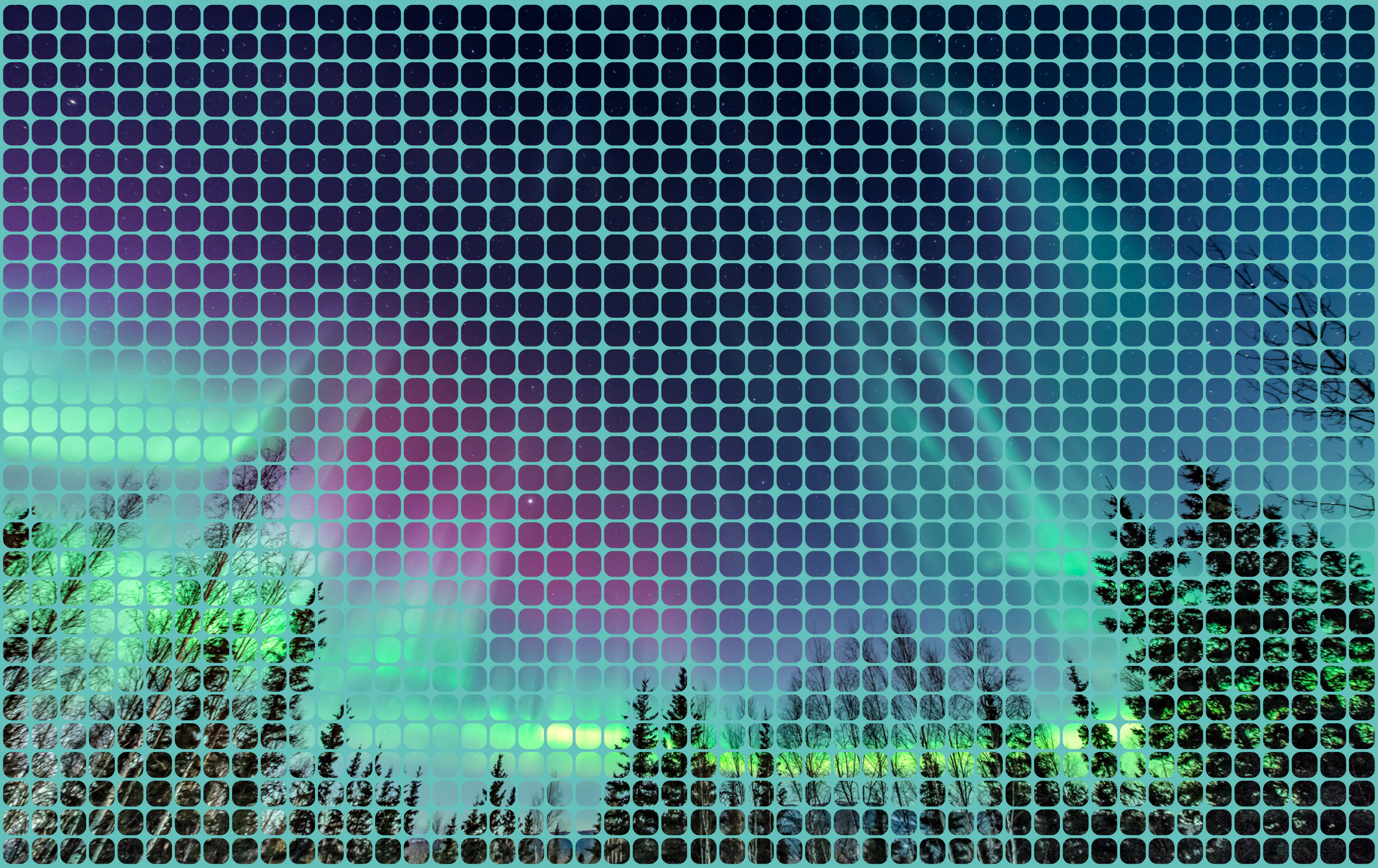


Digital Science Report

The State of Open Data 2018

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

Foreword by Ross Wilkinson

October 2018

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, Dimensions, Figshare, ReadCube, Symplectic, IFI Claims, GRID, Overleaf, Labguru, BioRAFT, UberResearch, TetraScience and Transcriptic. 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**, in particular Dan Penny for his expertise with survey design and hosting, and the **Springer Nature** marketing team for distributing the survey globally. We'd also like to thank **Wiley** for their assistance.

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, 4 Crinan Street, London N1 9XW, UK. info@digital-science.com
Figshare, 4 Crinan Street, London, N1 9XW, UK, info@figshare.com

DOI: https://doi.org/10.6084/m9.figshare.7195058
ISBN: 978-1-9993177-0-6

SPRINGER NATURE
WILEY

Contents

Foreword

Ross Wilkinson, Australian Research Data Commons (ARDC)

"The value of research is no longer taken for granted."

The value of research is no longer taken for granted. In a world where news is questioned and authority is no longer relied upon, research must be able to demonstrate its value transparently and clearly. Increasingly funders of research are requiring verifiable quality and the translation of research into outcomes of value to society. The publication of a research paper in a prestigious journal is no longer enough. Communication that is only between peers, and not available beyond, is not adequate. There is a pressing need for open research that delivers value beyond answers.

"A key to both quality and translation is trust in the underlying data."

Open data is a key element of open research, and that data should be FAIR (Findable, Accessible, Interoperable, Reusable). It should be discoverable, available, but importantly it has to be able to be used in new ways – typically via robust services as scale increases. It is very important that data is open – but that is not enough! The quality of the data has to be able to be assessed. The data should have rich descriptions that enables them to be integrated for further investigations, and they should also be available for machine interrogation, and interaction. A key to both quality and translation is trust in the underlying data, so systems and process are also very important.

Figshare's State of Open Data report is an important yearly landmark to reflect on our progress against these pressing needs. It is clear that a partnership is necessary to achieve the dramatic change that is needed quite quickly. Such a partnership needs to span from funders who provide new infrastructure and incentives to make open data services available, to companies that provide critical products and services, to research institutions to establish the right environment, to the data professionals and researchers who need to form new partnerships to enable research to be conducted differently.

Such a breadth of partners is critical because we are seeing not simply a change to making research more open, but a transformation of how research is conducted. New infrastructure is necessary, new professionals are required, new incentives are needed, if research is truly to deliver the promise of high quality research that is more widely used and relied upon. We see these changes taking place around the world. The African Open Science Platform is being established, and the Japanese Open Science strategy, the US NIH data commons, Australia's research infrastructure providers FAIR data strategy, supported by an Australian Research Data Commons, and the European Open Science Cloud all demonstrate enormous commitment to this change.

