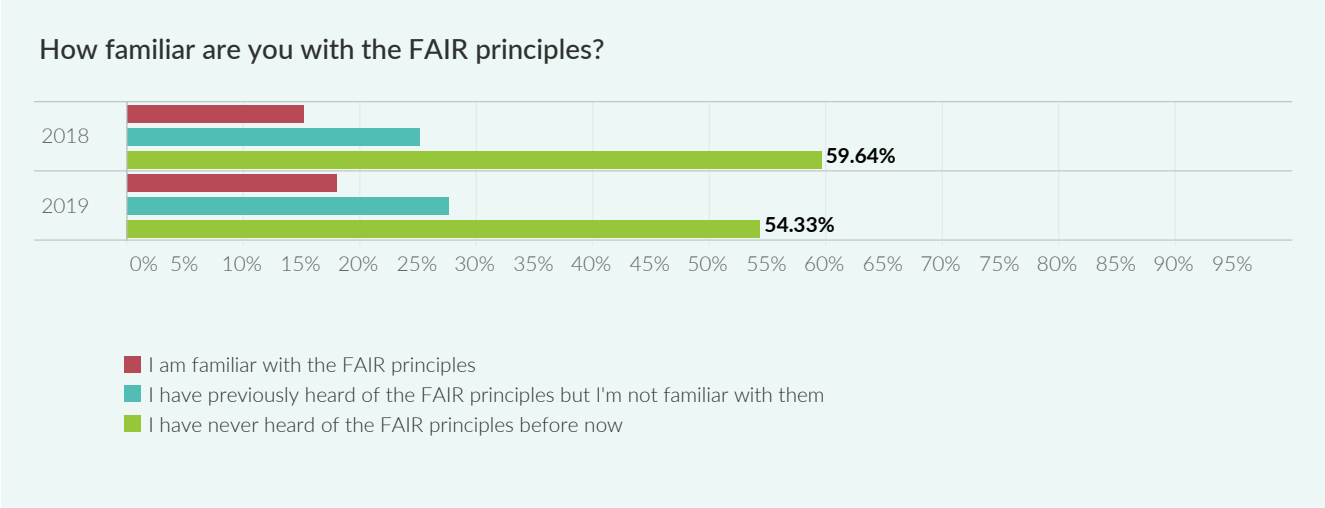


Figure 4: Familiarity with the FAIR principles 2018 versus 2019

Given the importance of FAIR principles in ensuring that open data can effectively be reused, while it is concerning to see such low awareness (Figure 4), it is however heartening that an increasing number of academics have used open data in their research work (Figure 5).

There is clearly a deep lack of understanding in the community around what makes data FAIR. For those reading this and seeing the term FAIR for the first time, the principles require data to be Findable, Accessible, Interoperable and Reusable. In our survey, 20% of respondents who have never shared their data stated that they think their data is compliant with FAIR principles either sometimes or mostly. Further efforts need to be put into educating the community about FAIR.

Of course, while the aim is that everyone will know about and apply FAIR data principles, the fact that colleagues are willing to not only reuse open data but also make data available is a big step forward. Our survey found that 60% of respondents who had never used open data in their research would be willing to do so.

"60% of respondents who had never used open data in their research would be willing to do so"

"It is clear that if we are to move the open data cause forwards then credit will play a key role"

