

## Digital Science Report

# The State of Open Data 2017

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

Foreword by Jean-Claude Burgelman

OCTOBER 2017

## About Figshare

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

## About Digital Science

**Digital Science** is a technology company serving the needs of scientific and research communities at key points along the full cycle of research. We invest in, nurture and support innovative businesses and technologies that make all parts of the research process more open, efficient and effective. We believe that together, we can change research for good. Visit [www.digital-science.com](http://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 Mathias Astell for assisting with the global distribution of the 2017 survey, and thanks to **Wiley**.

**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, The Campus, 4 Crinan Street, London N1 9XW, UK, [info@digital-science.com](mailto:info@digital-science.com)  
Figshare, 4 Crinan Street, London, N1 9XW, UK, [info@figshare.com](mailto:info@figshare.com)

Copyright © Digital Science and Figshare

# Contents

<b>1. Foreword</b> .....	<b>2</b>
Jean-Claude Burgelman, European Commission	
<b>2. Outputs Sharing: A Perspective from the Wellcome Trust</b> .....	<b>4</b>
Robert Kiley & David Carr, Wellcome Trust	
<b>3. Open Data - Sooner or Later? 2017 Survey Findings and Comparisons with 2016</b> .....	<b>7</b>
Jon Treadway & Briony Fane, Digital Science	
<b>4. 2 Out of 3 'Aint Bad. When Will Open Research Education Make it Through to Those Generating the Research?</b> .....	<b>12</b>
Mark Hahnel, Figshare	
<b>5. What Open Data Means for an Institution - The University of Cape Town</b> .....	<b>15</b>
Dale Peters, University of Cape Town	
<b>6. Collaboration and Concerted Action are Key to Making Open Data a Reality</b> .....	<b>17</b>
Grace Baynes, Springer Nature	
<b>7. Biographies</b> .....	<b>20</b>

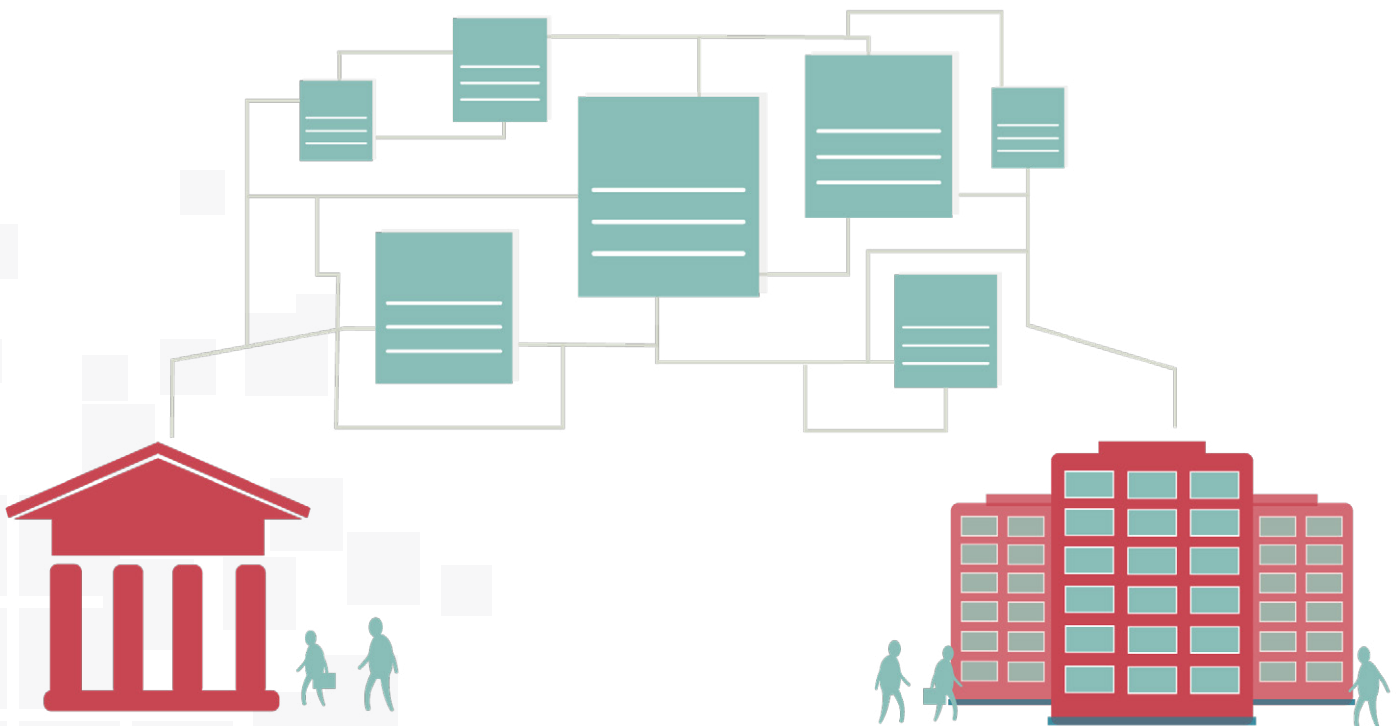
# Foreword

I am very pleased to introduce this year's State of Open Data Report. Since [Commissioner Moedas](#) took office and made open science one of his key priorities, open research data has become a centrepiece in research policy discussions.

Last year all European Commission Member States spoke out in favour of full open access to publications and data, reiterating the added value of the Horizon 2020 (H2020) [Open Research Data Pilot](#). From 2014 to 2016 a total of 68% of projects covered by the Pilot remained open access. This is encouraging and I am particularly pleased that outside the areas initially covered, 9% of projects voluntarily participated. There is real enthusiasm for research data sharing amongst researchers.

Therefore, the 2017 Open Research Data Pilot has been extended across the entire Horizon 2020 program: it makes open data the default in one of the biggest research funding programs in the world. Default opt-out possibilities still remain, however, and there are good reasons - IPR, personal data protection and national security. €30bn of research funding (the remainder of H2020) will produce open research data in the coming years.

Sir Nigel Shadbolt made an important argument in last year's State of Open Data Report. We should not assume that the open research data arguments have been won. One of the challenges is reluctance and, on occasion, skepticism amongst researchers, much of which stems from the fact that many scientists are still not well informed regarding open science (it is NOT free science), open research data and open access to publications. Communication and training efforts are so important.



We are in a situation analogous to the early 1980s when computing was introduced into labs and classrooms. It did not proceed 'sui generis', and skepticism for the large new machines did not fade even though clear benefits were demonstrated and research institutions set up dedicated training schemes. Here too, the soft "bottleneck" was the end user, in this case the researcher. Now again, much more needs to be done to provide our researchers with clear incentives for data sharing, most notably by inclusion in project and career evaluation. Open research data should pay off not just for science, but also for scientific careers.

The Open Research Data Pilot has shown us that making data open is an important component in the data ecosystem. In order to make data accessible, we also need our data to be findable, interoperable and reusable - the famous FAIR guiding principles. We are institutionalising this through Horizon 2020 data management guidelines, including a template for a FAIR data management plan. In doing so, we want to contribute to a truly European open research data culture as it is the DNA of the European Open Science Cloud and one of the center pieces of Commissioner Moedas' policy priorities.