This change is likely to occur increasingly rapidly as the need becomes ever clearer, and there is an increasing international consensus that this has to be done together. The Research Data Alliance is providing a forum for developing consensus on data interoperability. The joint meeting with CODATA[1] and WDS[2] taking place in Botswana this year indicated the need for a global approach to the challenge and opportunity. Figshare's 2018 State of Open Data report is a great chance to reflect on our progress, and where we need to focus our attention next.

[1] *http://www.codata.org*

[2] *https://www.icsu-wds.org/organization*

Fundamental Change in Academia Without Anyone Needing to Die

Mark Hahnel, Founder and CEO, Figshare

To paraphrase Max Planck “Academia advances one funeral at a time”.

Fortunately, due to funder policy changes, the underlying intention of the academic is irrelevant as researchers both old and new are sharing, citing and reusing data more than ever before.

What the 2018 State of Open Data report demonstrates, is that whilst there are myriad reasons for early-career researchers to make their data available from ethical and moral to career advancement, the naysayers really do not have a choice in the matter and are along for the ride.

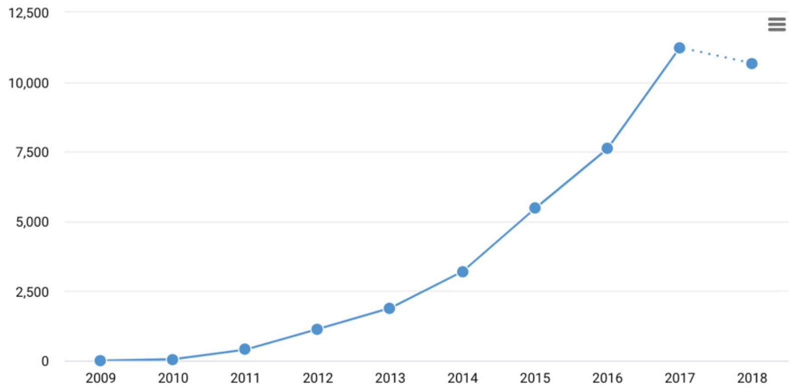
Driven increasingly by funder mandated open access to research data, we see a sustained increase in the percentage of academics making data openly available. 64% of respondents revealed they made data openly available in 2018 - up 4% on 2017 and 7% on 2016.

Early career researchers are focussed on the credit they receive for making data available, with regards to career progression opportunities. Many more researchers who first published in the 10s would be motivated by co-authorship as credit for sharing their data than those first publishing in previous decades.

"Early career researchers are focussed on the credit they receive for making data available."

Perhaps this is why early career researchers are more concerned about educating themselves in the fundamentals of the space. As an example, our reports shows these researchers are more likely to understand the licences with which they have make their research data openly available.

This also mirrors data from Dimensions.ai[3], which illustrates the growth of citations to generalist repositories Figshare.com, Dryad and Zenodo. The hockey stick graph indicates the exponential growth of datasets that are being made available.



"Librarians the world over have been working hard to enlighten researchers at their institution of their responsibilities."

This is great for the goals of open research, the ability for anyone anywhere to build on top of the research that has gone before - the democratization of knowledge. However, what the data also shows is that the logistical pathway to really make the research useful has many hurdles yet to pass.

Namely, 'FAIR' data. The four principles — Findability, Accessibility, Interoperability and Reusability (FAIR) provide a guideline for data producers and publishers to enhance the reusability of academic data. FAIR has been pushed as a major ambition for funders around the globe. While some of the definitions of these principles are still being worked out, it was a shock to see that nearly 60% of respondents 'had never heard of the FAIR principles before now'. The 2017 State of Open Data report highlighted the need for education at the level of the researcher. Librarians the world over have been working hard to enlighten researchers at their institution of their responsibilities, from Data Management Plans to suitable repositories.

Interoperability was highlighted by those who had heard of the FAIR principles to be the one that caused most confusion. In a recent article[4], I commented on the four long standing big picture challenges to be addressed in order to achieve useful open research;

1. Research files need to be put on the web in a persistent manner
2. Files become accessible in a standardized way
3. Files need to be indexed in a consistent manner
4. Tools for analysis need to exist

"We are at the dawn of the democratization of data without barriers and the transparency and the subsequent knowledge leaps that come with it."

Points 2 and 3 highlight the need for FAIR principles, with indexing requirements bolstered by Google Dataset's launch in the space, complementing previous academic dataset search engines, such as that provided by DataCite[5].

The last 12 months has seen huge organizational (International Data Week), technological (Google) and policy driven strides (Go FAIR[6]). The next year will see a clarification of definitions and hopefully a wealth of education, especially to those about to be hit by a wave of 2020-based funder mandates that are being led by the European Commission.

The future challenges remain around the continuing education for researchers whose funding could be pulled for non-compliance.

At Figshare, we will continue to build technological solutions to reward researchers in a way they care about, whilst bolstering easy compliance. The combination of technology and ongoing policy updates continues to ramp up progress in the Open Data space. As such, gone are the days that funerals dictate progress. We are at the dawn of the democratization of data without barriers and the transparency and the subsequent knowledge leaps that come with it.

[3] *Dimensions.ai*

[4] *https://figshare.com/blog/What_Google_Dataset_Search_means_for_academia/422*

[5] *https://search.datacite.org/*

[6] *https://www.go-fair.org*

The FAIR Future of Repositories as a Critical Component of the Internet for Machines (and People)

Barend Mons, Leiden University Medical Center, **Erik Schultes**, Leiden University Medical Center & Luiz Olavo **Bonino da Silva Santos**, GO FAIR International Support and Coordination Office

Open Science is emerging as a solution to a nice problem we created ourselves

As technical and scientific advances continue to bulldoze their way through society, exciting possibilities, alongside severe challenges emerge. The explosive growth of data and computer resources promise revolutionary modes of discovery and innovation not only within traditional knowledge disciplines, but also between them. The challenge, however, is to build, curate, re-use, properly fund, and maintain the large-scale, widely accessible, and automated infrastructures. These will be necessary for navigating and managing the unprecedented complexity of exponentially increasing quantities of distributed and heterogeneous data. This will require innovations in both the technical and the social domains. Consequently, both science and innovation are in a methodological phase transition.[7]

Because of this phase transition, we move from a closed, individual-privilege-patent- and 'center of excellence' based system to a system that has to support fully distributed, collective human and machine intelligence much more effectively. This is generally seen as the core of the hip term 'open science', which already enjoys many definitions. Moreover, on top of *the Internet for People*, we now need an *Internet for Machines*, in which machine actionable data and services will play a central role.