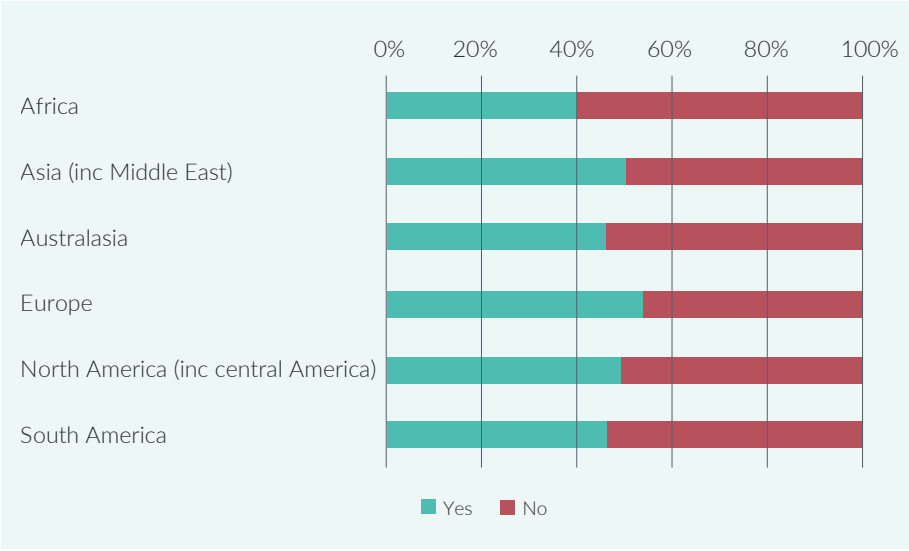


Figure 5: Geographic distribution for reuse of open data

Big takeaways from this year's survey

This year 65% of respondents reported that they curated their data for sharing either privately or publicly. This figure is similar to that for respondents in 2016 (67%) but less than those who reported the same in 2017 (74%) and 2018 (74%).

79% of 2019 respondents were supportive overall of a national mandate for making primary research openly available. This is an increase of 18% from 2018, an increase of 25% from 2017, and matches the findings in 2016 (79%).

"Respondents who first published in the 2000s and 2010s appear to be the most motivated to share their data if it resulted in them being credited as a co-author"

It is clear that if we are to move the open data cause forwards then credit will play a key role. When asked about mechanisms that would encourage more researchers to share their data, full citation (61%), co-authorship (42%), consideration in job reviews (45%) and financial reward (38%) all ranked highly as important mechanisms for researchers as credit for sharing their data openly. Digging a bit deeper, we see that respondents who first published in the 2000s and 2010s appear to be the most motivated to share their data if it resulted in them being credited as a co-author.

View the raw survey data

Full survey data and questionnaire can be found at dx.doi.org/10.6084/m9.figshare.10011788. An interactive visualization of all the data can be found at https://knowledge.figshare.com/articles/item/state-of-open-data-2019.

Data Mandates and Incentives: Steps Publishers Can and Should Take Today

"Linking articles to their supporting data in a repository was associated with on average a 25% increase in citations"

Grace Baynes, VP, Research Data and New Product Development, Springer Nature

This year's State of Open Data survey wanted to understand how researchers' views are evolving with respect to both mandates for open data, and motivations for sharing data. We learned that support for mandates is growing, and credit is increasingly an important motivator. Collaboration across funders, research organizations and publishers is needed to effect real change on both counts, but there are concrete steps that publishers can take today to make a difference.

Motivations to share: a shift to credit and mandates

As in previous years, we asked about motivators to share data. While "Increased impact and visibility of my research" and "Public benefit" remain the top two reasons, "Getting proper credit for sharing data", "Journal/publisher requirement" and "Funder requirement" are notably higher on researchers' agendas. When we asked "Which one of the circumstances would motivate you the most to share your data?", the top seven responses were the same.

Which circumstances would motivate you to share your data?	2019 RANK	2019	Count	2018 RANK	2018	Count
Increased impact and visibility of my research	1	62%	3,659	1	62%	841
Public benefit	2	60%	3,522	2	59%	802
Getting proper credit for sharing data	3	54%	3,172	4	46%	621
Journal/publisher requirement	4	51%	3,009	5	44%	599
Transparency and re-use	5	48%	2,817	3	48%	652
Funder requirement	6	47%	2,767	9	33%	453
Institution/organization requirement	7	44%	2,592	7	38%	522
Trust the person requesting my data	8	43%	2,510	6	41%	561
It was made easy and simple to do so	9	36%	2,102	8	36%	485
Freedom of information request	10	30%	1,789	10	26%	352
It was a field/industry expectation	11	22%	1,272	n/a	n/a	n/a
Other (please specify)	12	3%	191	11	5%	63
I would never share my data	13	2%	94	12	1%	17
TOTAL		**100%**	**5,886**		**100%**	**1,359**

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

What does 'credit' mean when it comes to data sharing?

Just 12% of respondents felt they received sufficient credit for sharing data. When asked "what credit mechanisms do you think would encourage more researchers to share their data?", the most popular responses were citation (61%), consideration in job reviews and funding applications (45%), co-authorship on papers (42%) and financial reward (38%). Based on these figures, it seems that we have a significant gap to close on the question of credit.