It has often been said that 'data is the new oil'. This has been helpful in stressing the value of data in our economy, but it is the wrong analogy. Contrary to oil, open research data is non-exclusive and non-rival and so it makes sense to have publicly funded research data openly available. Open data is more like a renewable energy source: it can be reused without diminishing its original value, and reuse creates new value. The EC's policies on open research data simply guarantee as many scientists and innovators as possible make, under equal conditions, the best use of this renewable source.

**Jean-Claude Burgelman,**

Head of Unit Open Data Policies and Science Cloud,  
European Commission, Belgium

*"Open data is like a renewable energy source: it can be reused without diminishing its original value, and reuse creates new value."*



# Outputs Sharing: A Perspective from the Wellcome?

**Robert Kiley, Head of Open Research, and  
David Carr, Programme Manager, Wellcome Trust**

*"Making data available in a timely and responsible way ensures others can verify it, build upon it and use it to advance knowledge and make health improvements."*

## Introduction

The Wellcome Trust, in common with many other research funders, has introduced policies which seek to ensure that the outputs which arise from the research we fund – including publications, data, code and materials – can be accessed and used in ways that will maximise the resulting health benefits.

Making data available in a timely and responsible way ensures others can verify it, build upon it and use it to advance knowledge and make health improvements. The [sharing](#) of data during the Ebola outbreak ultimately helped to bring the epidemic under control, whilst the finalists in the [Open Science Prize](#) competition – which we ran in partnership with the National Institutes of Health and the Howard Hughes Medical Institute - demonstrated the huge potential for data to be reused to develop new applications and uncover new knowledge. Similarly, making software or research materials – cell lines, reagents etc. – available to the research community supports reproducibility and can underpin further research.

This short piece describes Wellcome's approach to sharing outputs, highlights some of the challenges we face and considers how these might be overcome.

## Outputs sharing policy

Earlier this year Wellcome rolled out a new policy on [data, software and materials management and sharing](#). Replacing our previous policy on data management and sharing, our new policy takes a more holistic view of outputs management and sets an expectation that researchers should maximise the availability of software and materials (as well as research data) with as few restrictions as possible.

As part of the policy, applicants for Wellcome funding are asked to provide an [outputs management plan](#) that sets out how they plan to manage and share significant data, software or materials to ensure the greatest benefit to health and research. We commit to review these plans – and the costs associated with them – as part of our funding decisions.

Crucially – and building on the experience of our open access to publications policy where we enjoy high levels of compliance – we also specify that we will actively monitor the implementation of the policy, including through a new check at final report stage to look at how the outputs from the grant have been made available. Although we have no plans to introduce sanctions (as we have done with our open access publications policy) at the moment, this is something we will keep under review.

## Open as possible, closed as necessary

In developing our outputs sharing policies and, in particular our requirements around data sharing, we have followed the principle that data should be “as open as possible, and as closed as necessary”.

A core element of our output sharing policy is that the data underlying research articles – which arise from our funding – must be made available to other researchers at the time of publication.

However, this does not mean that every piece of data underpinning a research article has to be on the open web, available to anyone and everyone. If data includes sensitive and personal information – eg, clinical trials – then these data need to be shared in a manner that protects the participants’ confidentiality, and respects the terms under which they consented to take part in the study. Normally, this would involve a managed access procedure – whereby requests to access the data are considered by an independent committee and agreements are put in place to specify how the data may be used.

### Key challenges

Although there are many challenges associated with the sharing of research outputs – equity, skills, and trust – arguably the two most significant ones relate to funding and researcher incentives.

#### Funding

We recognize that sharing outputs is not a cost-free activity. Data needs to be collected, curated and made accessible in standardized formats in order to be useful to the scientific community. In addition, we need to support the development of resources that enable researchers to deposit, find, discover, and use the data.

At Wellcome we have adopted a number of approaches to help address these issues. As discussed above, costs related to the implementation of output management plans are considered as part of the funding decision.

In addition, we support the development of resources to facilitate output sharing. Our [Biomedical Resource and Technology Development Grants](#) scheme supports researchers who want to establish or maintain resources or technology for the benefit of the wider scientific community. We are working in partnership with other funders to support the development of the [ClinicalStudyDataRequest.com](#) platform to provide managed access to clinical trial data, and are actively considering establishing a Wellcome-branded repository to host data generated by our researchers in cases where no suitable community resource exists.

Ensuring data resources are sustainable remains a key challenge, especially if the ambition is to ensure these resources continue to be made freely available to both data depositors and users. There are no easy answers here, but Wellcome is working in partnership with a number of other life sciences funders to explore long term solutions. A preprint outlining the issues and an approach for addressing these is available [here](#).

*"Data should be “as open as possible, and as closed as necessary”."*

*"We recognise that sharing outputs is not a cost-free activity."*





*"There is a need to support initiatives that encourage standardised practices for citing data and other outputs, which allows downstream reuse to be acknowledged and tracked."*

### **Incentives**

A [survey](#) in 2016 of over 500 Wellcome funded researchers showed that although about 50% of respondents make their data available for use by others, relatively few report any direct benefits from sharing their data. Furthermore, many researchers are concerned about the possibility of misuse or misinterpretation of their data, loss of publication opportunities, and the effort required to prepare and deposit data.

Addressing these concerns is something that funders are seeking to address. Small steps – such as committing to check at the end of a grant that outputs have been made available – sends a message to the community that this is something we are serious about. Equally, ensuring that grant application forms encourage researchers to describe outputs other than journal articles and that reviewers/panel members take account of these outputs – is something we are championing. Specifically, panel members are reminded of Wellcome's commitment to the [San Francisco Declaration on Research Assessment](#) (DORA) and asked to focus on the "content and quality of publications, rather than their number, the venue of publication, or the impact factors of the journals in which they were published" and to take into account "the diverse range of possible research outputs".

Looking more broadly, there is a need to support initiatives that encourage standardized practices for citing data and other outputs, which allows downstream reuse to be acknowledged and tracked. The adoption of other tools, such as badges for open practices, are also worthy of further exploration.

### **Conclusion**

Our vision at Wellcome is a world where there are transformative improvements in human health because research outputs are managed, shared and used in ways that unleash their full value. To realize this we need to ensure that researchers are equipped, empowered and motivated to make their outputs findable, accessible, interoperable and reusable and to use these outputs to accelerate research and improve health.

We still have a long way to go to there, but the momentum towards open research is growing. Wellcome is committed to working with our research communities, partner funders, journals and other stakeholders to ensure we can unlock the full value of research outputs.



# Open Data - Sooner or Later? 2017 Survey Findings and Comparisons with 2016

**Jon Treadway, Chief Operating Officer, and  
Briony Fane, Data Analyst, Digital Science**

Full survey data and questionnaire can be found at [https://figshare.com/articles/State\\_of\\_Open\\_Data\\_survey\\_2017/5480710](https://figshare.com/articles/State_of_Open_Data_survey_2017/5480710). An interactive visualization of all the data can be found at <https://knowledge.figshare.com/articles/item/state-of-open-data-2017>

Last year, to mark Open Access Week and to try and garner some detailed insight into attitudes and experiences of researchers working with open data - sharing it, reusing it, redistributing it - Figshare released a report collating findings of an extensive survey it had undertaken in partnership with Springer Nature. (<https://doi.org/10.6084/m9.figshare.4036398.v1>). Entitled 'Open Season for Open Data', we presented a whole range of data that showed open data was already a reality, and that while researchers were unsure and lacked confidence on some specifics and particulars, we could see indications that the future would likely be more open.