Data Stewardship is at the core of Open Science

Unfortunately, our ability to deal responsibly with data as the principal first-phase output of the scientific process has not kept pace with generation and storage-capabilities. The current reality is a glaring lack of expertise, a crippled practice of cottage-industry with incompatible and fragmented data stewardship approaches. Combined with an almost complete lack of interoperability of data in domain silo's and a hopelessly outdated scholarly publication and reward system, it is effectively slowing down the transition to open science and innovation. Numerous studies indicate that data scientists both in academia and industry spend 70-80% of their time on mundane, manual procedures to locate, access, and format data for reuse.

[7] *Schultes E, Strawn G, Mons B. Ready, Set, GO FAIR: Accelerating Convergence to an Internet of FAIR Data and Services. Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018), Moscow, Russia, October 9-12, 2018*

"Our ability to deal responsibly with data as the principal first-phase output of the scientific process has not kept pace with generation and storage-capabilities."

We need a new science and innovation ecosystem

The urgency for automated, commonly usable data infrastructures (i.e., an Internet for Machines) is increasingly recognized by numerous national and international organizations, science funders and industry. Despite this need, building a generalized, ubiquitous data infrastructure that is widely used by diverse stakeholders is an inherently difficult process to direct.

In the past three years, we have seen a rapid development of machine-oriented initiatives such as the formulation of the FAIR principles, describing how data should be framed, published and stewarded in a way that supports optimal reuse in open science and innovation for both machines and humans. It is high time for modern data and tool repositories to become champions of publishing machine-actionable (meta)data, according to FAIR principles, supplemented with narrative for humans, to help science and innovation.

This is (like science itself) a global issue

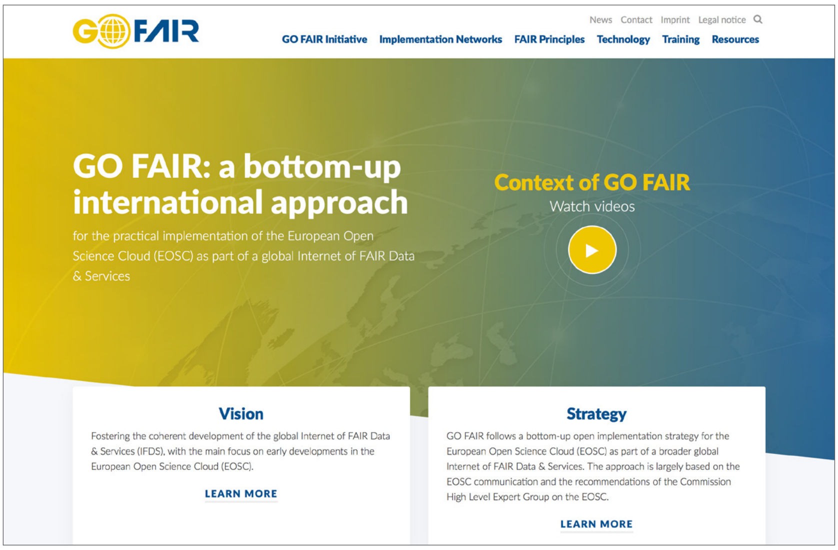
Open Science and Innovation need a global, equally accessible, democratic and performant infrastructure, comparable to the Internet and the WWW as we know it. Where possible it should be built on proven technical components and protocols. The infrastructure should also allow for proper security and privacy of data and services when 'open' is simply not an option. It is therefore of critical importance to re-emphasize here that FAIR principles cater for both open and restricted data and services access.

Globally mandated data-focused organizations are likely to play a role here, but could also suffocate the developments when overregulation, standard setting, governance and regulation seize power. A number of global players have various kinds of mandates and niches in the realm of data. Three complementary key organizations, CODATA, RDA and GO FAIR are all international and cross-disciplinary in scope, mandated and poised to support the global science enterprise including Pan-European, and global, domain specific research and industry infrastructures and e-infrastructures.

How does FAIR play into this pattern?

Even before the 2000s, visionaries had already anticipated the need for a general-purpose data infrastructure. Digital Object Architectures (DOA), systems supporting Persistent Identifiers (PIDs) and the Semantic Web (a framework for knowledge representation built on top of existing Internet and WWW infrastructures) appeared as important components, ensuring both data interoperation and machine readability. Since then, difficult problems in this space have been confronted resulting in a plenum of new, co-existing methods, languages, software and specialized hardware, producing by now, a protracted period of Creolization (numerous experimental implementations resulting in an uneven landscape of independently developed prototypes).

By 2012 the Attraction phase (attracting the attention of others working in the field) was underway with public discussions about component specifications, principles and procedures for semantically enabled data



GO FAIR: a bottom-up international approach

for the practical implementation of the European Open Science Cloud (EOSC) as part of a global Internet of FAIR Data & Services

Vision

Fostering the coherent development of the global Internet of FAIR Data & Services (IFDS), with the main focus on early developments in the European Open Science Cloud (EOSC).

Strategy

GO FAIR follows a bottom-up open implementation strategy for the European Open Science Cloud (EOSC) as part of a broader global Internet of FAIR Data & Services. The approach is largely based on the EOSC communication and the recommendations of the Commission High Level Expert Group on the EOSC.

www.go-fair.org

> "It is high time for modern data and tool repositories to become champions of publishing machine-actionable (meta)data, according to FAIR principles."

infrastructures. By early 2014, in a workshop hosted by the Lorentz Center (Leiden), this discussion culminated in the generalized and broadly applicable FAIR Principles for data reuse. In a now widely cited commentary (indicative of the Attraction Phase), the FAIR approach had been defined as “Data and services that are findable, accessible, interoperable, and re-usable both for machines and for people” and 15 high-level Principles had been articulated. Immediately following their publication (April 2016), the FAIR Principles have been acting as a powerful attractor in the emerging data infrastructure, but the Convergence phase (a compelling globally operational infrastructure) is the necessary next step, before wide exploitation of FAIR compliant components can rule the exploitation phase (what was hard and cost-prohibitive now becomes easy and affordable) which will commence once a ‘critical mass’ of users commits to particular, minimal specification for automatic routing of FAIR data and services. This is also where professional repositories play a major role.

How does GO FAIR fit into this pattern?

Given that many different combinations of technology and use of standards choices could conceivably be made when implementing the FAIR Principles, the GO FAIR[8] initiative was launched in late 2017 by the Dutch, German and French governments as a means to pragmatically accelerate community convergence, based on the vision for the European Open Science Cloud. Following the examples of the Internet and WWW, GO FAIR operates through voluntary stakeholder participation attempting to reach a ‘critical mass’ of users committed to a set of absolute minimal technology specifications. Beyond these minimal specifications, there is unrestricted room to innovate. GO FAIR is stakeholder governed, and works with researchers from specialized knowledge domains and also policy bodies, publishers, repositories, and funding agencies. The number of implementation networks is rapidly growing. Repositories also start to converge on FAIR implementation choices.

