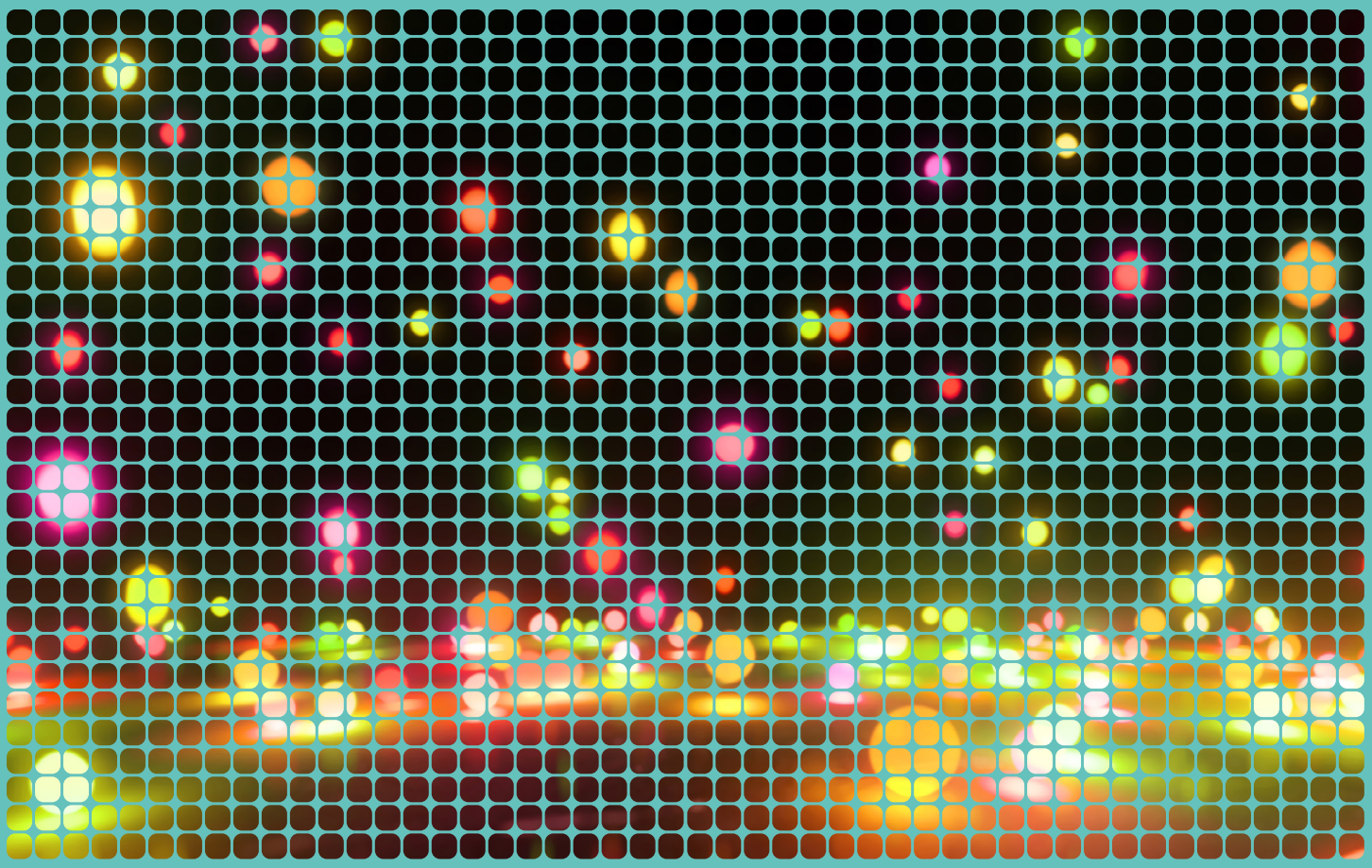


Digital Science Report

The State of Open Data

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

Foreword by Professor Sir Nigel Shadbolt

OCTOBER 2016

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 serving the needs of research. At the centre of our mission is the support of researchers within research institutions, funding bodies, publishers and governments, for whom we provide a range of software, content and consultancy solutions. We believe passionately that tomorrow's research will be different and better than today's.
Visit www.digital-science.com

Acknowledgements

Figshare and **Digital Science** are grateful to **Springer Nature**, in particular Dan Penny and Anna Gallagher for their expertise in assisting with the compilation of the survey and distributing it globally.

Figshare and **Digital Science** are also grateful to all the contributors for their thought leadership pieces included in this report.

This report has been published by Digital Science, which is operated by global media company, the Holtzbrinck Publishing Group.

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

DOI: https://dx.doi.org/10.6084/m9.figshare.4036398
ISBN: 978-0-9956245-1-1

Contents

Foreword

"We need look no further than the success of the human genome, released as open data for all, to understand how innovation can flourish around open research output."

I am delighted to introduce ***The State of Open Data*** report. As a long time advocate for open data I have always regarded the participation of the research sector as fundamentally important. The real prize for society is not simply producing open data but facilitating open innovation. Open data enables a situation where the collective genius of thousands of researchers produces insights and analyses, inventions and understanding beyond what isolated individuals with their silos of data could produce. We need look no further than the success of the human genome, released as open data for all, to understand how innovation can flourish around open research output. Open innovation needs open data, open standards, open licences and open participation to really flourish.

We all have our favourite examples of the power of open data and this report provides more from which to choose. However, it is important to understand how the research community views the opportunities and challenges of open data. This report highlights the extent of awareness around open data, the incentives around its use and the perspectives that researchers have about making their own research data open. We must appreciate that researchers do need time to reflect on their results, to determine if they are at all reliable and to be able to extract the value in the data they have laboured hard to produce. We must also have consistent policies that determine how data is to be made openly available if it is funded by the public purse and if there are no significant issues that ought to preclude its release.

If we are to really change custom, practice and culture in the research sector we do have to recognize that we need incentives as well as mandates. Increasingly, data is deposited along with the textual narrative of the research. Increasingly, that data is referenced, acknowledged and cited. Increasingly, data is the principal resource that drives aspects of the research agenda. The very availability of some data makes certain investigations possible at all.

"If we are to really change custom, practice and culture in the research sector we do have to recognize that we need incentives as well as mandates."

The call for governments and business, public and private organizations to open up data has been a refrain for some years. League tables of the comparative performance of countries are now available. However, we should not assume that doing well in these exercises means that the arguments have been won. It is essential that we highlight the benefits of open data policy. It is important that we understand the best way to promote and incentivise the production and consumption of open data. It is crucial that consistent policy is developed between research funders and where possible jurisdictions. It is vital that best practice is shared and that the data on open data made widely available. I therefore welcome this report and commend it to you.

Sir Nigel Shadbolt FREng
Chairman and Co-Founder Open Data Institute
Principal of Jesus College & Professor of Computer Science
University of Oxford

Open by Default?

Dr Mark Hahnel, CEO, Figshare, London
& Dr Daniel Hook, Managing Director, Digital Science, London

Recently, at a Digital Science Spotlight event in Denver, the subject of discussion that evening was open data. One of the panelists spoke in an impassioned way about the practical issues that face every researcher in making their data open. It was not the usual talk of ethical concerns, confusion over licensing or whether their contractual arrangements either demanded or indeed forbade the release of their data, but rather the career implications of their choice to make their hard-won data openly available.

"Opening up research and research data promises a plethora of benefits."

On the one hand, modern research is a highly collaborative endeavor, in which many minds work to solve complex problems requiring diverse skills and expertise. In such an environment, it is almost impossible to believe that you are the only person or the only group of researchers capable of addressing your problem; or that you are the best people to address the problem in all of the facets that are required to develop the problem further. Logically, you should share because that's the best way to develop a solution.

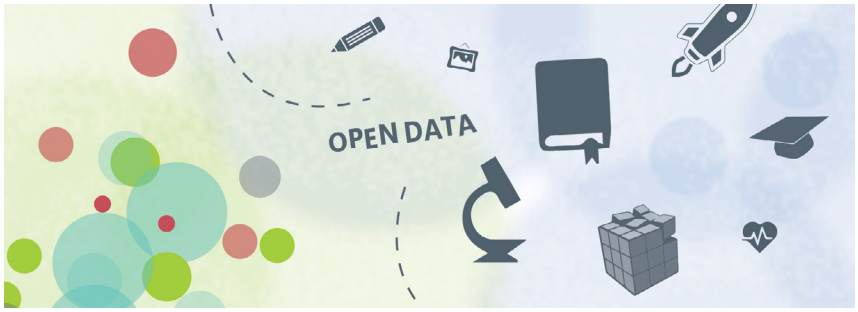
On the other hand, there is an uncomfortable reality – if everyone shares their data but you don't share your data, you have an advantage. You are more likely to be successful in the next funding round due to that advantage – after all, you will be standing on the shoulders of more giants!

The rise of the open data movement represents just one developing aspect of the modern research environment, but, it is the component that highlights some of the greatest sociological problems that we need to deal with in the sector. We are not the first, nor will we be the last to point out that our academic system is fundamentally broken – publication in particular journals assures funding and job prospects; no-one becomes a professor for sharing their data.

Nevertheless, opening up research and research data promises a plethora of benefits, not only to the funders of research, but also to society at large. As these benefits become more tangible and funders face increasing pressure to demonstrate return on public money invested, it's clear that the academic landscape is being driven to change to support openness. However, not all areas of the academic landscape are evolving equally rapidly.

Academics will soon have to share all the digital outputs of their publicly-funded research: their lab data, computer code, survey data – everything that's ethically appropriate. How are these changes going to affect research going forward? Do researchers have concerns?

At Figshare, we speak to academics on a daily basis and we believe that we have a good sense of the concerns of our users. Of course, different users in different subjects and different geographies have different concerns. But broadly, we can classify concerns into the following two major categories:



1. Structural – *Do I have to make my data open? Am I allowed to? How do I do it? What does my data need to look like for me to share it? Doesn't this mean even more work for me without recognition? What happens if I make the wrong data available – am I going to get sued?*

2. Cultural – *If I make my data open, couldn't someone progress the research at a faster rate? Would that mean that I'm helping others to progress their careers and win grant funding at my expense? If I make my data open, couldn't someone question my analysis or conclusions or more easily detect errors?*

Of course, as humans some of these fears are understandable. But, as humanity, some of these fears raise serious concerns. We don't have the space to discuss these issues in detail in a short article like this one, but it is important to recognize the situation.

Here we focus on the practical and ask: What can be done in the short-term to change the status quo?

Changing the infrastructure

In our survey, more than half the respondents and 62% of early career researchers said that they would welcome more guidance on compliance with their funder's policy. Given the complexity of the landscape with the interplay of institutional, contractual and funder policies uncertainty is understandable, but more and more help is at hand.

The role of the librarian is now much more multifaceted than it has been in the past. Over the last decade the role has transitioned from the traditional role of classifying, locating and accessing content, to a role that some refer to as the 'information professional' - assisting research staff to better manage and to better disseminate their content and, in doing so, to fulfill funder expectations. Librarians have become an indispensable source of knowledge around all things to do with data, code and policy; cementing their role at the heart of the research institution as key facilitators of the research process.

Beyond the compliance landscape there are technical problems that also need to be addressed. However, before there is a simple recipe to provide to researchers with what to share and how to share it, we need to revisit what it means to write a paper in modern research. The 'standard' for a paper has moved on considerably in the last 20 years – gone are the days when papers are static, printed artifacts: journals require supporting data and computer code (sometimes even logical copies of whole computers) in order to ensure reproducibility; they increasingly allow video and sound content that they render in an online experience that constitutes an evolution to the scholarly

record unlike anything that we've seen in the last 350 years. Journals themselves are fighting for relevance as a way to organize research output. One of the next walls that is likely to fall is the idea that a paper has a definite publication date and only a single 'version of record' exists rather than a constantly developing narrative as data and analysis are added to a corpus of work with many contributors. In this world data, although significantly different from different areas, needs to achieve a level of homogeneity that experts can come back to it in the future and continue to find meaning in it.

So how can we ensure the appropriate, useful and legal release of files? Ideally, we would have an army of data curators ready to go, but no such army exists, and such an option doesn't scale very well when we haven't yet established accepted community standards on what constitutes good quality peer review of data.

Over the last five years the team at Figshare has been thinking about how to meet the challenges that this new environment brings. Topics that keep coming up in conversation are: metadata, curation and peer review.

"Librarians have become an indispensable source of knowledge around all things to do with data, code and policy; cementing their role at the heart of the research institution as key facilitators of the research process."

There is no simple solution to gather quality metadata around research outputs, or to provide context in terms of quality. However, we believe that this needs to be a multi-step process.