We expressed the hope that we could use the survey to track the evolution of how researchers deal with data in future years. One year on, we have begun to do just that.

We saw a marked growth in respondents from just over 2000 to almost 2300, in part due to the willingness of Springer Nature to ensure large numbers of researchers were invited to respond. When researchers were happy to disclose where they were based, we saw a strong growth in responses from researchers based in Asia (20% up to 29%) with increases from Africa (1% to 6%) and South America (5% to 8%) which are welcome, but sample size is still too small to draw specific conclusions. We also saw a growth in the percentage of older researchers responding, from 8% to 14%.

*"We can see strong signals that open data is becoming more embedded; the trends are positive."*

## **We can see strong signals that open data is becoming more embedded; the trends are positive:**

- Respondents have become more aware of open data sets (82% up from 73%) than in 2016. Age does not appear to be a major factor in this trend - younger researchers (25-34 year olds) showing no larger increase (75% to 85%) when compared with older age groups, notably 55-64 year olds (up from 70% to 80%).
- Willingness of researchers to reuse open data sets in their own research has grown by a similar amount, a 10% increase to 80%, with the increase again replicated across age groups.

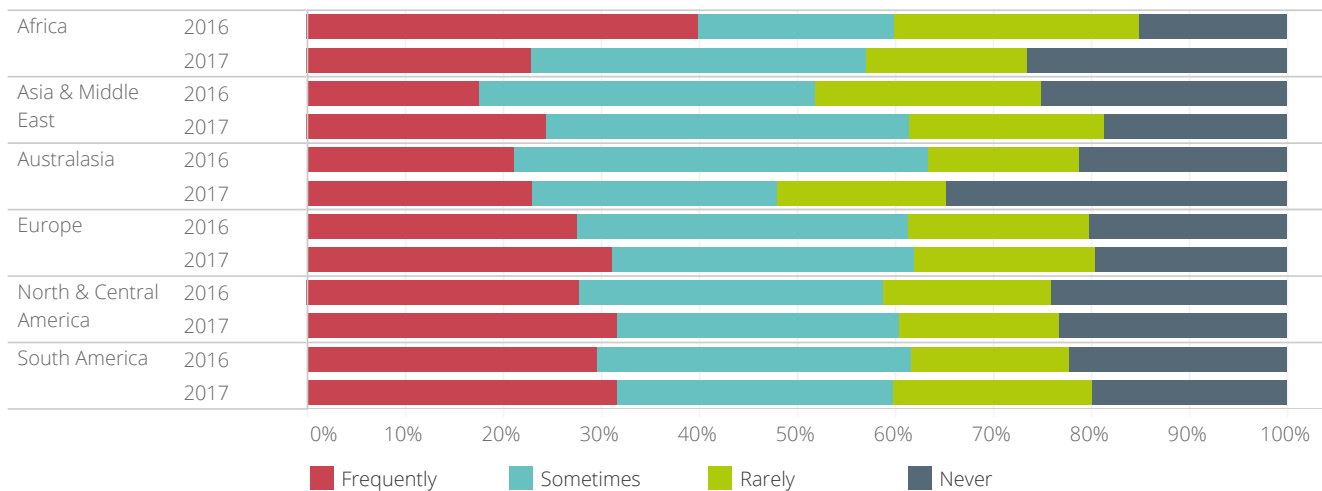


Figure 1 - Frequency with which researchers share data

- Those researchers who routinely share their data (either frequently or sometimes) has also grown since 2016, although by a smaller amount, from 57% to 60%. The proportion of researchers who have never made a data set openly available has reduced by 3%, from 24% to 21% (see Fig. 1). Looking deeper we can see further promise for the future of open data, as 70% of these researchers are now willing to reuse open data sets in their own research (up from 65%), and 63% are aware of open data sets (up from 51%).
- More researchers are curating their data for sharing. We saw an increase from 67% in 2016 to 74% in 2017.
- 29% of respondents who frequently share data do not know where funds to action this were coming from - a result entirely consistent between 2016 & 2017, but positively those who share data less regularly saw a reduction from 43% to 38%. Those who are not aware of data sets are also more likely to know who would pay for it, down from 54% to 46%.

**Digging deeper, we can see shifting patterns in the debate on open data, and a richer picture of behaviors:**

- When asked where they have published data, most commonly respondents had done so as an appendix to an article (just over 30%) with a data repository close behind (just under 30%) and 20% having published in a data journal.
- The use of open data sets was also revealing, with large numbers of respondents motivated by a desire to validate their own results (>50%), to avoid duplication, or because it complemented their own data (50%). The fostering of collaboration was another common motivation, with 35% of respondents citing it.
- 36% of respondents have lost data on which they were working and there is, unsurprisingly, a high correlation between the vehicle for storing data and where it was lost - computer hard drives were the most common culprit here.
- Fewer respondents are in favour of national mandates for open data, with notable growth in those neutral to the possibility (24% up from 8%), with this change visible across the regions and age groups.
- There is no change in the high number of researchers valuing a data citation the same as an article - from 78% in 2016 to 77% in 2017.

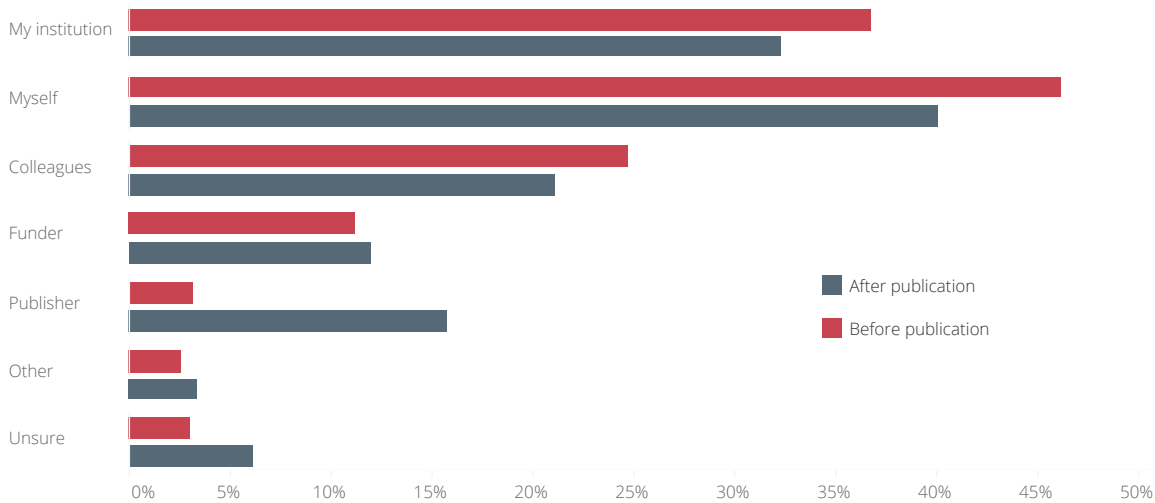


Figure 2 - Ownership of data before and after publication

**There is still a lack of confidence around open data, but again here we are seeing a shifting picture rather than any one clear trend:**

- Respondents who are extremely or very confident in citing secondary research data and referencing secondary research data has seen a decline from 42% to 36%. Confidence referencing secondary data sets similarly reduced from 40% to 34%.
- We have seen reductions in those who say they don't know whether they are required to make data open by different stakeholder mandates - those uncertain about funder mandates reduced from 25% down to 19%, those uncertain about institutional mandates was down from 20% to 17%, and with publisher mandates the figure was down from 31% to 24%.