[8] *https://www.go-fair.org/*

What is the State of Open Data in 2018?

Trends and comparisons with 2016 and 2017.

Briony Fane, Data Analyst, Digital Science, London, UK
and **Jon Treadway**, COO, Digital Science, London, UK

What do we know about anything

2018’s State of Open Data survey showed a decline in the total number of respondents, which makes some analysis harder, but still allows us to be confident in the validity of our results.

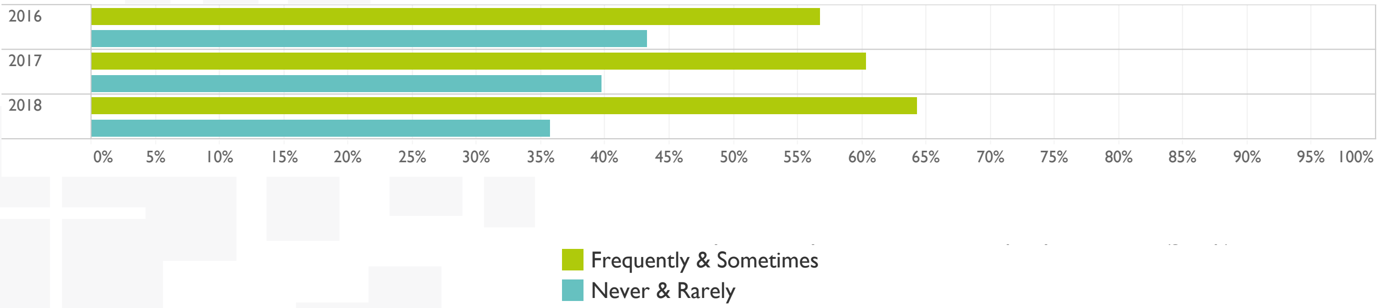
We have also asked less information about the demography of respondents, notably removing the age category. We are instead focusing on when respondents first published research as a proxy for the stage in career they have reached.

In headline terms:

- Awareness of open data remains high.
- More researchers are making data openly available.
- Fewer researchers are losing data.
- Attitudes to credit for data publication are more nuanced.

And yet trends are not uniformly in the direction we would expect for a sector where open data is an increasingly common topic of discussion.

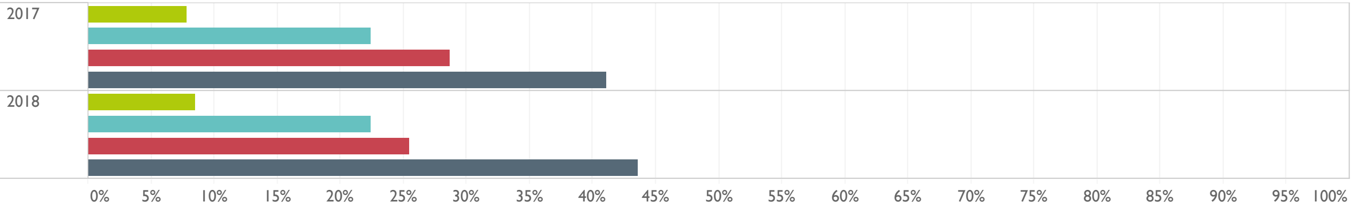
Fig 1. How often researchers have made their data openly available



Movers and shakers

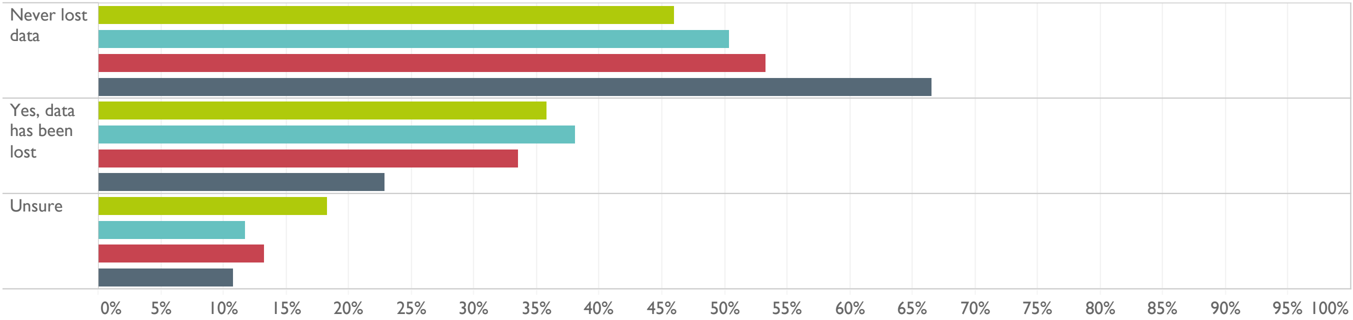
64% of respondents revealed they made data openly available in 2018 - up 4% on 2017 and 7% on 2016 (see Fig. 1). Data citations are motivating more respondents to make data openly available, increasing 7% from 2017 to 46% (see Fig. 2). We see no change in researchers' awareness of open datasets based on geography. Increased impact and visibility of research and public benefit remain the most potent motivators behind a researchers decision to share.

Fig. 2 Getting citations as motivation for making data openly available



Respondents having lost research data has decreased from 2017 (36% versus 30% in 2018). The increasing opportunities to store their data in spaces where it is much harder to lose it would be a good candidate to explain this result. When analysed against when respondents first published a peer reviewed article, those who first published in the last decade were least likely to have lost data - 66% have never lost any (see Fig. 3). External hard drives (16%) and PC hard drives (45%) were the most common places where data is lost.

Fig 3. Date of first publication and losing data



The percentage of respondents in support of national mandates for open data is higher at 63% than in 2017 (55%) but remains down on 2016 where 78% were in support.

Various metrics appear to have plateaued after growth between 2016 & 2017. 80% of respondents stated that they were aware of open data to reuse in 2018, compared to 81% in 2017. A similar number state that they are willing to reuse data - with 83% in 2017 and 84% in 2018.