Some of the key points of interaction with the user are:

- At point of file save
  - User generated metadata
  - Automated acquisition of provenance data – filetype, machine parameters
- At point of public release
- Additional specialist curation post 'publication'
- Automated curation through linked open data

Each of these opportunities adds different metadata that can serve different purposes. If well-structured and with appropriate context included, these additions can be tremendously valuable in downstream activities such as discovery or computer-aided inference. Figshare is actively building or enhancing functionality at each of these steps in the research publication process, from the creation of custom metadata schemas in our institutional edition, to collaborations such as *Link it up*[1] and novel user-developed functionality based on the Figshare API.

The limiting factors in the progression of the use of open data are principally in the quality of the description and metadata surrounding the data. Once these challenges have been addressed it is key to ensure that data can 'move' Open data can only be as powerful as the flexibility of the APIs in the software that helps digital files persist. As pointed out in the recent Digital Science white paper entitled, 'A New 'Research Data Mechanics'[2], in the short term, allowing files, metadata and identifiers to flow between institutional systems is an achievable goal and one that has already been identified by several universities. In the longer term, we aim to harness the power of APIs to query datasets in the browser, allowing researchers to build on work gone before without needing to download and parse open and accessible data.

"For those who embrace openness, there should be a tangible reward and there should be recognition of positive contributions to the new research paradigm."

The Data FAIRport[3] initiative has set up 'Guiding Principles' for FAIR data publishing, focusing on principles for Findability, Accessibility, Interoperability and Reusability. The final aim is to create minimal models for grouping results and linking data with analytics which are both human and computer actionable. By adhering to these principles, we aim to make the content on Figshare available to any computer or human searching for academic data through any system.

Changing the culture

As data and metadata become structured using the infrastructure described above and the context provided by linked open data, it is possible to utilize these structures to become more efficient. As a result of this underlying framework the web itself will evolve to return more accurate data in response to any question that is posed.

As the world's largest driver of knowledge, the Academy has a responsibility to lead by providing data to better answer queries at all stages of the learning and educational process. At the current rate of progress, by 2020 all of the developed world's research funding bodies will require openness by default.

In this vein, the biggest advancement in 2016 has been the announcement by the European Commission that all papers and digital objects created as a result of their funding need to be, 'as open as possible, as closed as necessary'. In terms of technology to support this, there are several existing solutions that fit the bill. In terms of outreach, learning and awareness of these changes, it seems as though there is going to be a very steep learning curve, with some real consequences for those who ignore the advice of their libraries and research offices.

Social change is always difficult, especially if it has been ingrained for decades, job security and career advancement are coupled to support the status quo in such a way as to make the behavior structural. However, change is afoot. Slowly but surely the research establishment is being driven to reassess its attitudes.

Even if the only argument on offer was to improve the efficiency of research, it would be difficult to deny that openness is a good thing given how research is now done. In this respect, even doing as little as openly sharing negative results would have a profound impact on the research that needs to be performed. But, this isn't the only argument: at the time of writing retractionwatch.com listed 684 articles just from PubMed tracked sources in 2015, a 37% rise over 2014. This represents a tiny percentage of the articles that were published in 2015 but it does make it difficult to sympathize with those who seek to protect their career progression over the central reason that most people go into science – namely, to progress the sum of human understanding.

There should be a fundamental social responsibility associated with being open in the same way that there is now a social responsibility associated with recycling your garbage or paying your taxes. For those who embrace openness, there should be a tangible reward and there should be recognition of positive contributions to the new research paradigm. Without these drivers, research will remain broken with openness being at odds with funding practice. Institutions have it within their power to make these changes…the tools are there.

1 Link it up: http://linkitup.data2semantics.org/
2 A New 'Research Data Mechanics: https://figshare.com/articles/Digital_Science_White_Paper_A_New_Research_Data_Mechanics_/3514859
3 Data FAIRport: http://www.datafairport.org/Mechanics_/3514859

Why Open Data Now? Big Data, Knowledge Production and the Political Economy of Research

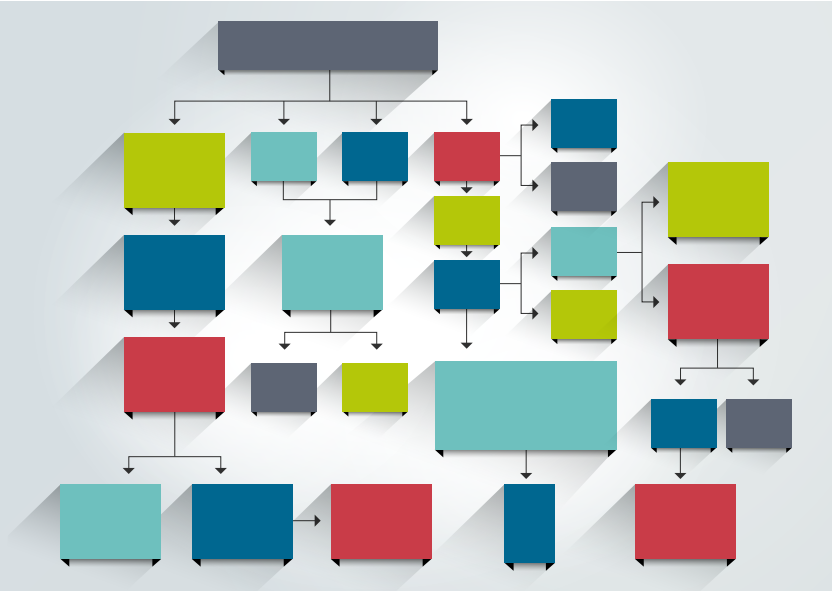
Dr Sabina Leonelli, Associate Professor, Department of Sociology, Philosophy and Anthropology, University of Exeter, UK

When data is used as evidence for scientific claims, the data becomes public objects, which should be widely scrutinized to assess the validity of the inferences drawn. Yet, the vast majority of scientific data generated in the 20th century have only been accessed by small groups of experts; and few of those data, selected in relation to the inferences made, have been made publicly available in scientific journals. This is tied to a view of scientific knowledge production as an esoteric and technical process, where even trained researchers become so specialized as to be unable to assess data produced by fields other than their own. Scientists invest time and effort in scrutinizing data only when they have reason to doubt their colleagues' interpretation or suspect foul play; and concerns with data production and interpretation, including issues associated with the emergence of 'big data', remain remote from global civil society.

Since the start of the new millennium, the open data movement has challenged this technocratic way of sharing data and their political, social and economic significance. The movement brings together scientists, policy-makers, publishers, industry representatives and members of civil society around the globe who believe that data produced by scientific research should be made publicly accessible online and freely accessible for reuse by anyone. The Internet provides a platform for scientists to exchange data, materials and opinions in real-time, no matter where they are geographically located. Participants in the open data movement embrace this opportunity. They typically advocate that data can, and should, travel beyond the specific setting in which they are generated, thus enhancing the possibility that people who have not been involved in their production will contribute to their interpretation. Accordingly, the production of scientific knowledge is portrayed as involving the centralized collection and 'mining' of datasets gathered by different research communities across the globe. Pooling together results, it is argued, maximizes the chances of identifying significant patterns in the data that are collected, and of transforming data into knowledge. This in turn may improve the quality, accessibility and transparency of research and speed up the rate of scientific discovery.

Whether research is actually being driven by data, rather than theories, hypotheses, models or policy challenges, remains disputable. What is clear, is that data are increasingly conceptualized as inherently valuable products of scientific research, rather than as components of the research process

"They typically advocate that data can, and should, travel beyond the specific setting in which they are generated, thus enhancing the possibility that people who have not been involved in their production will contribute to their interpretation."



which have no value in themselves. This involves viewing data as open to several possible interpretations, whose validity and usefulness depend on the questions, interests and materials characterizing the specific context in which data are adopted. It also involves viewing data as research outputs that can be published and cited without necessarily having been used as evidence for a specific claim (as required within traditional journal publications).

Over the last decade, funding agencies such as the National Institute of Health, National Science Foundation, European Research Council and Research Councils UK have endorsed this innovative perception of how data should be managed. They are actively promoting open data as key to the advancement of basic research and its translation into applications with immediate social impact, such as therapeutic or agricultural innovations. They are pressuring their grantees to release data to public databases – a move that affects how scientists set up their research, and measure and develop their outputs. Many researchers now invest considerable time and resources into donating data to public repositories; and regard the consultation of online databases as a first step towards the development of new lines of inquiry.

Why the open data movement has acquired such prominence in contemporary scientific and public discourse is an important question. Given the enormous achievements of 20th century science, where data sharing was confined to small sections of the (predominantly Western) scientific community, why are funding bodies insisting on open data as crucial to 21st century research? A standard answer to this question points to open data as a crucial way for scientists to exploit the emergence of new technologies, such as genome sequencing and Internet-based social media. It is true that the availability and widespread uptake of new information and communication technologies, as well as the introduction of new methods of data generation, play a crucial role in making it possible to produce and share information on the scale advocated by the open data movement. And yet, the emergence and political impact of the open data movement are not mere consequences of technological advances in data production and communication, nor are their implications restricted solely to science.

Scientific concerns underlying the open data movement need to be evaluated in relation to four other sets of factors.

1. Open data provides a common platform for scientists, scientific institutions and funders (in both the private and the public sphere) to discuss and tackle the practical difficulties involved in making data travel and be re-used. Whether scientific data are shared, among whom and to what effect, depends on the existence of appropriate regulatory, social and material infrastructures, such as workable databases, guidelines on data donation, and servers in safe locations where data storage can be guaranteed in the long term; as well as well-coordinated networks of individuals, scientific groups, companies and institutions that take responsibility for developing, financing and enforcing those infrastructures and the related instruments, computers and software. The resources and skills required to achieve such coordination are clearly not only technical, but also social.

2. Open data feeds into concerns with transparency, legitimacy and return on investment on the part of science policy and funding bodies. Public institutions responsible for science funding are under pressure from national and international policy. They have an interest in fostering public trust in science as a source of reliable knowledge and thus as a legitimate source of information. Perhaps the most blatant recent case of public mistrust in science is the controversy following the public release of emails exchanged by researchers at the Climatic Research Unit of the University of East Anglia in 2010 (an episode often referred to as ClimateGate). This was a case where a perceived lack of transparency in how climate data were handled fuelled social mistrust in the scientific consensus on global warming. This in turn affected public support for the implementation of international measures against climate change. Many national governments and international organizations like the European Research Council support the free circulation of data in the hope that it will increase the transparency and accountability of scientific research - and, potentially, its trustworthiness and social legitimacy. Similarly, the Royal Society has pointed to open data as an opportunity to prevent scientific fraud and disclose the evidence base for scientific pronouncements to the general public, so as to avoid the kind of miscommunication and misunderstanding underlying ClimateGate.

3. Open data aligns with the challenges posed by the globalization of science to new parts of the world, beyond traditional centres of Euro-American power. Open data are implicated in transforming the geographies of science and its relation to local economies, as illustrated by the rise of centres of research excellence in the global South. Institutes such as the Beijing Genomics Institute, interact with researchers across the globe largely through digital means, and do not see themselves as requiring the support of extensive local or even national research infrastructure and traditions. Thanks to widespread data dissemination over the Internet, they can quickly learn from results produced elsewhere and contribute their own share of data to international databases and research projects, thus gaining visibility and competing with established programmes in the United States, Japan and Europe. Nations that have not figured as prominent producers of scientific knowledge throughout the 20th century, such as China, South Africa, India and

Singapore, are devoting increasing financial support to research, in the hope of attracting a highly skilled workforce to boost their industrial productivity and economic prospects.

One might think that laboratories in poor or underfunded regions would strongly support data sharing, for it makes data produced with expensive technology accessible to them, raising their chance to produce cutting-edge science; and that rich laboratories, which regard the possession of such technologies as providing them with a competitive edge, would be reluctant to donate data – particularly since donation requires additional labour. However, taking account of the considerable resources and diverse expertise needed to transform data into new knowledge helps to acquire a more realistic view on the benefits and costs of data sharing. Underfunded laboratories actually struggle to access online resources, appropriate bandwidth, adequate expertise and computers powerful enough to analyze data found online; and are coming to terms with the difficulties involved in developing resources and standards for data donation. By contrast, many rich laboratories have found that data donation offers the opportunity to participate in international networks and receive help with data analysis, thus accruing their own prestige, visibility and productivity. Even major pharmaceutical companies like GlaxoSmithKline and Syngenta are contributing to the development of public databases, in the hope of outsourcing their R&D efforts, improving their public image and gaining from the availability of data produced through public funding.