*"There is still a lack of confidence around open data."*

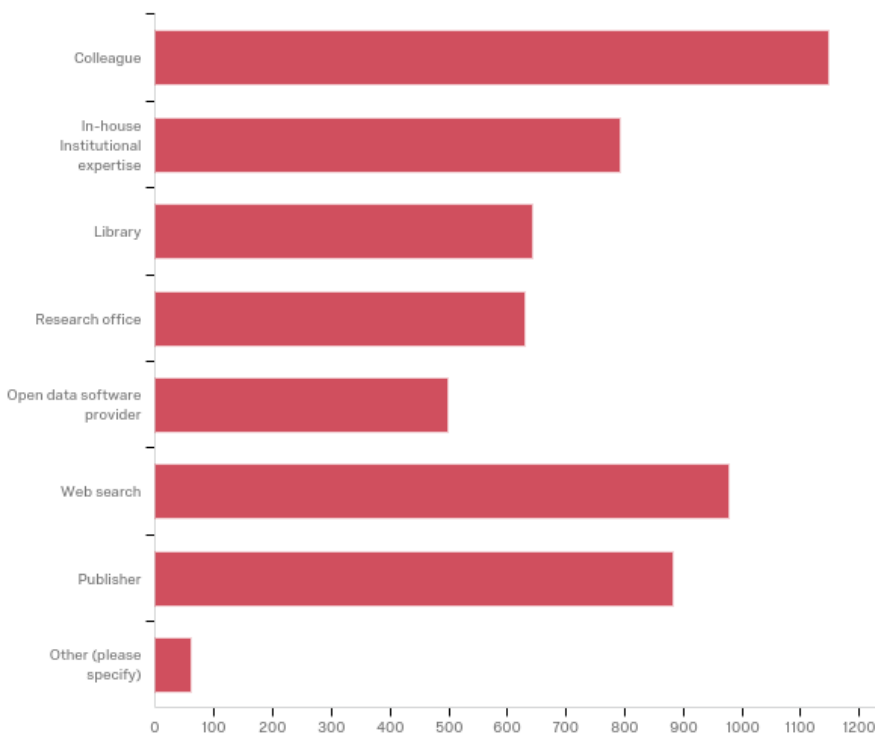
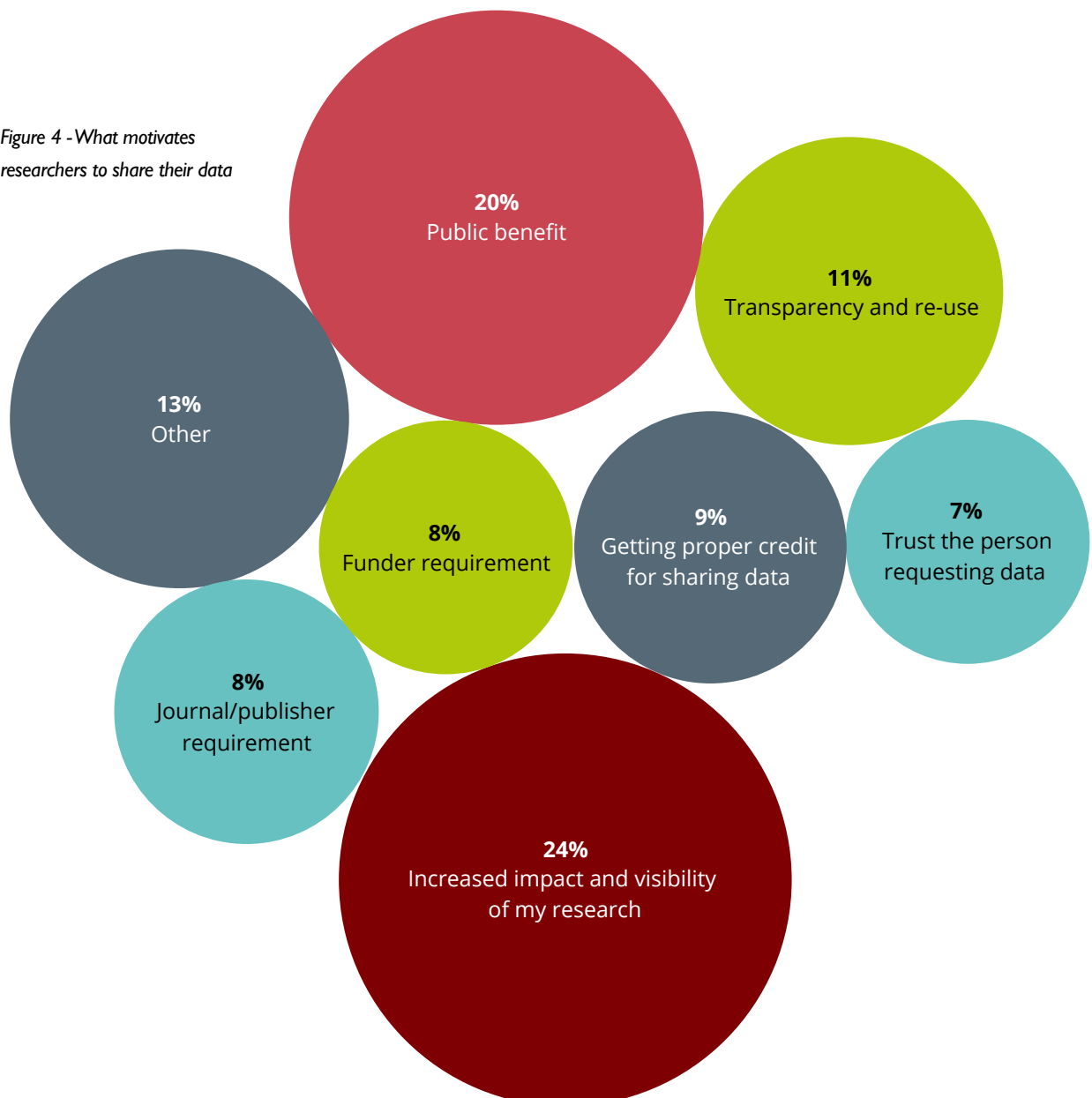


Figure 3 - Sources researchers look to for support with open data

- This appears to be driven by a growth in awareness in respondents that, in many cases, those stakeholders do NOT have such mandates - funders lacking a mandate is up to 48% from 40%, institutions without mandates is up to 58% from 53%, for publishers the figure is 44% up from 33%.
- While many researchers are still not clear under which licenses they have made research openly available, there has been a marked reduction. 37% of our respondents report that they are unsure about what licence covers the data they share compared to 64% last year.
- 49% of respondents think they own the data they produce before publication, with 40% believing they own it *after* publication (see Fig. 2).
- Most commonly, respondents look for support on open data via a web search (32%) with publishers the next most common result (29%) ahead of libraries (21%) and research offices (20%) (see Fig. 3).