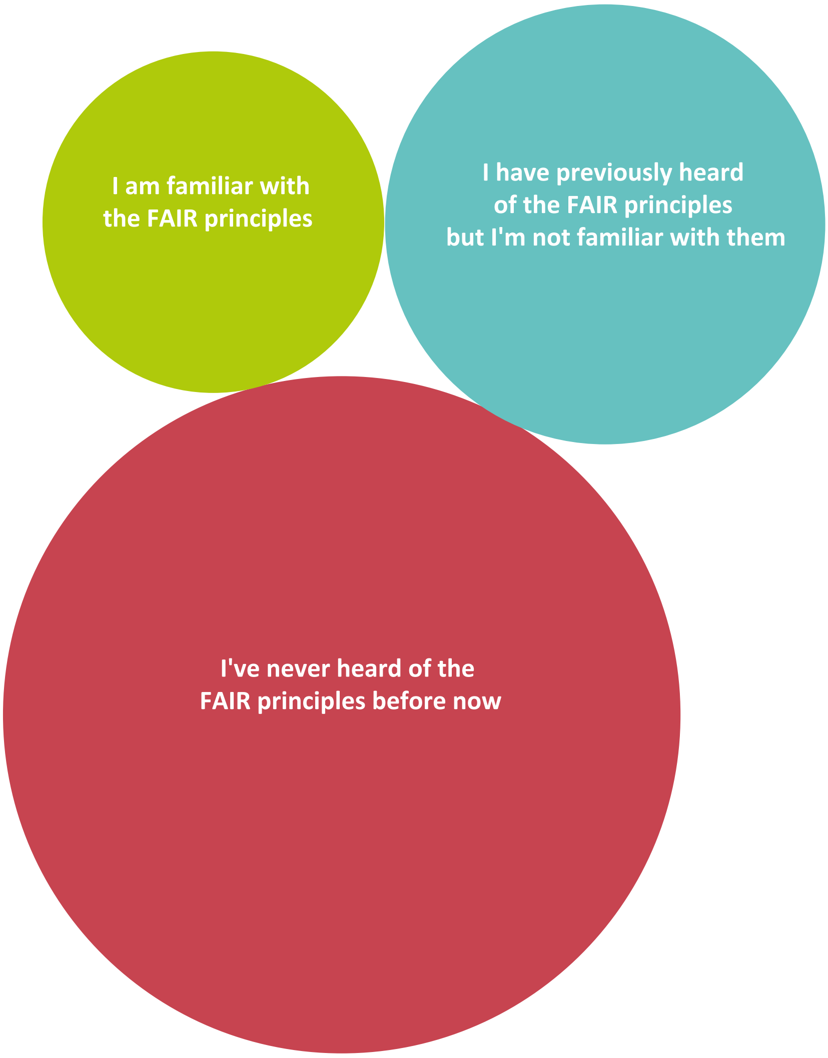
And then there are areas where, instead of progress and consolidation of attitudes to open data, we see signs of regression. Respondents who revealed that they had reused open data in their research continues to shrink. In 2018 48% said they had done this, whereas in 2017 50% had done so, with 2016 57% in 2016.

Comparing how researchers think that others' research data would benefit them has not changed.

There is a marked drop in researchers valuing a data citation the same as an article citation. 55% of researchers responded this way in 2018, but this is a decrease from 2017 and 2016 where 67% and 68% of researchers valued both types of citation equally. There is no change across all three years in those researchers valuing a data citation more than an article citation - 10% - but those valuing a data citation less has increased to 30% up from 20% in 2017.

Understanding these trends is not simple, and further investigation is warranted.

Fig. 4 How familiar are you with the FAIR principles



All is FAIR in love and war

We have asked a number of questions about the FAIR principles this year - i.e. the notion that data should be Findable, Accessible, Interoperable, and Reusable.

The percentage of respondents who reported being familiar with the FAIR principles was just 15%, with 25% having previously heard of FAIR and 60% never having heard of them (see Fig. 4). This lack of awareness is concerning as the FAIR principles are being rapidly adopted by publishers, funders and institutions worldwide but there is a crucial gap in educating researchers on what is expected of them.

With regard to how well defined the FAIR principles are, of respondents who reported being *familiar* with the principles, 41% felt that Interoperable needed better definition, 13% felt that Findable needed better definition, 19% felt that Accessible and 26% felt that Reusable needed better definition. What is very clear is that respondents who report being familiar with the FAIR principles are also more likely to make their data available in compliance with them (see Fig. 5).

These results confirm the need for initiatives like GO FAIR, which gives researchers clear instructions on how to be FAIR compliant and needs wider awareness and adoption.

> "This lack of awareness is concerning as the FAIR principles are being rapidly adopted by publishers, funders and institutions worldwide but there is a crucial gap in educating researchers on what is expected of them."

Fig. 5 Familiarity with FAIR principles and the extent with which researchers make their data available in compliance with the FAIR principles

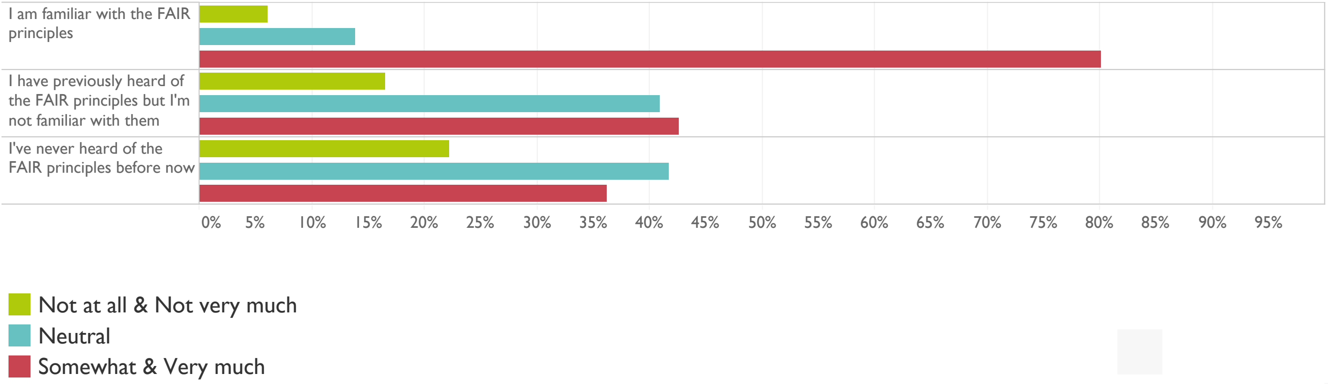
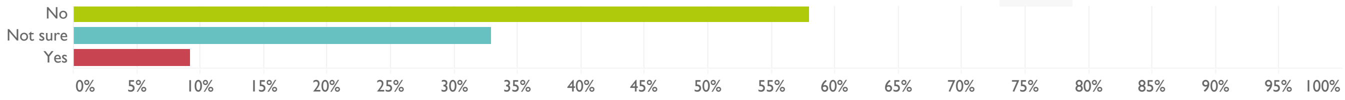


Fig. 6 Do researchers think they get sufficient credit for sharing data



[8] *https://www.go-fair.org/*

What else is new?