The importance of citations and how publishers can help

Researchers citing datasets in their research articles not only makes data easier to find, but puts data on a par with research articles in terms of importance. This is the first principle of the Joint Declaration of Data Citation Principles.[1]

There is growing evidence of a citation advantage in both sharing data and ensuring it is linked to from an article. A recent study[2] classified the data availability statements of over 500,000 articles in PLOS and BioMed Central (BMC) journals. Linking articles to their supporting data in a repository was associated with on average a 25% increase in citations.

While citations are an imperfect measure of impact, they are evidence of visibility. Increasing the impact and visibility of their research is the leading motivator for researchers to share data. Publishers can play a key role by encouraging the use of data citation in reference lists and data availability, by helping authors to ensure they use unique and persistent identifiers, and by ensuring dataset references are well-marked up in article metadata to maximise discoverability. The 2018 *Scientific Data* article, 'A data citation roadmap for scientific publishers',[3] lays out the steps that publishers can take to support this.

This does require time and effort to implement and support, but can be tackled step by step and would pay dividends for journals, researchers and research alike.

Mandates matter

Data sharing policies are becoming more important to researchers, based on the results of this year's survey. We see that in the motivators for sharing, 69% of respondents agreed that funders should "make the sharing of research data part of their requirements for awarding grants".

Journal and publishing requirements are also increasingly important to researchers, which may reflect the growing number of journals that have introduced data policies in recent years. Publishers do not need in-house data policy experts to develop and implement a sound research data policy. The Research Data Alliance's Data policy standardisation and implementation interest group has released a flexible draft framework[4] that any publisher can adopt.

[1] Martone M. (ed.) San Diego CA: FORCE11, 2014 , Data Citation Synthesis Group: Joint Declaration of Data Citation Principles.; *https://doi.org/10.25490/a97f-egyk*

[2] Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K. and McGillivray, B., 2019. The citation advantage of linking publications to research data. arXiv preprint arXiv:1907.02565.

[3] Cousijn, H. et al. A data citation roadmap for scientific publishers. Sci. Data. 5:180259 doi: 10.1038/sdata.2018.259 (2018)

[4] Hrynaszkiewicz, Iain; Simons, Natasha; Hussain, Azhar; Goudie, Simon (2019): Developing a research data policy framework for all journals and publishers. figshare. Preprint. *https://doi.org/10.6084/m9.figshare.8223365*

Following Fogg's Behavioral Model[5] (**Behavior = motivation x ability x trigger**) if motivation is sufficiently high, providing we have the ability to act, we can easily be prompted to do so. We need to continue to make it much easier for researchers to manage and share data, increasing their ability to take action. At the same time, there are clear steps publishers can take today to increase motivation to share, and make it worth a researcher's time and effort to open up their research.

"We need to continue to make it much easier for researchers to manage and share data"

[5] Fogg, B.J., 2009, April. A behavior model for persuasive design. In Proceedings of the 4th international Conference on Persuasive Technology (p. 40). ACM.

"While funder and publisher policies are important drivers, there is growing recognition that credit for data sharing strongly increases the incentive for researchers to make their data open"

University Press Scaffolding for Open Data Credit Mechanisms

Emily Farrell, Library Sales Executive, MIT Press[1]

As an institutionally-based mission-driven publisher, the MIT Press (MITP) stands in a unique position to promote open data. The Press is seeking and finding ways to incentivize researchers to cite and attribute the data they use, and to make their own data available for others to reuse. The trusted relationships that MITP fosters with its editors and authors, as well as its position within the Institute and Libraries provides a valuable place from which to encourage both bottom-up and top-down approaches to open data.

While funder and publisher policies are important drivers, there is growing recognition that credit for data sharing strongly increases the incentive for researchers to make their data open. In an incredibly competitive academic job market, much is at stake. Existing credit mechanisms in research processes, primarily those most central to tenure, promotion, and grant allocation assessment focus on incumbent models of output: journal articles and books. These systems focus less on complex and varied research objects like datasets. Researchers and their institutions continue to highly value citation and authorship, something that is confirmed in Figshare's 2019 State of Open Data survey. While the greatest majority in this year's survey favor full citation (60.9%) of their data as the credit of choice, many view co-authorship highly (42.4%). This is not surprising, as full citation and co-authorship both feed into existing mechanisms for research assessment. There are additional factors of lesser importance, for example financial incentives, that are part of credit for open data. Alongside this is the researcher's belief that 'they receive too little credit' for open data (63.9% of respondents).