Figure 4 - What motivates researchers to share their data



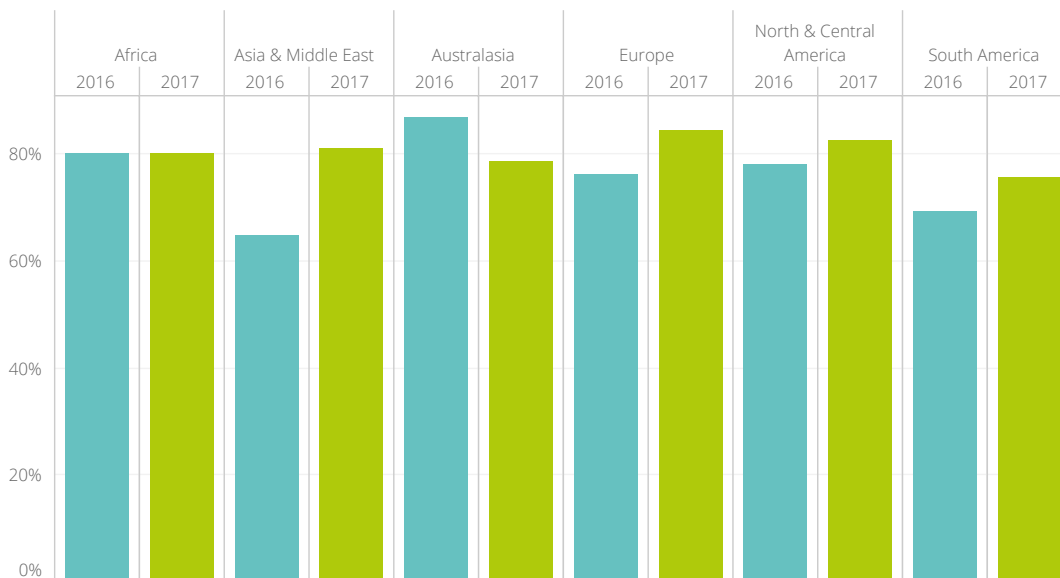


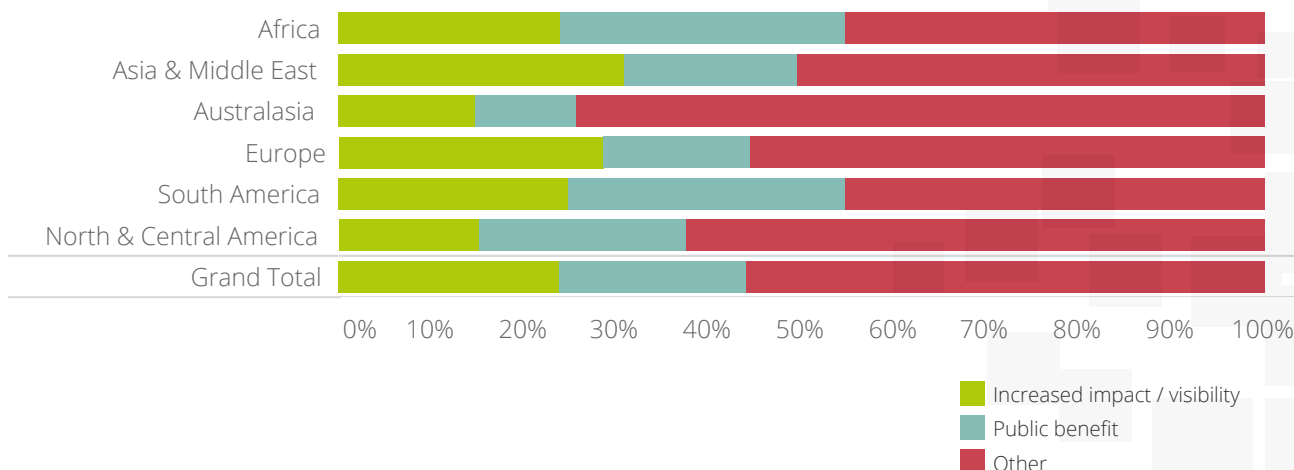
Figure 5 - Researchers' awareness of open data

**It is also worth noting that while trends are generally positive, key indicators are disproportionately positive in Asia, along with some other striking results:**

- Awareness of open data sets has grown in respondents from Asia, from 65% up to 80% - an increase of 15% compared to 9% globally (see Fig. 5).
- Respondents based in Asia who have shared open data sets have risen 10% to 62%, compared to 3% globally.
- Respondents based in Asia who are neutral about a national mandate for open data have grown markedly, up 22% to 29% of respondents, compared with a 16% rise globally.
- Achieving impact is a much bigger motivator for Asian researchers when sharing data openly - 30% cite this compared with 15% in North America, where public benefit is the main factor. (see Fig. 6).

There are more findings to be uncovered, more suggestive data points to be scrutinized, and we would invite the community to help us extend the analysis presented here. The data from the 2017 survey is available on Figshare at <https://doi.org/10.6084/m9.figshare.5480710>.

Figure 6 - Motivations for sharing data



## 2 Out of 3 'Aint Bad.

# When Will Open Research Education Make it Through to Those Generating the Research?

**Mark Hahnel, Founder and CEO, Figshare**

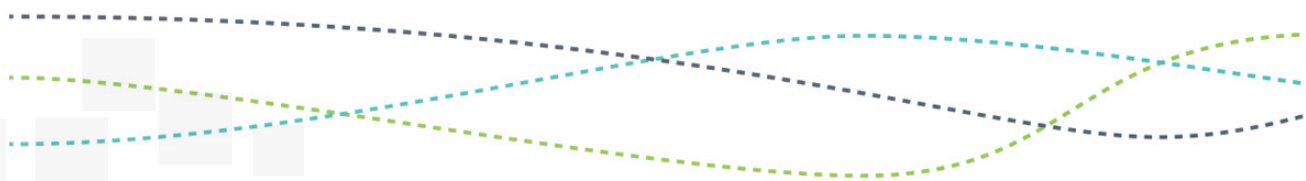
### *'Academics want to be famous!'*

*"Citations to these outputs are growing year on year."*

Not in the traditional sense perhaps - academia has more than its fair share of introverts - but to progress their career, or win Nobel Prizes, an academic's name and work needs to be well known in the community. This is traditionally achieved by publishing novel research in high impact journals. However, evidence in the 2017 State of Open Data report continues to suggest that academics are happy to get their credit wherever they can.

For the second year running, more than 75% of researchers surveyed stated that they value a citation to their non-traditional research outputs (NTROs) as much as, if not more, than to a traditional output. This is consistent with indications that outputs other than publications and their impact will be rewarded at an [equal level in funding decisions](#).

This year Figshare added citation information for every DOI that is minted across the system, whether it's on an institutionally branded repository we support, or on figshare.com. What we are finding is that citations to these outputs are growing year on year. We're also seeing a disproportionate amount of citations for code or software, an area that the traditional academic publishing systems have struggled to provide a solution that adequately distributes that much sought after credit. Being the first system to do this means that we are just scratching the surface on citation trends around NTROs.