When asked where researchers publish their data, 35% of 2018 respondents revealed that they published their data as an appendix to a research article, with little change from the 34% in 2017. Slightly fewer respondents reported publishing data in a data journal in 2018 (18%) than they did in 2017 (20%). 33% reported publishing data in a specific data repository compared with 29% in 2017.

53% of respondents in 2018 who routinely make their data open do not know where the funds to do so come from. This result is markedly higher than for 2016 and 2017 where the figures were 30% and 36% respectively. Respondents are approximately as likely to use their own funds (37%), money from a funder (39%) or funds identified in their grant (41%) to make data openly available.

There were a number of additional areas of insight due to new questions.

Most researchers felt that they did not get sufficient credit for sharing data (58%), compared to 9% who felt they do (see Fig. 6).

Researchers felt they need most help with copyright and licencing, ahead of data curation (see Table 1). Moreover, researchers on the whole are willing to let others help with curating their data, only 5% said they would not let anyone else work on their data.

Table 1 Areas in which respondents need help to make data openly available

Finding appropriate repositories for deposition of data	Curation of data	Data management policies	Copyright and licencing	Data management plans
17%	16%	16%	22%	15%

Tied to the 90s

We found good balance in the percentage of respondents first publishing in different time periods, at least when grouping into decades. Irrespective of when a researcher first published, they routinely make their data open for sharing and are equally aware of open data.

Researchers who first published in the last two decades were more willing to let others support them in curating their data for sharing publicly (see Table 2).

Table 2 Date of first publication and willingness to let others support them curate their data

First publication	Publisher	Peers	Library	No-one
Before 1990	42%	57%	39%	12%
90s	43%	47%	33%	26%
00s	59%	61%	47%	9%
10s	53%	63%	42%	10%

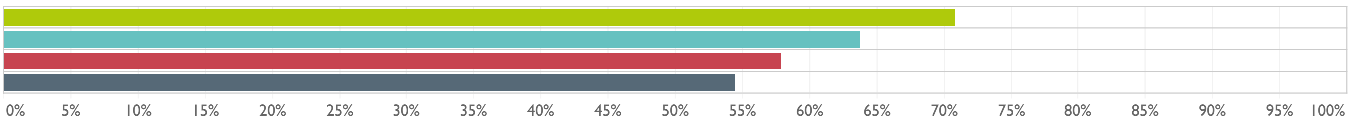
Table 3 Co-authorship as a motivation to make data openly available	
Date of first peer reviewed publication	**Percentage of respondents who would be motivated to make their data openly available for being credited as a co-author**
Before 1990	7%
90s	11%
00s	18%
10s	27%

Table 4 How researchers value data citations			
Date of first peer reviewed publication	**Respondents who value a data citation more than an article citation**	**Respondents who value a data citation the same as an article citation**	**Respondents who value a data citation less than an article citation**
Before 1990	1%	8%	5%
90s	2%	9%	6%
00s	2%	13%	10%
10s	4%	24%	11%

We found that the more recently a researcher first published, the more motivated they would be to make their data openly available if they were credited as a co-author (Table 3). We also found that respondents, irrespective of when they first published, did not necessarily value a data citation more than an article citation. However, those who first published most recently were slightly more likely to value both equally (see Table 4).

Looking at the licences under which researchers make their data open, we see virtually no change in behaviour from 2017, most researchers are still unsure under what licence they have made their data openly available. When looking at date of first publication and licencing, 71% of researchers who first published before 1990 were unsure under what licence they had made their data available under. This reduced to 64% in the 90s, 58% in the 00s and 54% in the 10s (see Fig. 7).

Fig. 7 Unsure of license and date of first publication



One of the proposed subtitles for this article was ‘FAIR. Huh. Yeah. What is it good for?’, but we rejected it in part because it seems clear that FAIR principles are good for a great deal of things. As they gain prominence and acceptance, we expect to see consolidation and expansion of awareness and acceptance of open data, and more clear trends in its widespread adoption.

From Green Shoots to “Grassroots”: How Can We Accelerate Data Sharing?

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

"Globally more than 50 funders now require data sharing, with the majority based in the US and Europe, particularly the UK."

Figshare tracking of researcher’s attitudes and actions in data sharing continues to bear new insights. Now in its third year, the 2018 survey shows some encouraging progress in respondents reporting making data openly available – up consistently year on year since 2016 to 64% in 2018. More researchers are also reporting publishing data in a specific data repository this year (33%) compared to 29% in 2017, which is great for making data more findable, accessible, usable and citable. Yet a closer look makes clear the work still to be done.

A large percentage of this year’s respondents do not feel they get credit for sharing data, and publications in data journals remain a fraction of the world’s annual publications. With funders and institutions seeking “grassroots” support for data sharing from the research community, the issue of credit for good practice in data management and sharing is a fundamental one, with no easy answers. What steps can we take to make data sharing worth a researcher’s time and energy, and accelerate progress?

Encouraging data sharing: policy is not enough

"Funders are increasingly committed to coupling policy with practical support for researchers."

This year’s State of Open Data survey shows an interesting trend in more support from respondents for national mandates (63%) than in 2017 (55%). China’s Ministry of Science and Technology this year introduced their “Notice of the General Office of the State Council on the Measures for Managing the Printing and Distributing of Scientific Data”, which effectively mandates data sharing at a national level. The European Commission Horizon Europe proposal will mandate open access to research data as well as publications. Globally more than 50 funders now require data sharing, with the majority based in the US and Europe, particularly the UK. Yet in Springer Nature’s *Practical challenges for researchers in data sharing*, a survey of more than 7000 researchers, we found self-reported levels of sharing below the global average of 63% by respondents in the UK (58%) and US (55%)[9].

Funders are increasingly committed to coupling policy with practical support for researchers. To give just a few examples, The European Open Science Cloud and NIH Data Commons pilot have significant funding and infrastructure behind them, as do investments by Wellcome and UKRI/JISC. Of note for funders from this year’s State of Open Data survey is the marked increase in lack of certainty about where funds will come from to support making data open.

[9] *https://doi.org/10.6084/m9.figshare.5975011.v1*

Wanting the drive for data sharing to come from the research community itself has been another common thread in conversations with funders and foundations in the US, UK and Japan this year. Rather than a top-down, policy enforcement approach, many of the funders we have spoken with want the research community to create the momentum to share data, and help define discipline-appropriate ways of sharing.

Some institutions are also taking this "grassroots" approach. TU Delft provides one case study, presented at the LIBER[10] conference this year. Embedding "data stewards" in every faculty, to support researchers in good data practice, they also provide training, additional funding for data management and data publication, as well as a data repository via DANS[11]. TU Delft is now in the process of developing its research data policy, which will be adapted by each faculty based on disciplinary needs. This is a long-term investment in the "bottom-up" approach, and will be worth watching over the next few years.