Systems for giving and receiving recognition in scholarly communication operate within a social framework. There is a growing body of work on the social aspects of data sharing as a part of citation and authorship workflows. As Borgman so incisively writes of the complexity of citation mechanisms: "Technical mechanisms for citation are only surface characteristics of the knowledge infrastructures in which they are embedded. Social conventions underlie citation practice, whether to publications, data, documents, webpages, people, places, or institutions."[2] Particularly useful, in regard to the power of this intersection of social and technical, is the notion, developed in recent qualitative research from Ithaka S+R, of data communities as drivers for open data. Rather than merely pushed forward through top-down mechanisms such as funder requirements or publisher policies for open data citation, 'data communities thrive when they cultivate formal or informal norms through which data sharing comes to be expected within the community.'[3]

[1] With thanks to Amy Brand (Director, MIT Press), Gita Manaktala (Editorial Director, MIT Press), Nick Lindsay (Director of Journals and Open Access, MIT Press), and Anne Ray (Senior Journals Editor, JSTOR) for input.

[2] Borgman, C. L. (2015). Credit, Attribution, and Discovery of Data. In Big Data, Little Data, No Data: Scholarship in the Networked World (pp. 241–270). *https://doi.org/10.7551/mitpress/9963.003.0015*. p. 242.

[3] Cooper, D., & Springer, R. (2019, May 13). Data Communities: A New Model for Supporting STEM Data Sharing. *https://doi.org/10.18665/sr.311396*

Using a framework of data communities to consider credit for data sharing allows for a rethink of the possibilities of collaborative research, and of how co-authoring may extend our ways of thinking within disciplinary boundaries. This framework also offers the potential to assess how open data promotes the social justice[4] concerns of the open science movement. As we build guidelines, principles, and protocols to support researcher-driven efforts to make their data open, with adequate credit, we owe it to the aspiration of equity in open science to look closely at potential biases in these new mechanisms. Recent studies that have examined, for example, gender biases in citation practices[5] deserve closer attention. The smaller, cross-disciplinary scale of current data communities opens a potential window to analyze the differences in mechanisms across research practices. Declarations like DORA (Declaration on Research Assessment), with its wide range of individuals, institutions, and publishers as signatories, call for a reconsideration of all research output, including datasets (https://sfdora.org/read/). In particular, DORA calls on publishers to diversify assessment metrics. Rethinking how research output is assessed provides a perfect opportunity to examine how each vested part of the system might scaffold open data, the way it is credited, and how we might build that support with equity in mind.

"Data communities thrive when they cultivate formal or informal norms through which data sharing comes to be expected within the community"

"We owe it to the aspiration of equity in open science to look closely at potential biases in these new mechanisms"

As a member of MIT's Open Access Taskforce, led jointly by faculty and the Libraries, the MIT Press has a seat at the table in Institute-wide discussions of open data. Because of the privacy concerns that accompany certain types of datasets, the MITP has not instituted a universal mandate for open data in our journal publications. Instead, the Press works with authors who want to link to open data, whether in books or journals. The Press's data policy,[6] which endorses Force11's Data Citation Principles,[7] asks that authors cite data and make that data available, where possible. The Press steers authors to DataCite's guidelines on data citation and authors are encouraged to contact the Press where they have questions. *The Review of Economics and Statistics*[8] frequently includes links to datasets. The Open Access journal *Data Intelligence*,[9] launched in 2019, includes traditional journal article formats as well as data-centric "data articles". In all of this, the Press works from a core strength: the trusted relationships with researchers. As the publishing landscape changes, so does the work of shepherding manuscripts from idea to publication. Attention to open data, as funders require it of publications, but also as scholars build data sharing into their research practices from the ground up, provides an opportunity for presses like MITP to do what it does best: scaffold the work of researchers to drive change in attitudes.

We are in the early days of the push from all sides to make data open and to ensure credit for open data. As Borgman notes, these early choices in innovation and change can have surprisingly long lasting effects.[10] We need to make open data matter, and one way to ensure that is the reuse of data that adequately credits the author, gatherers, creators, and even participants of those datasets. Another is to ensure that we build, support, and promote credit mechanisms that pay attention to the social mechanisms embedded in citation and authorship; not something that publishers, or funders, or researchers can do on their own.