This growth in the incentivization of researchers through credit is the first of three big trends that sum up the landscape around NTROs globally. The number of funder policies grows, along with suggestions about how to enforce compliance – yet the majority of researchers still don't think they have a publisher, funder or institutional mandate to share data.

Lack of researcher knowledge is the second trend. Institutions continue to hire research data librarians, big publishers are employing data curators, and yet the majority of researchers still are unclear about licensing requirements. If we're looking for an acceleration in open research globally, and what will drive it, it seems like there is still a lot of potential in the stick-led compliance approach, as policies and mandates proliferate and grow in precision across the world.

The third big trend in the space has been the buzz around preprints. With [Physics](#), [Chemistry](#) and [Biology](#) all having strong community-driven solutions, the concept of open access to all research outputs looks ever more likely. A rebrand of the institutional repository to the institutional preprint server may encourage compliance with open access mandates in a way that incentivizes the researchers. This all then becomes an infrastructure issue, one that is at least technically resolvable.

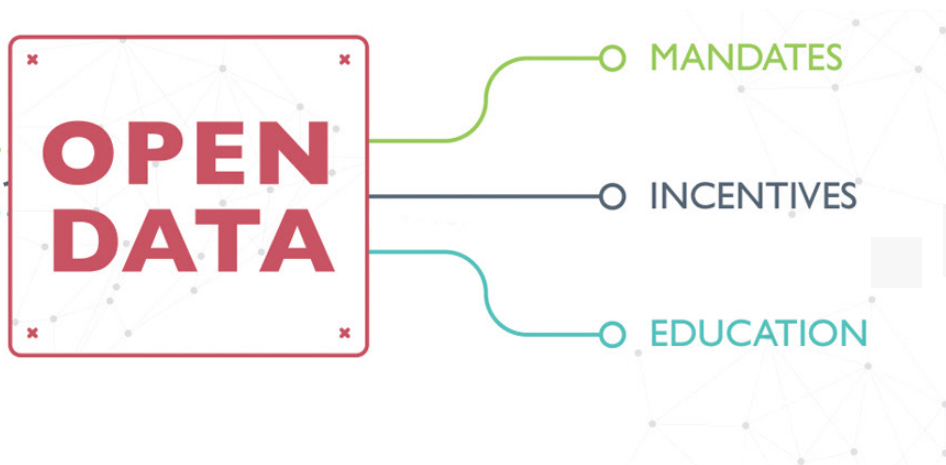
The State of Open Data is really The State of Open Academic Research Outputs, but that isn't quite as catchy. Herein lies opportunities. The [FAIR](#) principles that have been lauded as the Shangri-La for all academic infrastructure can also be applied to open papers. All digital files, including preprints, or papers, should be thought of as 'data' in this respect.

*"The majority of researchers still don't think they have a publisher, funder or institutional mandate to share data."*

Our internal discussions put the general state of affairs to be consistent with:

- data that is FAIR for humans
- data that is FA for machines

With all of the above considered, the Figshare team has come up with a set of guiding principles that can be adopted by publishers, funders and institutions as we work towards a FAIR-er future:



*"2017 has been good for bringing open data into the mainstream. Infrastructure and credit problems are on their way to being solved."*

- Academic research outputs should be as open as possible, and as closed as necessary
- Academic research outputs should never be behind a paywall
- Academic research outputs should be human and machine readable/query-able
- Academic infrastructure should be interchangeable
- Academic researchers should never have to put the same information into multiple systems at the same institution
- Identifiers for everything
- The impact of research is independent of the type of output and where it is published

### **So where will the next 12 months take us?**

It seems that a 'happy' mixture of carrots, sticks and education is needed to move academia forward, faster. Each stakeholder has their own responsibilities. Carrots provided by funders and infrastructure providers, sticks evolving with the growing number of mandates. Perhaps most critical is education, both getting the message to academics and providing curation expertise. Here is the biggest unknown - where will the education come from? I believe we'll see moves from universities and publishers. The balance of how content is disseminated and who ultimately gets credit for all these new citations, be it academics, publishers or institutions will be decided by how much resource and effort can be thrown into the education vertical.

2017 has been good for bringing open data into the mainstream. Infrastructure and credit problems are on their way to being solved. I for one hope that in 2018 we can say the same about education at every level of the academic pyramid, with a global focus on making things FAIR becoming our moon landing – a noble, ambitious target to aim for.

# What Open Data Means For an Institution - The University of Cape Town

**Dr Dale Peters, Director, UCT eResearch**

## **Introduction**

Amidst a growing number of mandates for data sharing and reuse, South African universities are scrambling to provide services to the research community to facilitate compliance with the complex data management requirements of numerous international funding agencies. The impending burden posed on institutions to fund this requirement is now brought sharply into focus with the introduction of a similar mandate by the National Research Fund (NRF), the very lifeblood on which the national academic enterprise is reliant.

At a time when nationwide student protest action has highlighted the financial constraints on the higher education sector, the need to act collaboratively is imperative to sustain levels of excellence. The award of a substantial grant to the University of Cape Town (UCT) as the lead institution on a regional consortium to establish and host a regional data node in the National Integrated Cyberinfrastructure System (NICIS) signals a new way of thinking about collaborative systems and services, based on open data principles, that marks a clear divergence from institutional infrastructure development in the past.

## **Institutional responses to Open Data**

An extensive evaluation process of data repository platforms was conducted by UCT eResearch, aimed at provisioning an effective research data service based on emerging standards and best practice. In comparing open source and licensed options, and taking into consideration the cost of infrastructure support staff, Figshare was identified as the most appropriate solution. The concurrent development of an institutional research data management policy together with an extensive advocacy programme gave substance to the evaluation, in the valuable feedback from the research community. The point was made that acceptance of the open data principle is less about funder compliance, and more about individual agency.

This is a valiant manifestation of academic freedom at UCT, a research-intensive university where scholarly communication- and more specifically data publication – has become the currency of the research enterprise, supported in a national subsidy system based on publication counts.

The research community acknowledge the societal benefits of open data, in driving greater scientific integrity, enabling a strategic response to societal challenges. They see the value of collaborative research and wider dissemination of their outcomes – but they insist on the right to make the decision whether to share openly.

*"The research community acknowledge the societal benefits of open data, in driving greater scientific integrity, enabling a strategic response to societal challenges."*

It is not surprising therefore, that the functionality provided by Figshare to separate data upload from data publication has found wide appeal. It is also not surprising that UCT has developed specific terms of deposit that recognize the responsibility of the researcher to determine the necessary limits on openness particularly relating to personal information and commercial considerations.

### **Conclusions**

The outcomes of the advocacy programme conducted at UCT suggest that while the research community fully support global infrastructure developments, many questions remain regarding data sharing and reuse. Primarily, the demand is for greater support for both staff and students to better understand and use open data to strengthen research practice as a whole.