4. Open data exemplifies the embedding of scientific research in market logics and contexts. To make it at all feasible for data to travel, market structures and political institutions need to assess not only their scientific value, but also their value as political, financial and social objects. The increased mobility of data is unavoidably tied to their commodification. The very idea of scientific data as artefacts that can be traded, circulated across the globe and re-used to create new forms of value is indissolubly tied to market logics, with data figuring as objects of market exchange. National governments and industries that have invested heavily in data production – through the financing of clinical trials or genome sequencing projects – are keen to see results. This requirement to maximize returns from past investments, and the urgency typically attached to it, fuels the emphasis on data needing to travel widely and fast to create knowledge that would positively impact human health.

Further, the open dissemination and reuse of data not only challenges notions of competition and property within established scientific communities, but also notions of property, privacy and effective communication in industry, government and civil society.

Data acquired from patients in clinical trials or participants in personalized genomics, for instance, have clear economic value, and some companies welcome the opportunity to access personal information unwittingly circulated by citizens who are not aware of its value as 'data' for medical research – a move widely disputed by legal scholars, advocacy groups and medical associations as an infringement of privacy. The dissemination of data of relevance to innovation in food security or bioenergy, such as molecular data on plants and plant pathogens, is similarly plagued by uncertainties about intellectual property, particularly in cases of public-private partnerships between governmental agencies and companies such as Monsanto or Shell.

Rajan[1] (2006) and Kelty[2] (2008) have shown how free data access has greatly helped to maximize exchange and downstream capital flows. At the same time, data mobility is not free in the sense of being devoid of financial and social costs. Data sharing requires human resources and capital: even the most successful initiatives are confronted with the exponential costs involved in maintaining and expanding data infrastructures in the long-term, and are struggling to produce sustainable business plans for their activities. Indeed, the European Union has denounced the costs associated with funding the current plurality of online databases in biomedicine as unsustainable in the long term, and is pushing for the centralization of facilities for data sharing as a possible solution (most prominently through ELIXIR, a gigantic effort currently underway at the European Bioinformatics Institute to coordinate and eventually integrate data sharing initiatives in biology and medicine; see the ELIXIR website: http://www.elixir-europe.org/). The National Science Foundation, which funded many successful data sharing initiatives at the turn of the millennium, is also attempting to rationalize its investments in this area and is now asking database curators to provide self-sustaining business models.

A critical assessment of the significance of the open data movement for contemporary society at large needs to take account of all these factors, which foreground the indissoluble ties of scientific research to global political economy. The emergence of technologies and related expertise that facilitate the production and dissemination of biological data on a large scale is certainly a key reason for the visibility and political support garnered by the open data movement in recent years. In turn, the development of technologies and expertise for the care of data, not to mention their production and use to create new biomedical knowledge and interventions, is made possible by the availability of institutions that help to define the financial value of data as commodities and the conditions under which data can be made to travel around the globe. What has propelled data into becoming protagonists of contemporary biomedicine is their ambiguous status as at once local and global, free commodities and strategic investments, common goods and grounds for competition, potential evidence and meaningless information. Openness, defined through the opportunities for dissemination associated with the Internet, is a defining characteristic of 'big data' science, policy and infrastructure. The vision underlying the open data movement is that data risk to remain meaningless if they are prevented from travelling far and wide, and that travel endows data with multiple forms of scientific as well as financial, social and political value.

"The vision underlying the open data movement is that data risk to remain meaningless if they are prevented from travelling far and wide."

Article adapted from: Leonelli, S. (2013) Why the Current Insistence on Open Access to Scientific Data? Big Data, Knowledge Production and the Political Economy of Contemporary Biology. Bulletin of Science, Technology and Society 33(1/2): 6-11.
http://dx.doi.org/10.1177/0270467613496768

1 Sunder Rajan, Kaushik. 2006. Biocapital: The Constitution of Post-Genomic Life. Durham, NC: Duke University Press.
2 Kelty, Christopher M. 2008. Two Bits: The Cultural Significance of Free Software. Duke University Press.

Open Season for Open Data: A Survey of Researchers

Dr Briony Fane, Metrics Researcher, & Jon Treadway, Director Operational Strategy, Digital Science
Anna Gallagher, Research Analyst, & Dan Penny, Head of Market Intelligence: Researchers and Audience, Springer Nature
Dr Mark Hahnel, CEO, Figshare

Figshare has garnered many insights from its users in the past, from formal surveys and informal feedback. However, these have been directed toward the working of its product, rather than the state of open data in the research sector as a whole.

Working with Springer Nature and Digital Science, we surveyed researchers about their attitude and experiences in working with data, sharing it and making it open.

The response rate was very strong, surpassing our best expectations; over 2,000 researchers responded to the survey, spread across continents and disciplines, from all types of institution and researchers at different career stages.

The survey reveals a number of compelling insights, some of which we present here - it is pleasing to note that the sector is far from procrastinating when it comes to open data.

1. For the majority of respondents, open data is already a reality

Researchers are aware of open data - Approximately three quarters of respondents are aware of data sets that are open to access, reuse, repurpose and redistribute (Figure 1).

Figure 1 - Awareness of data that is free to access, reuse, repurpose and redistribute, n=1915

Researchers in the social sciences demonstrate the highest level of awareness by subject area (Figure 2), while, by geography, researchers in Asia demonstrate the least familiarity (Figure 3).

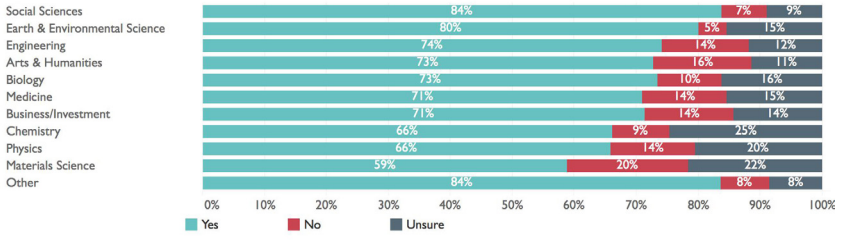


Figure 2 - Awareness of data that is free to access, reuse, repurpose and redistribute, by subject area, n=1436

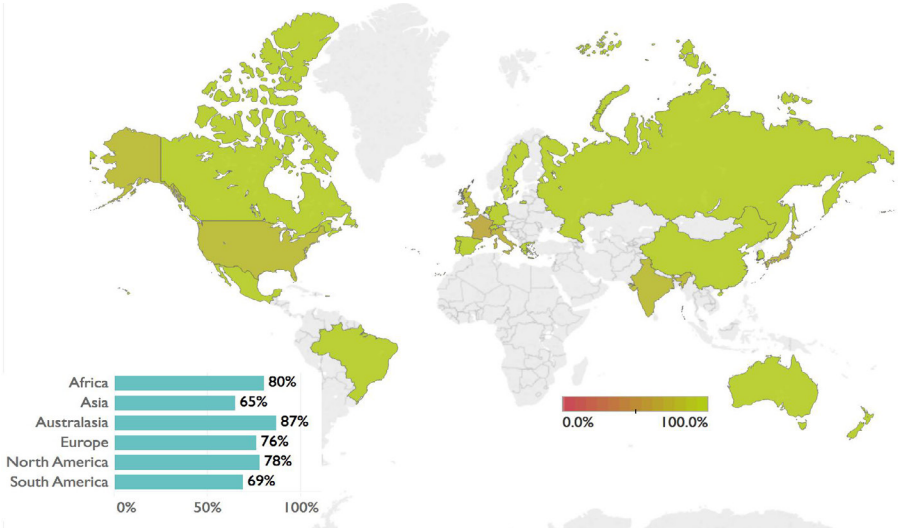


Figure 3 - Respondents who are aware of data that is free to access, reuse, repurpose and redistribute, by continent and by country where n>20, n=1915

Researchers are making data openly available - Approximately three quarters of respondents have made research data open at some point; of these, 24% do so frequently and 33% do so sometimes (Figure 4).

Researchers are reusing open data in their own research - A clear majority of respondents have reused data made open by other researchers (Figure 5); and that data was important to them over 80% of the time (Figure 6).

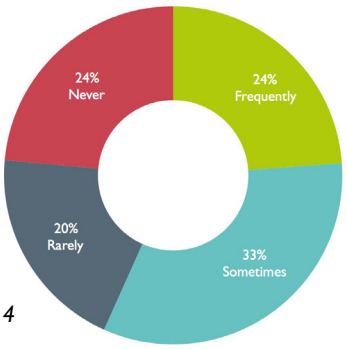


Fig 4

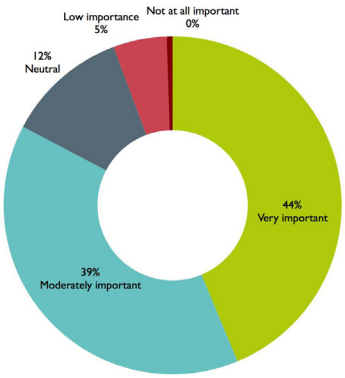


Fig 6

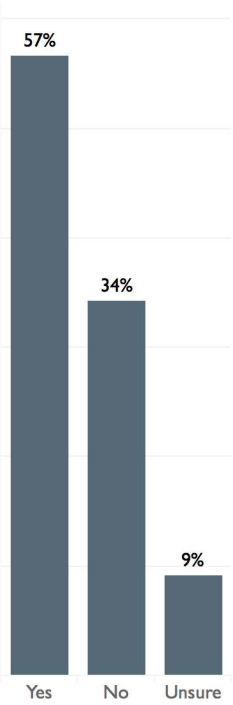


Fig 5

Figure 4 - Regularity with which respondents have made data free to access, reuse, repurpose and redistribute, n=1869

Figure 5 - Respondents who have reused data made free by others, n=1777

Figure 6 - Importance of freely available data to those who have reused it, n=1006

Researchers place value on credit they receive for making data open - Nearly 70% of researchers value a data citation as much an article citation. A further 10% value a data citation more than an article citation. Only 2% do not place any value on a data citation (Figure 7).

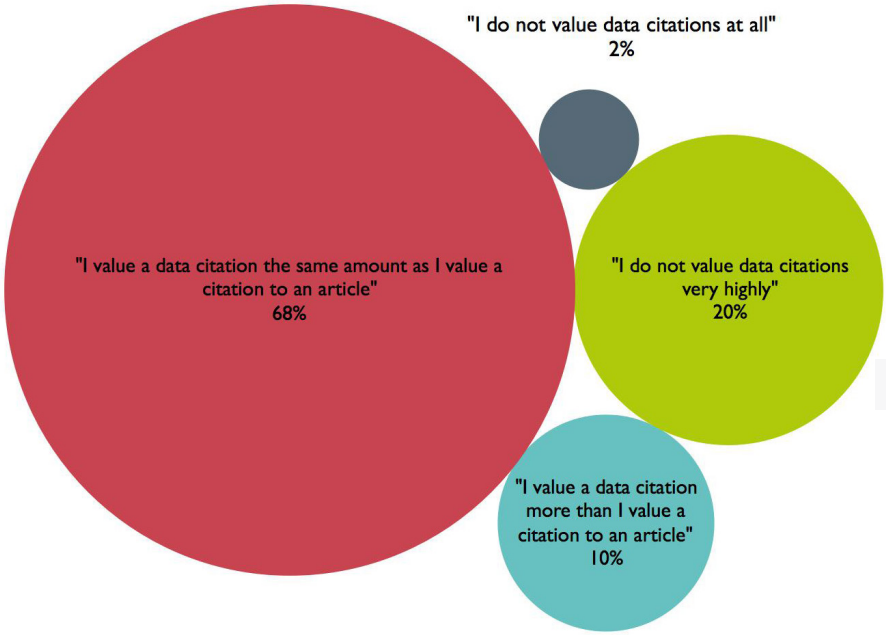


Figure 7 - Value placed on data citations, n=1714

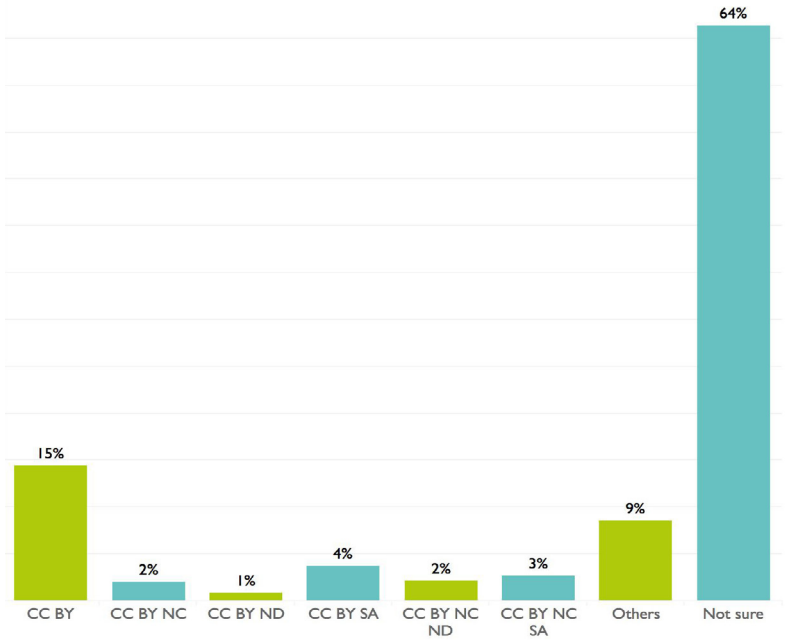
While the average across all respondents in all countries was just under 75% from countries in which we received a significant number of respondents, we found that developed research economies typically had 80%-90% of respondents replying that they were aware of open data. The top countries were Netherlands, Russia and Denmark with Australia, Switzerland, Argentina and Greece all scoring in the high-80 or low-90 percentiles. The US, Germany and the UK (the three largest producers of research publications) scored 79%, 82% and 82% respectively. Among the most developed research countries, the lowest levels of awareness were found in France (64%), Czech Republic (63%) and Italy (60%).