> "Researchers would share data more routinely, and more openly, if they genuinely believed they would get proper credit for their work."

Finding the keys to grassroots support for data sharing

In a number of fields, data sharing is already the established norm, supported by community standards, dedicated repositories and long standing funder mandates. Yet in Springer Nature's "*Practical challenges for researchers in data sharing*" survey, we found that only 54% of respondents who produce specific biological and medical data (e.g. DNA and RNA sequences), are using existing dedicated community repositories to share their data. Making it easy to find out **where** to share data is clearly still important.

Responses to the question “Which one of the circumstances you chose would motivate you the most to share your data?” would suggest that visibility of research findings and the public good are the keys to making data sharing the status quo. “Funder requirements” were stated by just 69 respondents, ranking below (in order of popularity) increased visibility and impact, public benefit, transparency and reuse, journal and publisher requirements and getting proper credit. In my view, this masks the real issue. Researchers would share data more routinely, and more openly, if they genuinely believed they would get proper credit for their work that counted in advancing their academic standing and success in career development and grant applications, and for subsequent work that builds on their data. As noted in the analysis, “58% of respondents felt they did not get sufficient credit for sharing data, as opposed to 9% who felt they do.”

The 600+ free text responses to the question: “What credit mechanisms do you think would encourage more researchers to share their data?” warrant further analysis. Common themes from an initial review include citation, co-authorship and collaboration, and credit in research assessment.

"We should not ignore the barriers and challenges that researchers experience in sharing data."

We should not ignore the barriers and challenges that researchers experience in sharing data. Here this year’s State of Open Data survey adds some interesting insight to the body of research on this topic. The top six responses to “What problems/concerns do you have with sharing datasets?” were “Concerns about misuse of my data”, “Unsure about copyright and licensing”, “Not receiving appropriate credit or acknowledgement”, “Unsure I have the rights to share”, “Organising data in a presentable and useful way” and “Contains sensitive information”. All were selected by more than 400 respondents. To my knowledge, this is the first time concerns about misuse of data has come out so strongly in a global survey.

By contrast, in *Practical challenges for researchers in data sharing*, “Organising data in a presentable and useful way” was the most stated reason for not sharing data (46% of respondents). Other common challenges were: “Unsure about copyright and licensing” – 37%; “Not knowing which repository to use” – 33%; and “Lack of time to deposit data” – 26%.

With regard to short term actions, we need to better understand researchers concerns about “misuse of data” much better. Perhaps simpler to tackle is making sure that researchers are clear about their rights to share, and the copyright and licensing options available to them. Helping researchers to deposit, describe and share their data, using good metadata, remains a priority for Springer Nature.

Credit mechanisms of today: Data publication and data citation

To provide true credit for good data practice, published, citable datasets need to be viewed as research outputs on a par with a research article in terms of career advancement and assessment. Realistically, routine inclusion of datasets, their citations and impact in grant assessments and CV evaluation is probably still years away.

In the meantime, we can encourage and measure the usage and citations of datasets. Initiatives such as the GO FAIR[12] metrics group, the FAIRdat project from DANS and MakeDataCount[13] are making strides in this area. Figshare and other repositories include download and citation statistics, and alternative metrics for datasets. They also provide DOIs or other unique identifiers for datasets, ensuring they are citable in their own right.

> "There is compelling evidence as to the benefits of managing and sharing data, including productivity and citation advantages."

Encouraging and enabling data citations is also critical. As noted in the analysis of this year's survey, "data citations are motivating more respondents to make data openly available, increasing 7% from 2017 to 46%".

Here there are also encouraging community initiatives we should support. For example, DataCite[14] provides DOIs for research data, and provides a searchable registry of datasets, and a citation formatting reference tool. FORCE[15] continues to make progress in implementing its Data Citation Roadmap with publishers and other stakeholder groups. Publishers are increasingly providing links to datasets on articles, and including dataset citations in article metadata.

Data articles provide an established credit mechanism - a citable publication - while making datasets easier to find, access and reuse. Yet uptake of publishing data descriptors in data journals continues to be low. In this year's survey, 18% of respondents reported publishing data in a data journal, compared to 20% in 2017. These percentages are high compared to the global research community. The two largest dedicated data journals by volume are Elsevier's *Data in Brief* and Springer Nature's *Scientific Data*. Both have grown strongly in 2018, on track to publish close to 2000 and 300 articles respectively. Together, that's just 0.1% of the estimated 1.8 million articles published in English language journals annually. Perhaps there is more we can do here to make it easier for researchers to write and publish data articles, and see the benefits to their research in doing so.

We need to tell more stories about the benefits of data sharing

There is compelling evidence as to the benefits of managing and sharing data, including productivity and citation advantages. I referenced these in my contribution to last year's State of Open Data report. I still include them in almost every talk I give, because they continue to be "new news" when I share them, and not widely known. We need to continue to provide this evidence to the research community. We also need to do a much better job of finding and telling stories about researchers who are sharing data, the impact on their work and on the fields they work in.

Coupling these real world examples and evidence with better credit, clear funding, practical help and answers to common questions are all essential factors in accelerating data sharing to an established norm. There are no easy answers, and no "silver bullet", but there is much we can act on now.

[10] *https://libereurope.eu*

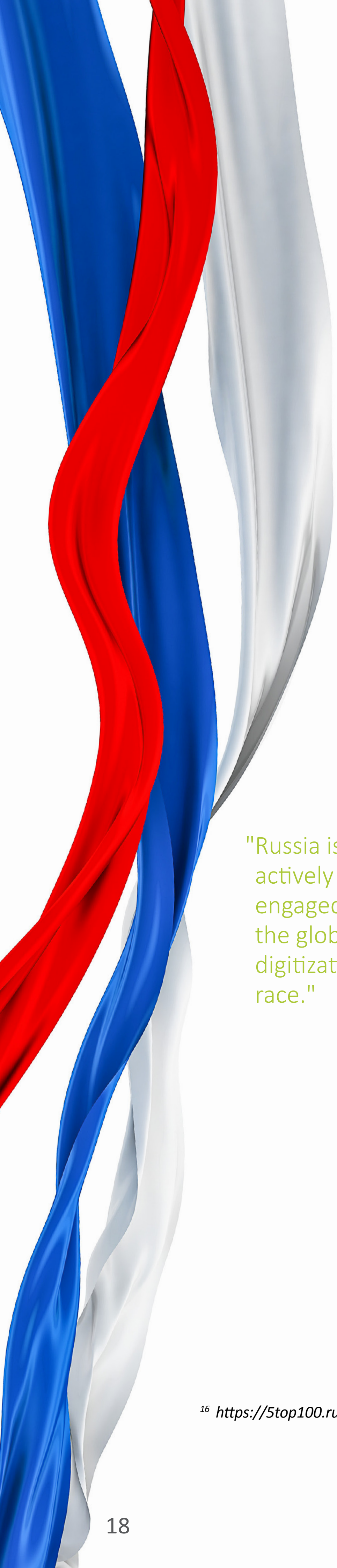
[11] *https://dans.knaw.nl/en&sa=D&ust=1538577624937000&usg=AFQjCNGUcGJun38lbLWOfWh3HQmJ0b_6Zg*