# Collaboration and Concerted Action are Key to Making Open Data a Reality

**Grace Baynes, Director, Data & New Product Development, Open Research Group, Springer Nature**

The case for good research practice and open data to research outputs is increasingly inarguable. Open access to research data can help speed the pace of advancing discovery and deliver more value by enabling reuse and reducing duplication. Good data practice also makes research more efficient, effective and fulfilling for researchers. As the data in this survey show, the research community recognize the value of open data, yet good data practice and data sharing are still far from the status quo.

Springer Nature and its publications have been advocating for good data practice for over a decade. Recent efforts have focussed on growing data publishing options to provide credit, and strengthening and simplifying our data policies. Our future focus is on support and incentives to enable data sharing, data management and open data, built in collaboration with the research community.

## **The case for data**

The argument for better data practice is made stronger by global concerns about reproducibility and research integrity, reducing fraud and improving patient outcomes. As much as 50% of preclinical research done in the US, at a cost of US\$56.4B a year, cannot be reproduced, estimates a 2015 study<sup>1</sup>. In the same year, a *Nature* survey found that 70% of over 1,500 respondents had tried and failed to replicate the work of others. More shocking was that 50% of respondents had failed to reproduce their own work<sup>2</sup>. There is evidence that data availability increases reproducibility, as reported in a review of *Nature Genetics* papers<sup>3</sup> and elsewhere.

There is also a proven productivity benefit to good data practice. Data archiving can double the publication output of research projects, according to a study of 7,000 National Science Foundation and National Institutes of Health funded research projects in social sciences<sup>4</sup>. Citation impact of research papers has also been shown to increase when data are made available – by as much as 50% in astrophysics<sup>5</sup>, and between 9-35% in gene expression microarrays<sup>6</sup>, astronomy<sup>7</sup> and paleoceanography<sup>8</sup>.

The data in this survey show that researchers are using others' research data (49%), or would be willing to do so (80%). Yet only 60% of respondents make their data openly available "frequently" or "sometimes". The most common ways of sharing data are still supplementary information in a journal article or peer-to-peer. Perhaps more concerning is data storage and data

*"The argument for better data practice is made stronger by global concerns about reproducibility and research integrity, reducing fraud and improving patient outcomes."*

*"The results of this survey would suggest that funder mandates are not a key motivator for open data."*

*"More than 50 funders now mandate or encourage data sharing, compared to 28 in 2015."*

management. Only 20% of respondents had prepared a data management plan, and the most common ways to store active and archived data were personal hard drives, external hard drives and institutional servers.

Researchers are intelligent, responsible, motivated people. They are also time-poor, and do not necessarily want to become data or licensing experts. So they need clear information, simple policies and advice. They also understandably prioritize advancing their field, their own research and building their careers. So they need tools to make data sharing and management easier, and credit and incentives to make good research data practice and open data worthwhile.

To effect change, government, funders, institutions, libraries, publishers and researchers themselves all have a role to play. Here are areas this survey has prompted us to think more about:

### **The role of government**

It is interesting to see the support for national mandates for open data in this survey (55% of respondents). Many countries have now made government data open, providing the best use cases to date for economic and social impact of open data. When it comes to research data, national approaches and infrastructures will continue to need similar long-term commitment, and to be balanced with fostering international collaboration, including through global discipline-specific data repositories.

### **The role of the funder**

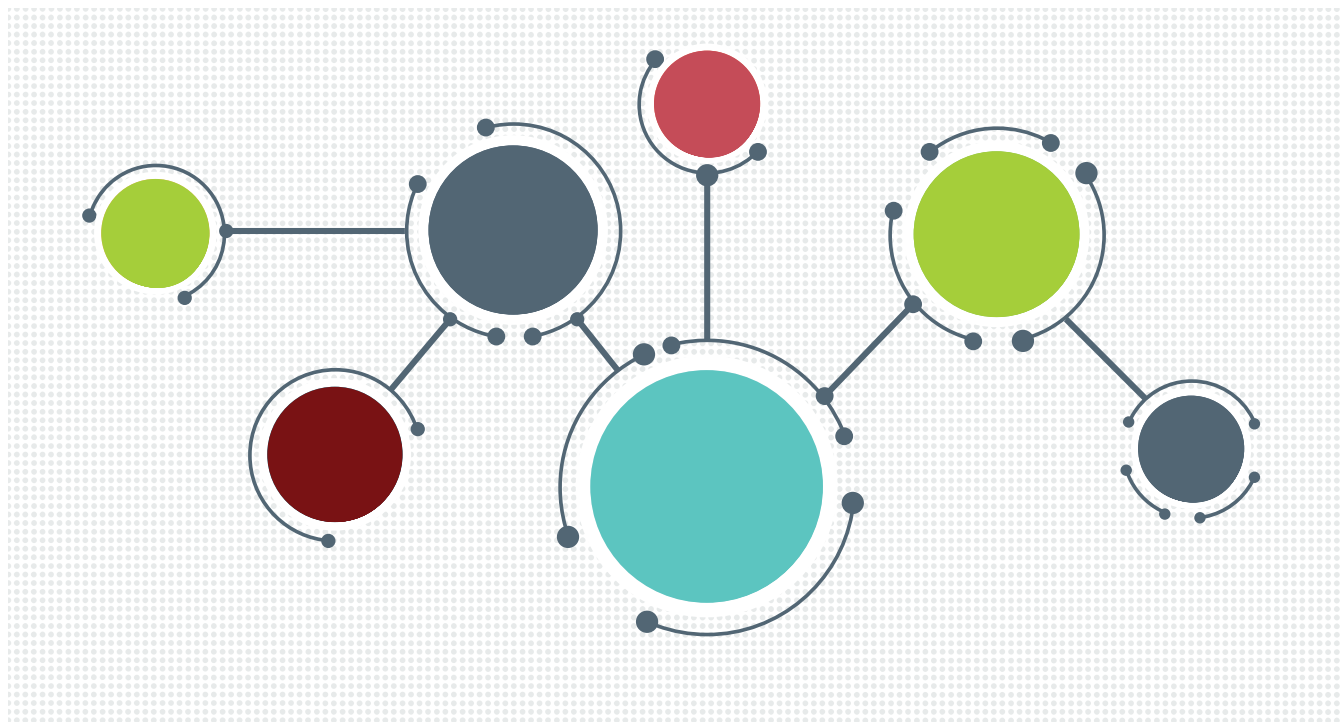
The results of this survey would suggest that funder mandates are not a key motivator for open data. This contradicts the findings of other studies<sup>9</sup>, and is contrary to what we see as funders' crucial role in effecting change. The growth of open access publishing was driven in part by funders issuing clear and specific mandates, explicitly making funds available and making compliance a requirement. Springer Nature tracks funder policies on data to help provide advice to authors on compliance. Encouragingly, more than 50 funders now mandate or encourage data sharing, compared to 28 in 2015. As yet, only a few funders have requirements for data management plans or data availability statements, or explicitly make funding available for data management, storage and curation.

### **The role of the institution**

Institutions and libraries have a key role to play in supporting researchers: helping them understand and comply with funder requirements, training, establishing local research data management solutions and support where needed<sup>10</sup>. Partnering with data initiatives, repositories and other useful parties, including publishers, will help reduce potential duplication of effort and ensure sustainability.