It is interesting to note that awareness of open data transcended age and career progression for the most part, with principal investigators and professors consistently scoring similarly to PhD students and post-doctoral fellows. Clinically-oriented colleagues tended to be less aware and those involved in research administration where, perhaps unsurprisingly, more aware.

2. Respondents admit to uncertainty and gaps in their knowledge; they want to know more

Researchers do not know how open they have made their data - 60% of respondents are unsure about the licensing conditions under which they have shared their data, and thus the extent to which it can be accessed or reused (Figure 8).

Figure 8 - Licenses used by respondents to make data freely available, n=1015



Researchers are uncertain as to who will meet the costs of making data open - 39% of researchers cannot identify a source of funds which will enable them to make data openly available (Figure 9); younger researchers in particular are uncertain as to where funds will come from.

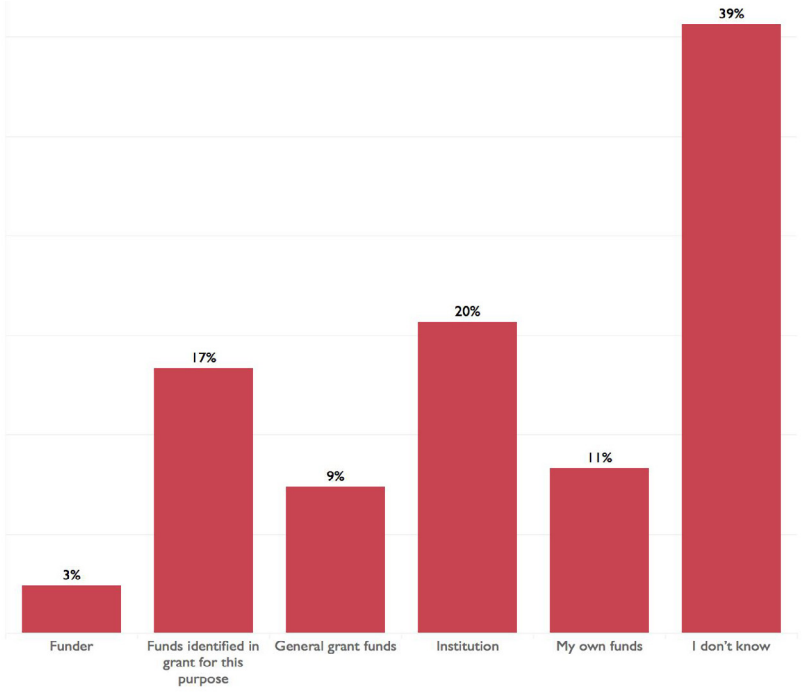


Figure 9 - Respondent's knowledge of source of funds for making data openly available, n=1554

A consistent proportion of researchers do not know what is required of them - Around 20% of researchers do not know whether their funders require them to make their data open. 25% do not know about their institution's requirements, and 31% do not know about publisher's requirements (Figure 10).

Funder requirements		Institution requirements		Publisher requirements	
I don't know	20%	I don't know	25%	I don't know	31%
No policy	53%	No policy	40%	No policy	33%
Policy exists	27%	Policy exists	36%	Policy exists	35%

Figure 10 - Respondent's awareness of open data policies, by Funder n=1451, by Institution n=1338, by Publisher, n=1401

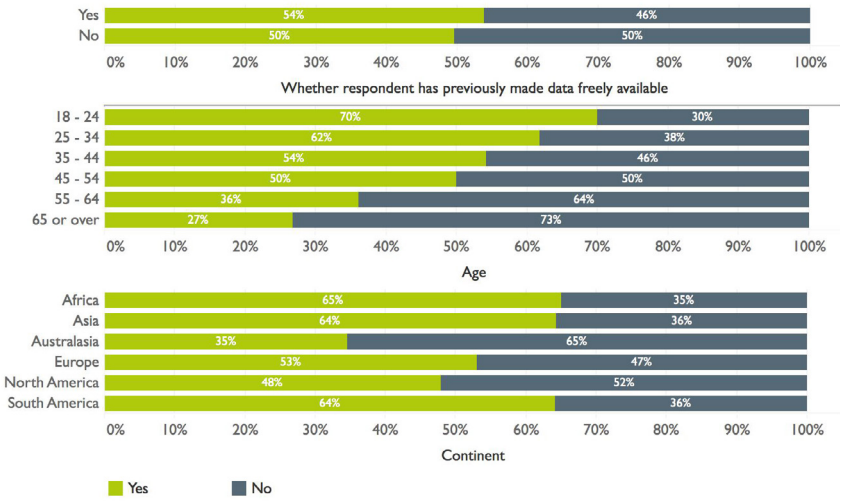
Respondents display remarkable consistency in their awareness of open data policies, irrespective of the stakeholder concerned (Figure 11).

		Funder requirements		
Institution requirements	Publisher requirements	I don't know	No policy	Policy exists
I don't know	I don't know	13%	0%	2%
	No policy	1%	0%	0%
	Policy exists	2%	1%	1%
No policy	I don't know	2%	5%	2%
	No policy	2%	23%	4%
	Policy exists	2%	8%	6%
Policy exists	I don't know	1%	1%	3%
	No policy	0%	1%	4%
	Policy exists	2%	1%	13%

Figure 11 - Consistency of respondent's awareness of Institution, Publisher and Funder open data policies, n=1252

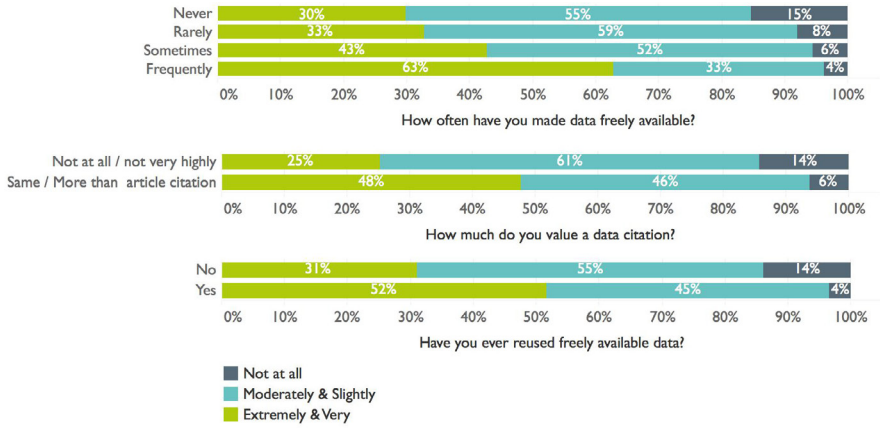
Researchers would like more guidance - Over half of respondents said that they would welcome more guidance on compliance with their funder's policy, with the desire for more information evenly distributed across those who have shared data openly in the past and those who have not. The desire is strongly correlated with age; and those in Asia are more likely to want more guidance than those in Europe or North America (Figure 12).

Figure 12 - Respondents desire for more guidance on meeting funder requirements, by data sharers n=1572, by age n=1428, by continent n=1430



Researchers are uncertain of how to cite datasets; their lack of confidence reflects their attitude towards open data - Less than half of respondents say they are extremely or very confident in how to cite a secondary research dataset; confidence levels correlate strongly with the value respondents place on data citations, how frequently they have previously shared data openly and whether they have used open data themselves (Figure 13).

Figure 13 - Level of confidence when citing secondary datasets by data sharers n=1737, value placed on data citation n=1714, reuse of data n=1577



While there is clearly a lot to understand about how data sharing works in practice, there are still many unanswered questions: What should be made open? What should be curated? How to cite data that you've used? Provenance and production principles? It is also clear that subject-based approaches to open data are key in changing perceptions and training the coming generations of researchers.

There are deep seated reasons for not sharing - some are cultural and some are systemic. Spending time to understand local subject-based effects will give us a clear path to tackle the blockers to openness.

3. We have some indication as to what the future holds, and it will be more open

Researchers who have never made data openly available are considering doing so - Of respondents who have not made any data open to date, 44% will definitely consider doing so in the future, and a further 46% might consider doing so (Figure 14).

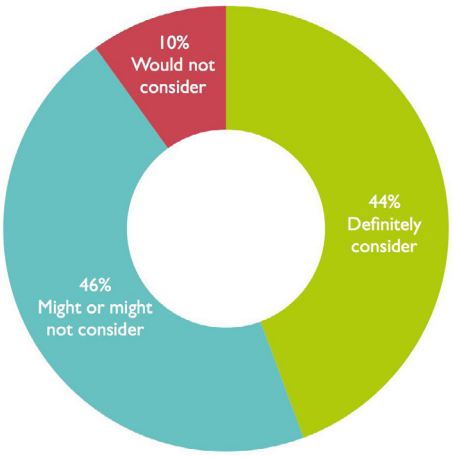


Figure 14 - For respondents who have never made data freely open, willingness to do so in future, n=233

Even researchers who have never made data open are reusing data made open by others - 35% of respondents reluctant to share data openly have reused open data in their work (Figure 15).

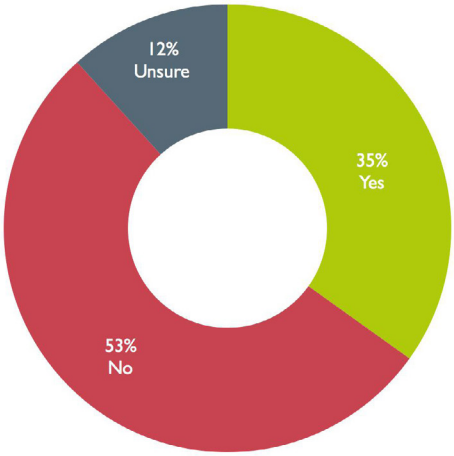
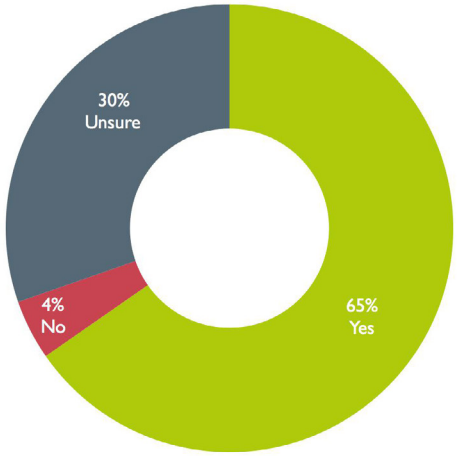


Figure 15 - For respondents who have never made data freely open to date, whether they have reused data others have made freely available, n=427

Of those who have not reused open data, two thirds would consider doing so in the future (Figure 16).

Figure 16 - For respondents who have never made data freely open to date nor reused data others have made freely available, willingness to reuse data in future, n=228



Regional differences exist and are likely to persist - North American respondents who have not made data open in the past or reused open data are most likely to do so in the future; Asian respondents are least likely to do so in both cases (Figures 17 & 18).

Figure 17 - Respondents who have never made data freely open or reused data that others have made freely available - willingness to reuse data in the future, by continent, n=161

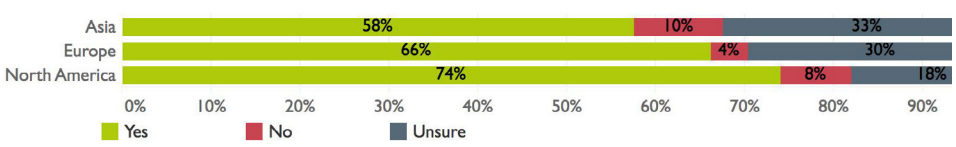
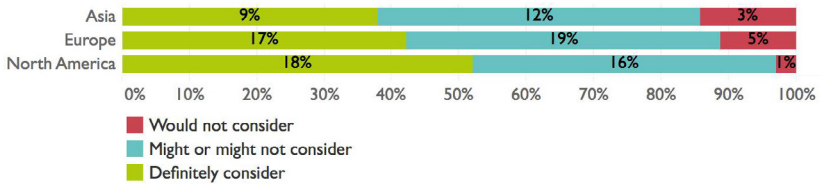


Figure 18 - Willingness of respondents who have never made data freely open to do so in the future, by continent, n=287



The appendix contains more detailed breakdowns from the survey. The data highlights many traits which correlate with how often respondents have made their data open and have reused open data themselves.