[4] See for example an overview here from April Hathcock, NYU Libraries: *https://aprilhathcock.wordpress.com/2016/02/08/open-but-not-equal-open-scholarship-for-social-justice/*

[5] See Vettese, Sexism in the Academy (2019), for an overview some of these studies: *https://nplusonemag.com/issue-34/essays/sexism-in-the-academy/*

[6] *https://www.mitpressjournals.org/data_policy*

[7] Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk

[8] *https://dataverse.harvard.edu/dataverse/restat*

[9] *https://www.mitpressjournals.org/loi/dint*

[10] Borgman, C. L. (2015). Credit, Attribution, and Discovery of Data. In Big Data, Little Data, No Data: Scholarship in the Networked World (pp. 241–270). *https://doi.org/10.7551/mitpress/9963.003.0015. p. 265.*

What does this mean for MIT Press?

We continue to work from our strengths. We continue to guide and shepherd our authors towards protocols and best practices for open data through the close relationships we foster between our editors and authors, our libraries and institutions. We continue to encourage authors to link to data, to offer suggestions on where to host that data and what policies may be relevant, and to guide them to relevant experts as needed.

Contributor Biographies:

Paul Ayris is Pro-Vice-Provost at UCL Library Services. He joined UCL in 1997 and was the President of LIBER (Association of European Research Libraries) from 2010-14. He is Co-Chair of the LERU (League of European Research Universities) INFO Community. He chairs the OAI Organizing Committee for the Cern-Unige Workshops on Innovations in Scholarly Communication. He is also Chair of JISC Collections' Content Strategy Group. On 1 August 2013, Dr. Ayris became Chief Executive of UCL Press. He is a member of the Provost and President's Senior Management Team at UCL. He has a Ph.D. in Ecclesiastical History and publishes on English Reformation Studies. In 2019, he was made a Fellow of the Royal Historical Society.

Email: p.ayris@ucl.ac.uk
https://orcid.org/0000-0002-6273-411X

Mark Hahnel is founder and CEO of Figshare. Mark created Figshare 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.

Email: mark@figshare.com
http://orcid.org/0000-0003-4741-030

Iain Hrynaszkiewicz is Publisher, Open Research at Public Library of Science (PLOS), where he leads the conceptualization and development of new products and services that add value to the PLOS portfolio by supporting and enabling open science. Iain was previously Head of Data Publishing at Springer Nature where he developed and implemented research data policies and services, and was publisher of Nature Research Group's Scientific Data journal. He has also been Outreach Director at Faculty of 1000 (F1000), and spent seven years at the first commercial open access publisher BioMed Central (BMC) in a variety of editorial, publishing and product/policy development roles. Iain is part of several research/publishing community projects related to data sharing and reproducible research. He founded and is co-chair of an Interest Group in the Research Data Alliance (RDA) that is setting standards for journal research data policy globally, and founder of the annual early-career researcher conference, Better Science through Better Data. He has published numerous papers related to data sharing, open access, and the role of publishers in reproducible research - one of which has been cited nearly 200 times.

Email: ihrynaszkiewicz@plos.org
https://orcid.org/0000-0002-9673-5559

Briony Fane is a Data Analyst at Digital Science. She came to Digital Science from a higher education background, having gained a PhD from City, University of London, and worked as a researcher and subsequently as a research manager. She coordinates and manages Digital Science's reports, playing a major role in Figshare's State of Open Data report where she analyses and writes up the outcomes from the annual state of open data survey. She is also Digital Science's Catalyst Grant Coordinator.

Email: b.fane@digital-science.com
http://orcid.org/0000-0001-6639-7598

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.

Email: g.baynes@nature.com
http://orcid.org/0000-0002-4933-3186

Emily Farrell is the Library Sales Executive for the MIT Press, where she manages North American library relations. She is a member of the press's Open Access Steering Committee and is involved in the 2019 Arcadia Foundation grant to develop a business model for open access scholarly books at the press. Her interest in open science stems from her current role with libraries, as well as her experience as a researcher and editor. She completed her doctoral work in sociolinguistics at Macquarie University, Sydney, where she has an ongoing affiliation working with the Language on the Move research group. Emily serves on the board of UnLocal, a legal services and educational outreach organization that works with undocumented immigrants.

Email: efarre@mit.edu
https://orcid.org/0000-0002-5364-2643

Part of **DIGITAL**science

digital-science.com