[12] *https://www.go-fair.org*

[13] *https://makedatacount.org*

[14] *https://www.datacite.org/*

[15] *https://www.force11.org/article/data-citation-roadmap-scientific-publishers&sa=D&ust=1538577624938000&usg=AFQjCNHwbNpiwU5gjTBnmhaSGXiTOiLjA*



Russia’s Move to Open Research

Pavel Arefiev, Principal Researcher, Scientific Electronic Library Ltd, Moscow, Russia

Igor Osipov, Far Eastern Federal University, Russia and CEO of Digital Science, Russia

Russia is actively engaged in the global digitization race – along with the EU, China and the US. This is manifested in nearly all aspects of national economy and public life, with massive volumes of open data and metadata now available on various state-funded portals. Elsewhere, science development and educational reform were made a priority in 2012. As a result of the new Digital Economy Strategy, research and education are now an important integrated element in the digitization paradigm.

Specific ideas and key models of open access to scientific publications came to Russia from the West in the late 90s. The first Russian open access journal, called "Investigated in Russia", was launched in 1998 by the Moscow Institute of Physics and Technology. The first significant database "Socionet" aggregated open access preprints and publications in 2000. The first open access university repository was opened in the computer network of the Ural State University (now Ural Federal University) in 2005. Nowadays, nearly all large universities have their own open access repository. According to the data from the Russian Index for Science Citation (the largest open access database of scientific publications in Russia) approximately 54% of all Russian scientific journals are open to the public as open access sources.

"Russia is actively engaged in the global digitization race."

The push for open data in Russia has been driven by Project 5-100. The goal of Project 5-100[16] is to maximize the competitive position of a group of leading Russian universities in the global research and education market. The project is focused on proactive internationalization of Russian science and education as well as rapid growth of relevant digital technologies.

The trend of mandating open access to research results is an opportunity for countries with open scientific policy and open scientific budget. This focus on long term improvements with a defined long-term strategy fits well with the open research agenda. Russia has a state-supported "national aggregator of Russian university Open Repositories" (NORA). The goal of the NORA project is to create a "single space for collecting information on the results of research by Russian scientists and providing access to materials published in the public domain" and several similar University-driven projects, which are based on university networks across the country.

The Russian government is still experimenting with the concept of open science, but the opportunity, progress thus far and obvious tie ins with the long term goals of the country suggest that open access and open data are firmly on the agenda. We will have to see how this will develop in the next few years.

[16] *https://5top100.ru/en/*

Contributor Biographies:

Barend Mons is a molecular biologist and biosemantics specialist at Leiden University Medical Center. He is known for innovations in scholarly collaboration, especially nanopublications and the FAIR data initiative.

Email: b.mons@lumc.nl
iD https://orcid.org/0000-0003-3934-0072

Erik Schultes is FAIR Data Scientific Projects Lead at the Dutch Techcentre for Life Sciences and at the Human Genetics Department at the Leiden University Medical Center. Erik's research interests focus on mapping the growth of biomedical knowledge and the development of tools that help biologists create, find, access, annotate, and share biomedical data.

Email: erik.schultes@go-fair.org
iD https://orcid.org/0000-0001-8888-635X

Luiz Olavo Bonino da Silva Santos is Lead Architect of the Linked FAIR data technology team, and International Technology Coordinator at the GO FAIR International Support and Coordination Office.

Email: luiz.bonino@go-fair.org
iD https://orcid.org/0000-0002-1164-1351

Ross Wilkinson is Director of Global Strategy at Australian Research Data Commons. Formerly, he was the Executive Director of the Australian National Data Service (ANDS), a program funded by the Australian Government to develop research data infrastructure and enable more effective use of Australia's research data assets. He is a Council Member of the Research Data Alliance, an international initiative aiming to build the social and technical bridges that enable better sharing of data.

Email: ross.wilkinson@ands.org.au
iD https://orcid.org/0000-0002-4192-1522

Grace Baynes is VP of Data & New Product Development for Open Research at Springer Nature. She is responsible for promoting open data and good research data practice; data publishing including the journal Scientific Data; data services; and new product development across open science and open research. Grace's passion for open science dates back to joining open access publisher BMC in 2003, and has flourished over 14 years at BMC, Nature Publishing Group and now Springer Nature.

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

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
iD http://orcid.org/0000-0003-4741-030

Pavel Arefiev is Principal Researcher at the Scientific Electronic Library Ltd in Moscow, Russia. He is responsible for analysis and evaluation of research outputs as well as developing academic careers in the universities. Pavel has been recognized as an expert in the areas of science policy and academic evaluation since he started work as the Principal Researcher in the national project “Russian Index for Science Citation” in 2005 and the Russian Academic Excellence University project 5-100-2020.

Email: arefiev64@gmail.com
iD https://orcid.org/0000-0003-3661-8497

Igor Osipov is CEO of Digital Science in Russia/CIS and VP Academic & Government EMEA, Digital Science UK. Having spent nearly 10 years in the Alaskan, Canadian and Russian Arctic as a student and researcher, Osipov held senior management positions in IT, STM Publishing, Consulting and Education. Most recently, he worked as Managing Director (Region) for Elsevier, working with funders, governments, and universities. He is Adjunct Professor at FEFU School of Economics and is an active member of the international research community, chairs and participates in various boards and working groups including FEFU Endowment Foundation, BMJ and UArctic Science Analytics.

Email: i.osipov@digital-science.com

Jon Treadway, Chief Operating Officer, Digital Science, London, UK.

Email: j.treadway@digital-science.com
iD http://orcid.org/0000-0001-9577-0283

Briony Fane, Data Analyst, Digital Science, London, UK.

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

Part of **DIGITAL**science

Altmetric

ANYWHERE ACCESS

BIORAFT

Dimensions

figshare

GRID

ifi CLAIMS

labguru

Overleaf

readcube

SYMPLECTIC

TETRASCIENCE

transcriptic

über RESEARCH

DIGITALscience Consultancy

digital-science.com