If a respondent has made their data open or reused open data, they are more likely to:

1. **Be aware of freely available data**
2. **Have experience of producing Data Management Plans**
3. **Produce large files from their research**
4. **Produce large numbers of files from their research**
5. **See file versioning as key to their research**
6. **Know where funds will come from to make data open**
7. **Actively annotate their data**
8. **Use Figshare and Github, versus other tools for sharing data**
9. **Collaborate with other researchers**

At Figshare, we have long contended that big data, while capturing the zeitgeist, is not the big problem that people think it is in the open data community. Big data tends to attract big attention and sometimes even big funding. That means that the problems of big data transfer, storage, provenance and a hundred other issues have been examined. The reason that we have the Internet, grid computing and ultra-fast graphics processes is all related to the big data revolution. Our data from this survey shows that correspondingly, it is more likely that you will share your data if you produce gigabytes of data rather than megabytes of data and more likely again if you produce terabytes (or more) compared with gigabytes. So, we still need to solve the "long tail" or "small data" problem. Only 22% of researchers producing megabytes of data share their data. Yet this is where most of the diverse and interesting data is: the negative results; the figures that didn't make it into the paper; the valuable runs of data that never made it into the tables that were published.

The survey and anonymized responses are both available on Figshare for reuse.

We hope that the results provide the sector with a foundation to track the evolution of how researchers will deal with data in the coming years.

"Only 22% of researchers producing megabytes of data share their data. Yet this is where most of the diverse and interesting data is: the negative results; the figures that didn't make it into the paper; the valuable runs of data that never made it into the tables that were published."

Open Data Will Save Lives – Notes from the AllTrials Campaign for Clinical Trials Transparency

Dr Till Bruckner, AllTrials campaign manager, London
& Beth Ellis, Intern at Sense about Science, London

"Opacity in medical research kills, and open data can save lives."

Discussions around open data can sound dry and boring. Often, they proceed from an abstract starting point – "transparency is good" – straight into highly technical discussions about software platforms, data scraping tools and a plethora of acronyms, leaving bystanders to wonder why they should care.

Here at the global AllTrials campaign for clinical trials transparency, we mobilize support around a far more compelling message: Opacity in medical research kills, and open data can save lives.

Clinical trials are at the heart of modern medical research. They are the best way we have of testing whether a medicine is safe and effective. Around the world, pharmaceutical companies, universities, government research institutes, foundations and charities conduct tens of thousands of trials a year to test new drugs, devices and treatments. Some of these trials can involve thousands of patients and take years to complete.

The problem is that the results of around half of all clinical trials currently remain hidden[1] because they are not posted or published anywhere. There is no complete list of all clinical trials, so we do not even know that some trials have taken place, never mind what was found in them. As a result, a huge amount of medical research goes to waste.[2] Potentially valuable findings are lost, different research teams unknowingly duplicate each other's work, and gaps in knowledge are hard to identify. This has a direct impact on doctors' and patients' abilities to make informed choices about treatment options.

Dr Aus Alzaid at his practice in Saudi Arabia, treats many patients with diabetes. There is a commonly observed link between diabetes and Alzheimer's disease, so he decided to examine the medical literature to see if any of the medications used to treat diabetes affected memory or caused dementia. The evidence on metformin, a widely used diabetes drug, seemed conflicting, with some studies showing a higher risk of dementia while others reported exactly the opposite. "The fact that the verdict on metformin was uncertain was somewhat unsettling for me," he wrote in a blog.[3] "I didn't want to give my patients anything to lower blood sugar if it meant them losing their minds too!"

Dr Alzaid discovered that a randomized, placebo-controlled, multi-year clinical trial had been conducted years earlier to study whether metformin caused dementia; but that its results had never been made public. "In the meantime, millions of people were taking the drug and could be at risk of dementia unless the work was published," he explained. The lead investigator of the trial eventually did publish the results – but only several years after the trial had been completed, and after having been prompted multiple times by Dr Alzaid.

Compounding the problem of non-reporting of results are systematic biases in the population of results that are reported. Academic researchers find it difficult to get the results of trials that yield negative results published in high impact academic journals. Research shows that trials with negative results are twice as likely to remain unreported as those with positive results.

The most (in)famous example is Study 329[4], a clinical trial of the antidepressant drug paroxetine conducted with teenage volunteers. A 2001 academic paper[5] based on the trial claimed that "Paroxetine is generally well tolerated and effective for major depression in adolescents." However, a subsequent analysis[6] by an independent research group showed a marked increase in suicidal behaviour among participants in the trial. Further scrutiny[7] of data from a wider range of trials suggested[8] that pharmaceutical companies had under-reported[9] instances of suicide in the documents they had submitted to regulators when applying for licenses to sell antidepressants. These revelations led to a series of court cases. In 2012, one of the companies involved agreed to pay a $3 billion fine[10] as part of a wider settlement.

In order to solve the problem of missing and biased trial evidence, the AllTrials campaign calls for all clinical trials – past, present and future – to be registered, and their methods and results to be fully reported. (Please note that AllTrials does not actively campaign for individual patient data sharing, a separate open data issue that is currently being hotly debated in medical research circles.)

AllTrials is campaigning for a future research landscape in which:

- all clinical trials are registered, with a full trial protocol, so that researchers can see exactly who is investigating what and nobody can bury unwelcome findings.
- a summary of results are posted where a trial was registered within one year of completion of a trial, so that other researchers, doctors like Dr Azaid, and their patients can find out what each trial discovered.
- all trial reports are posted online in full, with only minimal redactions, so that independent researchers can check whether their methodology was sound and their findings correctly presented.
- all the findings of trials conducted in the past are made available - most of the medicines we take today were developed decades ago.

In recent years, the medical research community has already made huge strides in the right direction. Rising awareness of the immense human (and financial[11]) cost of hidden trials has spurred governments, regulators, research

Image kindly provided by Sense about Science

"Open data in medical research matters to you – because it can save lives, including your own."

institutions and professional bodies into action. Governments have set up online clinical study registers to facilitate trial registration and results sharing. Trial registration and reporting is rapidly becoming seen as a must, both ethically and legally.

Pharmaceutical giant GlaxoSmithKline has voluntarily committed[12] to being transparent with its clinical trial data, "...to help advance scientific understanding and inform medical judgment". The International Committee of Medical Journal Editors has announced that it will no longer publish papers based on non-registered trials. In the UK, anyone planning to run a clinical trial must commit to registering it in order to receive ethics approval. Laws and regulations in the European Union, the United States and beyond are consistently evolving towards mandating greater clinical trial transparency[13]. An initiative called OpenTrials[14] is about to launch a platform integrating data from multiple trial registers to make it easier to locate trials and their results.

Large gaps remain, notably regarding trials conducted in the past, but the trend towards more open data in the sector is clear. To maintain and accelerate this positive momentum, the AllTrials campaign keeps repeating the same mantra to patients, doctors, researchers, regulators and politicians: Open data in medical research matters to you – because it can save lives, including your own.

1 http://www.alltrials.net/news/half-of-all-trials-unreported/
2 http://blogs.bmj.com/bmj/2016/01/14/paul-glasziou-and-iain-chalmers-is-85-of-health-research-really-wasted/
3 http://www.alltrials.net/news/diabetes-alzheimer-link-clinical-trials/
4 https://www.newscientist.com/article/mg22730394-500-new-look-at-antidepressant-suicide-risks-from-infamous-trial/
5 http://www.ncbi.nlm.nih.gov/pubmed/11437014
6 http://www.bmj.com/content/351/bmj.h4320
7 http://www.bmj.com/content/352/bmj.i65
8 https://www.scientificamerican.com/article/the-hidden-harm-of-antidepressants/
9 http://www.telegraph.co.uk/science/2016/03/14/antidepressants-can-raise-the-risk-of-suicide-biggest-ever-revie/
10 http://www.nytimes.com/2012/07/03/business/glaxosmithkline-agrees-to-pay-3-billion-in-fraud-settlement.html
11 https://www.theguardian.com/business/2014/apr/10/tamiflu-saga-drug-trials-big-pharma
12 http://www.alltrials.net/supporters/organisations/gsk-statement/
13 http://www.alltrials.net/news/un-calls-for-global-action-on-clinical-trial-transparency/
14 http://www.alltrials.net/news/opentrials-clinical-trials-transparency/

Practical Steps for Increasing the Openness and Reproducibility of Research Data

Natalie K. Meyers, MA, MLIS, Partnerships and Collaborations Manager, Center for Open Science, USA

This ***State of Open Data*** report provides a much needed snapshot of open data sharing practices today and provides a timely complement to a previous study: ***Science Gateways Today and Tomorrow:*** Positive perspectives of nearly 5,000 members of the research community[1]. Taken together these two surveys paint a broad picture of those producing, re-using, and making research data more open. It is encouraging to see in this burgeoning context that most major metrics are improving. 74% of ***State of The Data*** survey respondents have made research data open at some point, and of respondents who have never done so, 90% would consider making data open in the future. This interest closely echoes the Science Gateways survey, where 75% of respondents indicated that data collections were important to their research/education work, ranking it highly alongside data analysis tools and computational tools (72% each) and their interest in being able to rapidly publish and/or find domain-specific articles and data (69%).

The Center for Open Science(COS)[2] convenes reproducibility focused efforts like the Transparency and Openness Promotion (TOP).[3] TOP provides eight actionable steps journals and organizations can take to reward best practices in open, reproducible science with increasing rigor[4]. COS also incentivizes transparency through Open Data, Open Materials, and Open Practices Badges[5]. To provide infrastructure COS offers the Open Science Framework (OSF) an open-source platform that enables collaborative, transparent and reproducible work. You can find out more about OSF and COS's many initiatives at cos.io.

COS has prepared the following list of five practical steps for increasing the openness and reproducibility of research data to inspire more transparent research practices:

1. The first step is to get informed and stay informed - Educate yourself with primers and quick-study guides:

- 23 Things: Libraries for Research Data[7] will help you start learning and keep up with trends in open data sharing.
- For Social Scientists, the Inter-university Consortium for Political and Social Research (ICPSR) also offers a very complete *Guide to Social Science Data Preparation and Archiving: Best Practice.*[8]

- DataONE offers a *Best Practices Primer*[9] for those new to data management and a *Best Practices database*[10] where users can select an area of the research lifecycle diagram they want to learn more about.

2. The second step is planning - Data Management Plans, Pre-analysis Plans & Pre Registration:

- **A Data Management Plan (DMP)** describes how you will acquire or generate data, how you will manage, describe, analyze, and store those data, and finally how you will preserve and make available your data. Many funders require data management plans. The DMPTool[11] and DMPonline[12] are two among a number of helpful tools for collaboratively authoring compliant plans aligned with funder requirements.
- **A Pre-analysis Plan** is a detailed description of the analysis to be conducted written in advance of seeing the data. It may specify hypotheses to be tested, variable construction, equations to be estimated, controls to be used, and other aspects of the analysis. The Registry for International Development Impact Evaluations (RIDIE) has a pre-analysis checklist helpful to those writing pre-analysis plans[13].
- **Registered Reports & Preregistration** can increase the credibility of hypothesis testing by confirming in advance what will be analyzed and reported, and even accepted for a journal article. Researchers can reduce the file drawer effect, prevent biased data analysis, and engage in peer review before results are known with Preregistration[14] and Registered Reports[15].

3. The third step is using data management best practices during data collection & analysis:

Accurately describe and report your research

- **Thoroughly describe methods providing open materials where possible.** Remember to provide sufficient explanation for an independent researcher to understand how your materials relate to your reported methodology.
- **Clearly distinguish between confirmatory** (hypothesis testing) **versus exploratory** (hypothesis generating) **analyses**. Both are important for scientific discovery.
- **Make clear your effect size and confidence interval** reporting
- **Use machine readable community standards for metadata** appropriate for your discipline and information systems.

Make data files & code understandable by documenting data and code as you work. Create and keep up-to-date: clearly labeled variables,codebooks,data dictionaries,well-documented and shareable software/scripts.