*"Concerted efforts by governments, funders, research institutions, publishers and researchers themselves are needed to make widespread open data a reality."*





### The role of the publisher

Publishers work closely with researchers at many stages of the research process, particularly when they are writing up and sharing their findings. Here are five actions publishers can take:

- 1) Continue to advocate for good data practice across different communities.
- 2) Encourage good research data practice and open data through journal policies and author information: see for example Springer Nature's standardized research data policies, Research Data Support Helpdesk and recommended repositories list.<sup>11</sup>
- 3) Provide credit mechanisms for good data management and open data: through data publishing, registered reports, data citation and linking, and new mechanisms such as badges for open practices.
- 4) Offer solutions to help researchers share their own data, and discover and use data: for example our pilot Data Support Services, which help researchers deposit and curate data, in partnership with Figshare.<sup>12</sup>
- 5) Partner with the research community to build shared solutions: for example, the global Research Data Alliance (RDA) interest group to improve research data policy<sup>13</sup> standards, data linking and citation.

A number of other publishers including PLOS, Wiley and Elsevier are also taking some or all of these steps.

Concerted efforts by governments, funders, research institutions, publishers and researchers themselves are needed to make widespread open data a reality, and make research data management the new normal. Collaboration and partnerships between these groups will make that happen faster, and more effectively. Springer Nature looks forward to further playing its part.

- 1 Freedman, L. P., Cockburn, I. M. & Simcoe, T. S. *PLoS Biol.* 13, e1002165 (2015) <http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002165>
- 2 Baker (2015) <http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970>
- 3 Ioannidis et al (2009) <https://www.nature.com/ng/journal/v41/n2/full/ng.295.html>
- 4 Pienta et al (2010) <https://deepblue.lib.umich.edu/handle/2027.42/78307>
- 5 Dorch et al (2015) <https://arxiv.org/abs/1511.02512>
- 6 Piwowar & Vision (2013) <https://doi.org/10.7717/peerj.175>
- 7 Henneken & Accomazzi (2011) <https://arxiv.org/abs/1111.3618>
- 8 Sears et al (2011) [https://figshare.com/articles/Data\\_Sharing\\_Effect\\_on\\_Article\\_Citation\\_Rate\\_in\\_Paleoceanography/1222998/1](https://figshare.com/articles/Data_Sharing_Effect_on_Article_Citation_Rate_in_Paleoceanography/1222998/1)
- 9 Schmidt et al. (2016). *PLoS ONE* 11(1): e0146695. doi:10.1371/journal.pone.0146699
- 10 <http://www.oclc.org/research/publications/2017/oclcresearch-research-data-management.html>
- 11 <https://www.springernature.com/gp/authors/research-data-policy>
- 12 <https://www.springernature.com/gp/authors/research-data-policy/support-services/12327144>
- 13 <https://www.rd-alliance.org/groups/data-policy-standardisation-and-implementation>

# Contributor Biographies:

**Grace Baynes** is Director of Data & New Product Development for Open Research at the publisher 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 in BMC in 2003, and has flourished over 14 years at BMC, Nature Publishing Group and now Springer Nature.

Email: [g.baynes@nature.com](mailto:g.baynes@nature.com)

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

**J C Burgelman** is Head of Unit Open Data and Science Cloud Policy. He joined the European Commission in 1999 as a Visiting Scientist in the Joint Research Centre (the Institute of Prospective Technological Studies - IPTS), where he became Head of the Information Society Unit in 2005. In January 2008, he moved to the Bureau of European Policy Advisers (attached to the president of the EC) as adviser for innovation policy. Since 1-10-2008, he joined DG RTD, as advisor and then Head of Unit in charge of top level advisory boards like the European Research and Innovation Area Board, the Innovation for Growth Group and the European Forum for Forward Looking Activities. Till 2000 he was full professor of communication technology policy at the Free University of Brussels, as well as director of the Centre for Studies on Media, Information Telecommunication and was involved in science and technology assessment. He has been visiting professor at the University of Antwerp, the European College of Bruges and the University of South Africa and sits on several academic journals. He chaired the World Economic Forum's Global Agenda Council on Innovation and was a member of its Science Advisory Committee.

Email: [Jean-Claude.BURGELMAN@ec.europa.eu](mailto:Jean-Claude.BURGELMAN@ec.europa.eu)

**David Carr** is Programme Manager for Open Research at the Wellcome Trust, where he is responsible for developing and taking forward a range of activities to maximise the availability and reuse of research outputs – including publications, datasets, software and materials. Previously, David worked as a Policy Adviser at Wellcome – leading on work to develop and communicate policy in several areas – including data sharing, open access publishing, biosecurity and genomics. In 2001, David worked on secondment at the World Health Organisation in Geneva, where he assisted in the preparation of the Advisory Committee on Health Research (ACHR) report on Genomics and World Health. Prior to joining the Trust in 1999, David worked as a project researcher at a scientific consultancy firm in Cambridge. He has undergraduate and master's degrees in genetics from the University of Cambridge.

Email: [d.carr@wellcome.ac.uk](mailto:d.carr@wellcome.ac.uk)

 <http://orcid.org/0000-0003-1435-307X>

**Mark Hahnel** is founder and CEO of Figshare, London. He is passionate about open science and the potential it has to revolutionize the research community. Figshare is looking to become the place where all academics make their research openly available, as well as producing a secure cloud based storage space for their outputs. By encouraging 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.

Email: [Mark@figshare.com](mailto:Mark@figshare.com)

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

**Robert Kiley** is Head of Open Research at the Wellcome Trust where he is responsible for developing and implementing an open research strategy. The overarching aim is to help ensure that researchers are equipped, empowered and motivated to make research outputs findable, accessible and usable – and to use these outputs to accelerate research and improve health. Over the past decade Robert has played a leading role in the implementation of Wellcome's open access policy and overseeing the development of the Europe PubMed Central repository. Robert also acts as the point of contact for [eLife](#), the open-access research journal, launched in 2012 with the support of the Howard Hughes Medical Institute, the Max Planck Society and the [Wellcome Trust](#), and more recently championed the work to create a new open publishing platform for Wellcome researchers – Wellcome Open Research. Robert is a qualified librarian and an Associate Member of CILIP. Robert also sits as a Director on the [ORCID](#) Board.

Email: [r.kiley@wellcome.ac.uk](mailto:r.kiley@wellcome.ac.uk)

 <http://orcid.org/0000-0003-4733-2558>

**Dale Peters** is Director of eResearch at the University of Cape Town. In this role she provides leadership in engaging networked information technologies to enhance and support innovative practice in scientific research; promoting multi-institutional collaboration and trans-and interdisciplinary research in the deployment of research data infrastructures.

Email: [dale.peters@uct.ac.za](mailto:dale.peters@uct.ac.za)

 <http://orcid.org/0000-0002-4160-3874>

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

Email: [j.treadway@digital-science.com](mailto:j.treadway@digital-science.com)

 <http://orcid.org/0000-0001-9577-0283>

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

Email: [b.fane@digital-science.com](mailto:b.fane@digital-science.com)

 <http://orcid.org/0000-0001-6639-7598>



**DIGITAL**  
science

Part of the Digital Science family



**BIOR**AFT



**GRID** 



**labguru**



**Peerwith**



 **SYMPLECTIC**