Track & Preserve digital files, computational workflow and/or data analysis

- Employ version control, registration, and documentation features during data collection and analysis.
- Adopt and consistently use file naming conventions
- Employ literate programming methods
- Create re-usable, shareable workflows and consider containerizing work with docker-centric tools to preserve your software and platforms.

4. The fourth step is to share and archive data and materials:

- **Use open file formats** to share and archive data to meet funder, publisher or repository requirements. If your discipline or toolsuite produces data in proprietary formats, if appropriate use conversion tools to create a parallel set of data in an open format. Preserve them along with the native file formats.
- **Claim a permanent author identifier** like an ORCID[17] and/or ResearcherID[18]. This helps distinguish you from other authors and is increasingly integrated into manuscript and grant submission systems' automated linkages.
- **Deposit and preserve data in an open repository**: Funders may specify a preferred repository, affiliated organizations may provide an institutional repository, or there may be a disciplinary repository that provides the widest audience for your data. Re3data[19] maintains a registry of repositories and sharing platforms so you can find research data repositories most appropriate for your work.

The abiding steps are to identify where to seek expert help and to get to know proximal and aspirational peer communities:

- The library is a good start. Most research libraries have data management consultancies. In 2013, 74% of responding U.S. and Canadian libraries in a survey of the Association of Research Libraries reported offering such services, and 23% of the others reported they were planning to begin offering such services[20].
- Identify disciplinary, government agency, and collaborative organizations aligned with your research to seek expert help and be informed by your peers, like: Berkeley Initiative for Transparency in the Social Sciences (BITTS)[21], CODATA[22], Data Curation Centre[23],DataONE[24],Data and Software Preservation for Open Science(DASPOS)[25], Federation of Earth Science Information Partners (ESIP)[26], ICPSR[27], Society for Improving Psychological Science (SIPS)[28]. Browsing the lists of SHARE[29] and TOP[30] signatories and contributors is another way to find aligned communities of practice.

- The Research Data Alliance[31] convenes an international group from 111 countries and provides a neutral space where its 4,000+ members come together online and in bi-annual plenaries across domain, research, geographical and generational boundaries to develop and adopt infrastructure that promotes data-sharing and data-driven research .This diversity of membership ensures you can find an expert and/or identify quickly with a community of practice at RDA.

Improving the reproducibility of research literature is a complex challenge, that will be most effectively accomplished through the collective will of researchers, data managers, publishers, universities, professional societies, and funding agencies pulling together around best practices.

1 *Science gateways today and tomorrow: Positive perspectives of nearly 5,000 members of the research community.(2015) by Lawrence, KA, Zentner, M, Wilkins-Diehr, N, Wernert, JA, Pierce, M, Marru, S, Michael, S. Concurrency and Computation: Practice and Experience 2015, DOI: 10.1002/cpe.3526.*
2 *Center for Open Science Available: https://cos.io/*
3 *Transparency and Openness Promotion Guidelines https://cos.io/top/*
4 *Incentivizing Transparency by Mellor, D. Editorial Office News 2016 Vol. 9 (8), DOI: 10.18243/eon/2016.9.8.1*
5 *Badges to Acknowledge Open Practices Available: https://osf.io/tvyxz/wiki/home/*
6 *Open Science Framework Available: https://osf.io*
7 *23 Things: Libraries for Research by Witt, M., RDA Libraries for Research Data Interest Group. DOI:10.15497/RDA00005*
8 *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle (5th ed.). Inter-university Consortium for Political and Social Research (ICPSR). (2012). Ann Arbor, MI. Available: http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf*
9 *DataONE Best Practices Primer Available: https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf*
10 *DataONE Best Practices Database Available: https://www.dataone.org/best-practices#search*
11 *DMPTool Available: https://dmptool.org/*
12 *DMPonline Available: https://dmponline.dcc.ac.uk/*
13 *Analysis Plan Checklist RIDIE Registration Help. Available: http://ridie.3ieimpact.org/index.php?r=site/page&view=registrationHelp#AnalysisPlanFileId_Section*
14 *Preregistration Challenge Available: https://cos.io/prereg/*
15 *Registered Reports Available: https://cos.io/rr*
16 *Literate programming by Knuth, D. The Computer Journal (1984) 27(2): 97-111.doi: 10.1093/comjnl/27.2.97*
17 *ORCID Available: http://orcid.org/*
18 *ResearcherID Available: http://www.researcherid.com/*
19 *Registry of Research Data Repositories Available: http://www.re3data.org/*
20 *SPEC Kit 334: Research Data Management Services by Fearon, D; Gunia, B; Lake, S; Pralle, BE; Sallans,A. (July 2013) Available: http://publications.arl.org/Research-Data-Management-Services-SPEC-Kit-334/*
21 *Berkeley Initiative for Transparency in the Social Sciences Available: http://www.bitss.org/*
22 *International Council for Science : Committee on Data for Science and Technology (CODATA) Available: http://www.codata.org/*
23 *Data Curation Centre (DCC) Available: http://www.dcc.ac.uk/*
24 *DataONE Available: https://www.dataone.org/*
25 *Data and Software Preservation for Open Science (DASPOS) Available: https://daspos.crc.nd.edu*
26 *Federation of Earth Science Professionals (ESIP) Available: http://www.esipfed.org/*
27 *Interuniversity Consortium for Political and Social Research (ICPSR) Available: https://www.icpsr.umich.edu/icpsrweb/*
28 *Society for the Improvement of Psychological Science (SIPS) Available: http://improvingpsych.org/*
29 *SHARE Available: https://share.osf.io/sources*
30 *TOP Signatories Available: https://cos.io/top/#list*
31 *Research Data Alliance Available: https://www.rd-alliance.org/*

Emerging Policies for Open Research Data in the United States

Heather Joseph, Executive Director, SPARC, Washington DC, USA

Background

In the United States, the federal government invests approximately $60 billion each year on basic and applied scientific research with the expectation that this investment will stimulate new ideas, accelerate scientific discovery, fuel innovation, grow the economy, create jobs, and, in general, improve the welfare and well-being of the public. Increasingly, research funders have recognized that by ensuring the outputs of this research – including data – can be freely accessed and fully used by the widest possible audience, progress towards these goals can be significantly accelerated. U.S. policymakers have also recognized, that in an era where improving the transparency and accountability of government has taken center stage, the need to create a policy framework that effectively supports all stakeholders in a transition to a more open system of sharing research results.

Policy Precedent and Evolution

The United States has a long history of information policy precedents that have created a strong foundation for the creation of an effective research data-sharing framework. Dating back to the mid-1960s, key policies have been issued which have become progressively more explicit in articulating expectations for sharing government-produced/funded information, ranging from the Freedom of Information Act of 1966, to the Office of Management and Budget (OMB) Circular No. A-130 in 1985, to The Digital Accountability and Transparency Act (DATA Act) recently passed by Congress in 2014.

While none of these policies specifically targeted digital research data, and many were drafted prior to the advent of the Internet, they nonetheless have played a key role in helping to define expectations for sharing data of all kinds. OMB Circular A-130, in particular, clearly outlines key principles for sharing government information of all kinds, and speaks to the heart of the current objectives of U.S. federal research funders, noting:

> *"…Government information is a valuable national resource, and the economic benefits to society are maximized when government information is available in a timely and equitable manner to all," and further calling for "Open and unrestricted access to public information at no more than the cost of dissemination..."*

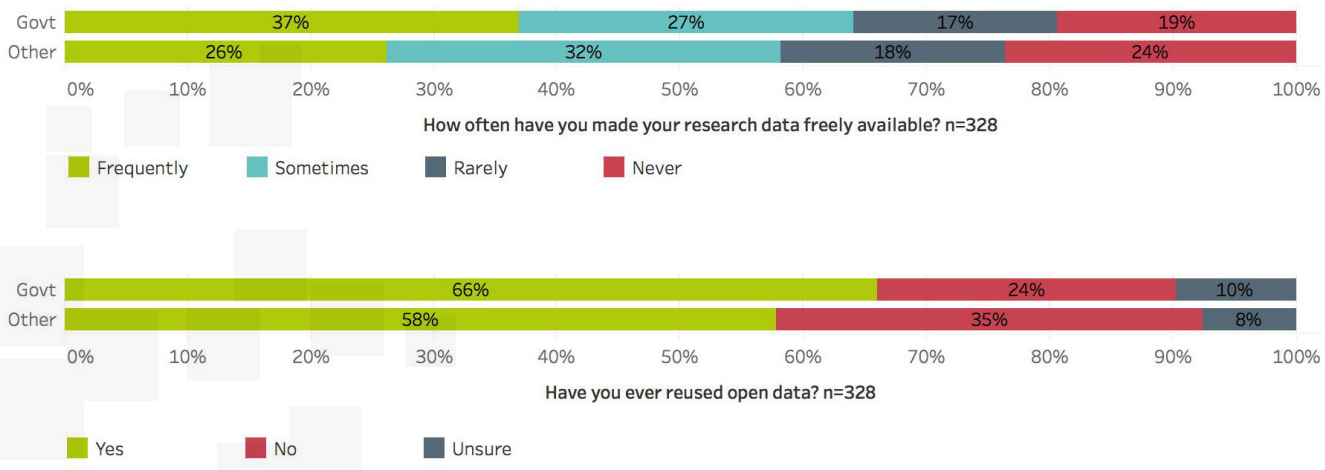
This has proven to be an extremely useful foundation for U.S. research funders to build data sharing policies upon, and one that the Obama Administration has taken full advantage of. On his first day in office in January 2009, President Obama issued a sweeping Open Government Directive, outlining guidelines promoting transparent and participatory government. The first concrete step that agencies were required to take was to publish government information online and in open formats to increase its accessibility and utility to the public. They have continued to issue ever-more granular policies focusing on all types of digital government data, culminating in a 2013 Executive Order making "Open and Machine Readable" the default for all government data. Just a few months later, this administration honed directly in on research data, with an additional Directive from the Office of Science and Technology Policy (OSTP) requiring public access to all federally funded research outputs.

Current Policy Environment

The OSTP Directive has been a landmark in U.S. research data policy development. It set out, for the first time, specific requirements for U.S. research funders to develop policies to make all digital data resulting from unclassified research supported by U.S. federal funding "accessible for the public to search, retrieve, and analyze." (Figure 19).

The Directive provides an interesting window into the current state of research data sharing policies and reflects the deep diversity of data generated in various research disciplines, as it, understandably, contains a high-level of ambiguity and what often seem to be contradictory directions. For example, the Directive tasks agencies with the goal of "maximizing access to data," while still fully protecting confidentiality and personal privacy; recognizing proprietary interests, business confidential information and IP rights; and balancing value of long-term preservation and access with costs and administrative burdens. These apparent contradictions actually serve as important – and useful - indicator of areas where tension between the potential benefits of full open sharing of data runs directly into potential negative consequences of such sharing. In order to solve this problem, additional input from the research community will be required.

Figure 19 - From Figshare's Open Data Survey: North American respondents who have been government funded who have made data freely available or reused open data.



Currently, draft or final research data sharing policies have been released by 15 of the 19 U.S. science agencies covered by the OSTP Directive. It's also clear from their content that creating final policies will be an evolutionary process, requiring significant community involvement and input to iterate towards working policies. Unlike in many other cases, creating research data sharing policies will not be a "one-and-done" policy drafting process.

All of the policies released differ somewhat both in interpretation of the OSTP guidelines, as well as in the implementation processes that they propose, however, there are a lot of significant commonalities. Most of the U.S. agency policies:

- require the submission of data management plans at the proposal stage
- provide direction for approved locations for data deposit/storage
- acknowledge the need for routine attribution for data
- require the creation of agency inventories of data to aid discovery
- support public and private collaboration to achieve aims of data sharing policies
- recognize the need for robust long-term preservation strategies

There is not yet a common set of standards for any of these policy components. For example, while all agencies require data management plans to be submitted at funding proposal stage, we don't have a common set of attributes and expectations for these plans. This will likely be addressed as agencies work through additional community consultations. In the short term, this will make policy implementation more labor intensive for both the funders and recipients to comply with.

With a high level of ambiguity still a hallmark of these emerging policies there is also an undercurrent of concern over compliance confusion from the institutions that are the primary recipients of federal research funding. There is also a general willingness on their part to work together with funders and the wider community on solutions to decrease compliance friction.

What can we do to keep things moving in right direction?

This complex and somewhat volatile environment highlights the need for regular and close collaboration not only among funders, but also with the academic and research community. There is a need to evolve research data sharing policies at paces and in directions that are acceptable to and sustainable by the research enterprise.

It also underscores the necessity of an evolutionary approach in order to produce effective research data policies for different disciplines. An approach that emphasizes regular pilots to test assumptions, and includes mechanisms to gather and incorporate community feedback on those pilots could be effective in this environment.

"We need to be able to effectively support and sustain the infrastructure needed to make data sharing a reality."

"Harmonizing" policy components across U.S. Federal agencies wherever possible to reduce operating friction should be a key shared goal. Creating consistent policy components, and ongoing interagency collaboration on implementation requirements, is also vital for the ultimate success of these policies.

We need to be able to effectively support and sustain the infrastructure needed to make data sharing a reality. Additional investments in the infrastructure needed to support access to and use of research data for the long term is essential. We must build an effective, sustainable infrastructure needed to support our vital national collective interests, and this will require the community to work together to secure additional investments to ensure we can achieve the laudable objectives of our emerging open research data sharing policies.

Building Trust - The State of Open Data in Burkina Faso

Malick Tapsoba, Deputy Manager, Burkina Open Data Initiative, Agence Nationale de Promotion des TIC (ANPTIC)

Burkina Faso, ranked as one of the poorest countries in the world, is the first African francophone country to take positive steps to bring significant, social, economic and environmental benefits to its population via an open data initiative. The initiative, Burkina Open Data Initiative' (BODI), is paving the way to successfully creating an open data environment, making available a wide range of datasets in a reusable format from sources including government, private sector and civil society. The data can be accessed and reused to bring added value through the creation of new services and improved service provisions in the country.

"The initiative, Burkina Open Data Initiative' (BODI), is paving the way to successfully creating an open data environment."

Since March 2013, BODI has been engaged in a process of releasing and making open non-sensitive data collected by public institutions to increase public access to information on key social policies and services. Three years since the start of this open data initiative, we take a look at the situation in Burkina Faso across a number of areas.

1. Data release

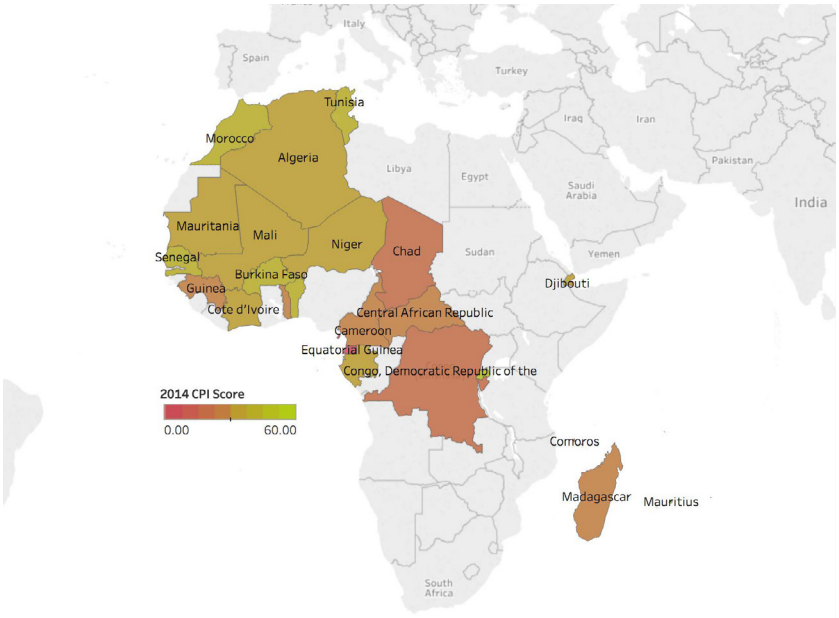
Burkina Faso government departments and institutions have a tradition of publishing data in the form of a statistical yearbook generated as part of their daily activities. This is a very significant publication for the country but has shortcomings in that: i) the yearbook does not provide an exhaustive list of institutional and ministerial data; ii) some of the yearbooks are not published in a reusable format; iii) some yearbooks are in the form of reports with statistical tables buried within them; and, iv) not all yearbooks are centralized.

The aims of the BODI project are to establish a catalogue for data collected by the ministry and institutions, and collect, process and publish the data, on a central platform, which can be re-used. So far the project has managed to create a central portal of data accessible at http://data.gov.bf; collect a large mass of government and civil society data; and, process and publish more than 260 directly reusable datasets.

2. On the reuse of data

BODI provides excellent examples of how data is being shared and reused in Burkina Faso. For example:

Open Elections - For the first time, the results of the 2015 Presidential election in Burkina Faso were made openly available in real-time. The official election website showed live results by district for each presidential candidate, and which candidate was leading in each of the provinces. Using 'Open



Corruption Perceptions Index data in francophone countries in Africa from 2014

Election', a mobile-responsive web based application run by BODI, anyone could follow the emerging results of the elections in Burkina Faso in 2015 and 2016. Partners of this initiative include the Open Data Institute and the World Bank. Open Election has contributed to the organization of free, transparent and accepted elections in Burkina Faso. The site allows users to follow the results of the presidential and legislative elections in 2015 and municipal elections in 2016 in Burkina Faso. Africa has a history of disputed election results and this application has contributed to the organization of free, transparent and accepted elections by all. The application has also enabled Burkina Faso to be invited to several international events to share its experience regarding: "Open Data and Elections".

'**Our Schools our Data**' Nos Ecoles Nos Donnée (NENDO) http://nendo.data.gov.bf) gives a detailed outline of educational structures in Burkina Faso and has developed a decision tool for education stakeholders. The tool allows users to find data on the education system at school level, including pass rates, gender ratios, and presence of running water. The application was highlighted during OpenDataCon 2015 as one of the best examples of data reuse in the world by the Vice-President of the World Bank for Africa, Mr. Makhtar Diop.

CARTEAU-BF is a decision support tool for all water stakeholders, from government to end users of these water points and sanitation through technical and financial partners. It aims to provide in real time the situation of water and sanitation in Burkina to enable better decision making.

With CARTEAU:

- the government has a tool to measure the effectiveness of the policy on water and sanitation;
- technical and financial partners can see the impact of their investment in the water sector and sanitation and also to better target areas in which to invest;
- private investors to identify business opportunities;
- the end user to know the situation of his community compared with the other communities.

Instilling the culture of systematic reuse of data still needs further

encouragement for it to be firmly embedded in Burkina Faso, especially for groups including developers, journalists, researchers, planners and economists. This continues to be a major part of BODI's work.

3. Ecosystem development

Since BODI launched, there has been a shift from the concept of how open data can help Burkina Faso, to real changes in societal behavior.

- The interest of open data appears increasingly in the speeches of politicians. The current Prime Minister, Paul Kaba Thieba, said in his State of the Nation Address that, "the Government has continued the implementation of project Burkina Open Data Initiative allowing, through the Open Election platform, to have the results of the elections of 29 November 2015 in real time."
- The creation of many associations working for government are strengthening the transparency efforts of the BODI. These include, among others, the Open Burkina association working in the field of transparency in the 'extractive industries' and the association 'Open Education' fighting for the accessibility of education data.
- The growing interest of development partners for open data initiatives. Indeed, many associations receive support from development partners for the implementation of their transparency projects

4. The future for open data in Burkina Faso

Despite the challenges facing Burkina Faso, the BODI is now established as one of the developments that the country embraces in its move towards becoming a transparent and open society. From a timid beginning in 2013, stakeholders of open data in Burkina Faso have made the concept a reality and the interests are now apparent in all spheres of society.

However, it remains that some nationally important projects will strengthen the open data initiative. These include:

- Burkina Faso's membership of organizations such as the Open Government Partnership (OGP) and the Global Open Data for Agriculture and Nutrition (GODAN).
- The adoption of a law requiring all data collected which has received public funds to be made openly available.
- The creation of an open data project incubator to promote effective reuse of data in order to create economic wealth for Burkina Faso.

Burkina Faso's commitment to open data is commendable and BODI is laying the groundwork towards transparency and accountability nationwide. Data driven initiatives aim to improve the quality of information available to its people, and in turn is increasing the availability of knowledge which will subsequently contribute to economic growth and international recognition.

The State of Australian Research Data – Systems are Ready but Where are the Incentives?

David Groenewegen, Director, Research at Monash University, Victoria, Australia

"The Australian Federal government has made a strong public commitment to opening up its data and is pushing data driven innovation in its Public Data Policy Statement."

Over the past ten years we have seen tremendous advances in the provision of infrastructure, both technical and human, to enable and encourage better storage, management and collaboration around research data.

Australian government infrastructure projects such as the Australian National Data Service, have resulted in most local universities and research organizations having not only a data management policy, but also an active data management infrastructure. This has resulted in thousands more datasets being made available, with an even larger number being effectively managed (Figure 20). This represents a substantial improvement on the situation ten years ago, and it continues to grow and develop. Well managed data enables better and more efficient research.

Additionally, the Australian Federal government has made a strong public commitment to opening up its data and is pushing data driven innovation in its Public Data Policy Statement[1]. The statement, which is supported by similar policies in many state governments, promotes the idea that 'open data has the potential to stimulate innovation and help grow the economy, improve service delivery and decision making for planning and policy development'. The belief here is that harnessing the value of public data will increase Australia's capacity to remain competitive in the global digital economy.

The Australian government has also acknowledged the need for better systems to discover and disseminate data, stating: "All new systems must support discoverability, interoperability, accessibility and cost-effective access to data". This applies to data produced by Australian government entities, but falls short of mandating similar principles to research data produced at Australian universities. Even though there are a growing number of mandates from funders and institutions requiring and encouraging data to be made open and institutions have policies and infrastructure in place it seems like not enough researchers are embracing open data.

Resistance to making data available remains high. My team at Monash University Library have contacted hundreds of research staff about making just the data related to their recent journal articles available, and we're lucky if one in 20 even want to consider it, let alone actually do it.

This is completely understandable. While there are many important reasons why data could not or should not be shared, there are also not enough

incentives to make it worthwhile. Open data sharing can be an added burden, especially given how busy most researchers are, and how many things they have to do to gain credit - or risk being penalized. Sharing data is one more piece of work, one more system to learn and possibly, a risk to their careers if something goes wrong.

So for those of us who believe in the value of open data, what do we need to do to encourage researchers to do that little bit more?

1. **We need to make it easier.** Here at Monash University we try to make it easy for all our researchers by providing a university wide data management platform. It still needs to be more connected to other systems that researchers use, so they can get the maximum return for the least effort, but it is a good start.

2. **We need better incentives** - preferably credit for making data open that contributes to career advancement and/or financial reward. Because people deserve to be rewarded for their efforts.

3. **We need more evidence of the benefits for individual researchers.** Researchers like evidence, that's what they look for. There isn't enough to show that the effort needed to make their data open and discoverable benefits them. As in all walks of life, saying this is the "right thing to do" will only achieve a certain amount.

4. **We need to be prepared to acknowledge the potential downsides**, and put in place processes to help mitigate them. For instance, the process of de-identification of data is difficult and important, so how can this be made easier and shown to be safe?

We have the policy and the infrastructures to make this happen, and increasingly this has led to better managed data, data that can be stored for the future and shared in a way that researchers can manage. But if we really want it to be made more open, the incentives need to change.

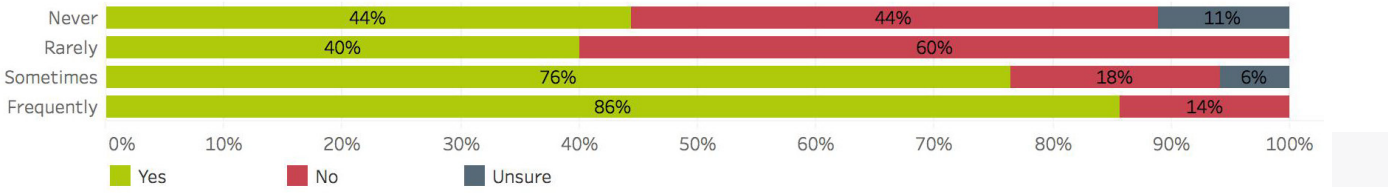


Figure 20 - From Figshare's Open Data Survey: Frequency with which Australian respondents have made data freely available versus whether they have reused data made openly available by others, n=1787

I Australian Government Public Data Policy Statement https://blog.data.gov.au/news-media/blog/australian-government-public-data-policy-statement

Can Japan Catch Up? Fostering Culture, People, and Community for Research Data

Dr Kazuhiro Hayashi, Senior Research Fellow, Science and Technology Foresight Center, National Institute of Science and Technology Policy, Japan & Nobuko Miyairi, Regional Director, Asia Pacific, ORCID Inc, Tokyo, Japan

"Not only are the philosophy and goals of open science being advocated; Japan has been steadily making strides towards building the infrastructure for research data."

Open science is a strong undercurrent of science policy discussion in many countries. Japan is no exception. In early 2015, the Council for Science, Technology and Innovation (CSTI) issued a report that illustrated guiding principles on promoting open science in Japan[1] based on expert panel discussions. With the global movements of open science and an array of initiatives as a backdrop, the CSTI urged Japan's research community not to lag behind, but to build a solid and sustainable framework to keep abreast of global open science trends. The report devoted a considerable amount of space to open research data; defining the scope and responsibilities of each sector. Publicly funded institutions must disseminate their research results and data data (Figure 21). The 5th Science and Technology Basic plan[2] started from 2016 and it clearly declared the promotion of Open Science based on the CSTI's report. And also, Japan raised an agenda of Open Science at G7 Science and Technology Ministers' Meeting held in Tsukuba, Japan, releasing the Tsukuba Communique[3] which described further commitments by organizing a working party to implement Open Science Policy.

At a grass-roots level, there have been researchers, research managers, librarians and information professionals who have engaged in and are contributing to the global open research data initiatives. However, the CSTI report has accelerated ongoing efforts and invited more people to participate in this initiative. The 7th Plenary of Research Data Alliance (RDA) was held in Tokyo in early March and attended by more than 100 Japanese participants, accounting for 30% of total attendees. A month later, a debrief session was organized at the National Diet Library, where key participants of the RDA plenary shared their views on what it takes to be involved in community initiatives like RDA.

Not only are the philosophy and goals of open science being advocated; Japan has been steadily making strides towards building the infrastructure for research data. The[1] [NM2] Japan Link Center (JaLC), a Japanese DOI agency, facilitated an experimental project of DOI registration for research data in 2014. 14 research institutions and universities participated, the project goal was to establish reliable workflows to assign DOIs to research data. During the course of this project, a number of issues and challenges were identified – the timing of DOI assignment, granularity of research data, the need for

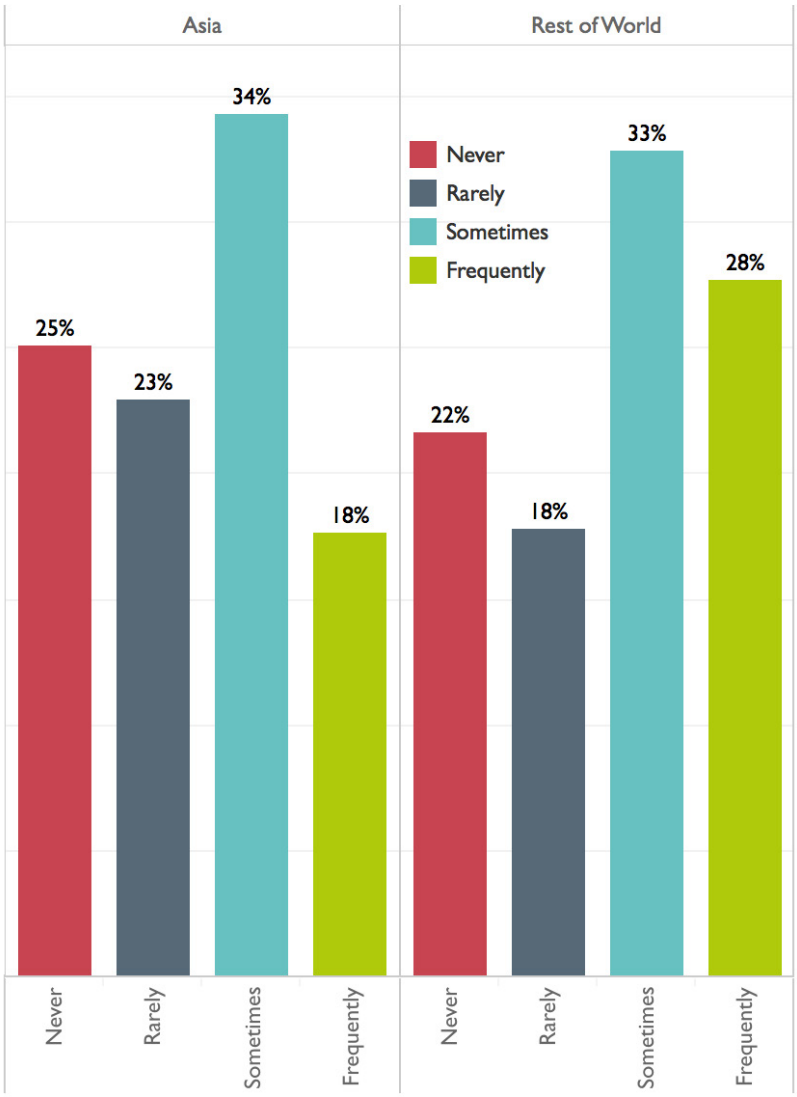


Figure 21 - From Figshare's Open Data Survey: Frequency with which Asian respondents have made data freely available versus all other respondents, n=1430

sustainable and consistent organizational policy, the lifecycle of research data, handling of dynamic data, versioning, and more – all incorporated into JaLC's guidelines for registering DOIs for research data[4].

With a large number of institutional repositories[5], training appropriate staff for managing research data is imperative. Because of this, Japan has a growing number of librarians who are able custodians of the institutions' data. JaLC's experimental project brought together a network of individuals from across multiple disciplines to facilitate this discussion. The Research Data Utilization Forum was initiated in June 2016, allowing these individuals to continue networking and to participate in ongoing discussions. Ultimately, this forum will serve as a vehicle for fostering the culture in research data management, thus harmonizing with the global landscape of open research data as recommended by the CSTI.

Without these recommendations being fully embedded, it continues to be an uphill struggle to motivate researchers to fully embrace open data.

1 Cabinet office, Government of Japan. Promoting Open Science in Japan: Opening up a new era for the advancement of science. March 30, 2015.http://www8.cao.go.jp/cstp/sonota/openscience/150330_open-science_en1.pdf

2 The 5th Science and Technology Basic Plan: http://www8.cao.go.jp/cstp/kihonkeikaku/5basicplan_en.pdf

3 The Tsukuba Communique, G7 Science and Technology Ministers' Meeting: http://www8.cao.go.jp/cstp/english/others/communique_en.html

4 Japan Link Center. Guidelines for Registering DOIs for Research Data. October 20, 2015. https:doi.org/10.11502/rd_guideline_en

5 496 institutional repositories are listed as of August 31, 2016 by the National Institute of Informatics at: http://www.nii.ac.jp/irp/en/list/

The Bird in Hand: Humanities Research in the Age of Open Data

Daniel Paul O'Donnell, Professor of English, University of Lethbridge, Alberta, Canada

"Humanities researchers rarely have an incentive (or capability) to prevent others from accessing their raw material and entire research domains."

Traditionally, humanities scholars have resisted describing their raw material as "data" [10].

Instead, they speak of "sources" and "readings." "Primary sources" are the texts, objects, and artifacts they study; "secondary sources" are the works of other commentators used in their analyses; "readings" can be either the arguments that represent the end product of their research or the extracts and quotations they use for support.

These definitions are contextual. The primary source for one argument can be the secondary source for another or, as in the case of a "critical edition" of a historical text, simultaneously primary and secondary. Almost any document, artifact or record of human activity can be a topic of study. Arguments proposing previously unrecognized sources ("high school yearbooks, cookbooks, or wear patterns in the floors of public places") are valued acts of scholarship. [1]

This resistance to "data" is a recognition of real differences in the way humanists collect and use such material. In other domains, data are generated through experiment, observation, and measurement. Darwin goes to the Galapagos Islands, observes the finches, and fills notebooks with what he sees. His notes (i.e. his "data") "represent information in a formalized manner suitable for communication, interpretation, or processing"[2]. They are "the facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors"[3]. Given the extent to which they are generated, it has been argued that they might be described better as capta, "taken," than data, "given".[4]

The material of humanities research traditionally is much more datum than captum, finch than note. Since the humanities involve the study of the meaning of human thought, culture, and history, such material typically involves other people's work. It is often unique and its interpretation is usually provisional, depending on broader understandings of purpose, context and form that are themselves open to analysis, argument and modification. In the humanities, we more often end up debating why we think something is a finch than what we can conclude from observing it.

Perhaps most telling is the fact that humanities sources, unlike scientific data, are usually practically as well as theoretically non-rivalrous[5]. Humanities researchers rarely have an incentive (or capability) to prevent others from accessing their raw material and entire research domains (e.g. Jane Austen studies) can work for centuries from the same few primary sources. Priority disputes that occur regularly in the sciences[6] are almost non-existent within the humanities.[1]

The digital age is changing one aspect of this traditional disciplinary difference. Mass digitalization and new tools make it possible to extract material algorithmically from large numbers of cultural artifacts. Where researchers used to be limited to sources in archives and libraries to which they had physical access, digital archives and metadata now make it easier to work across complete historical or geographic corpora: all surviving periodicals from 19th century England, for example, or every known pamphlet from the Civil War. In the digital age, humanities resources can be capta as well as data.

Such changes allow for new types of research and improve the efficacy of some traditional approaches. But they also raise existential questions about long-standing practices. Traditionally, humanities researchers have tended to work with details from a limited corpus to make larger arguments: "close readings" of selected passages in a given text to produce larger interpretations of the work as a whole; or of passages from a few selected works to support arguments about larger events, movements or schools. In one famous but far from atypical example, author Ian Watt uses readings from five novels and three authors as the main primary sources in his discussion of the Rise of the Novel. [7]

In the age of open data, it is tempting to see this as being, in essence, a small-sample analysis lacking in statistical power.[8] But such data-centric criticism of traditional humanities arguments can be a form of category error. Humanities research is as a rule more about interpretation than solution. It is about why you understand something the way you do rather than why something is the way it is. It treats its sources as examples to support an argument rather phenomena to be observed in the service of a solution. While Watt's title, "The Rise of the Novel," can be understood as implying a historical scope that his sample cannot support, his subtitle, "Studies in Defoe, Richardson, and Fielding," shows that he actually was making an argument about the interpretation of three canonical authors based on his understanding of the novel's early history - an understanding that by definition always will be provisional and open to amendment.

The real challenge for the humanities in the age of digital open data is recognizing the value of both types of sources: the material we can now generate algorithmically at previously unimaginable scales and the continuing value of the exemplary source or passage. As the raw material of humanities research begins to acquire formal qualities associated with data in other fields, the danger is going to be that we forget that our research requires us to be sensitive to both object and observation, datum and captum, finch and note. In asking ourselves what we can do with a million books[9], we need to remember that we remain interested in the meaning of individual titles and passages.

1 Borgman, Christine L. 2007. Scholarship in the Digital Age: Information, Infrastructure, and the Internet. Cambridge, Mass: MIT Press.
2 Consultative Committee for Space Data Systems. 2012. "Reference Model for an Open Archival Information System (OAIS)." CCSDS 650.0-M-2. NASA. http://public.ccsds.org/publications/archive/650x0m2.pdf.
3 National Research Council. 1999. Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases. Washington: National Academies Press. http://public.eblib.com/choice/publicfullrecord.aspx?p=3375284.
4 Jensen, H. E. 1950. "Editorial Note." In Through Values to Social Interpretation: Essays on Social Contexts, Actions, Types, and Prospects, vii – xi. Sociological Series. Duke University Press.
5 Kitchin, Rob. 2014. The Data Revolution. Thousand Oaks, CA: SAGE Publications Ltd.
6 Casadevall, Arturo, and Ferric C. Fang. 2012. "Winner Takes All." Scientific American 307 (2): 13. doi:10.1038/scientificamerican0812-13.
7 Watt, Ian P. (1957) 1987. The Rise of the Novel: Studies in Defoe, Richardson, and Fielding. London: Hogarth.
8 Jockers, Matthew L. 2013. Macroanalysis : Digital Methods and Literary History. Urbana, IL: University of Illinois Press.
9 Crane, Gregory. 2006. "What Do You Do with a Million Books?" D-Lib Magazine 12 (3). doi:10.1045/march2006-crane.
10 Marche, Stephen. 2012. "Literature Is Not Data: Against Digital Humanities." Los Angeles Review of Books, October. https://lareviewofbooks.org/article/literature-is-not-data-against-digital-humanities/.

Appendix

Figure A - Demographics of respondents

Total respondents n = 2061

Africa	20
Asia	283
Australasia	52
Europe	576
North America	418
South America	81
Grand Total	1,430

Female	342
Male	1,027
I'd prefer not to say	56
Transgender	5
Grand Total	1,430

Under 18	2
18 - 24	40
25 - 34	560
35 - 44	410
45 - 54	222
55 - 64	125
65 or over	71
Grand Total	1,430

Arts & Humanities	44
Astronomy & planetary sc..	15
Biology	402
Business/Investment	14
Chemistry	64
Earth & Environmental Sci..	110
Engineering	135
Materials Science	51
Medicine	225
Physics	88
Social Sciences	123
Other (please specify)	159
Grand Total	1,430

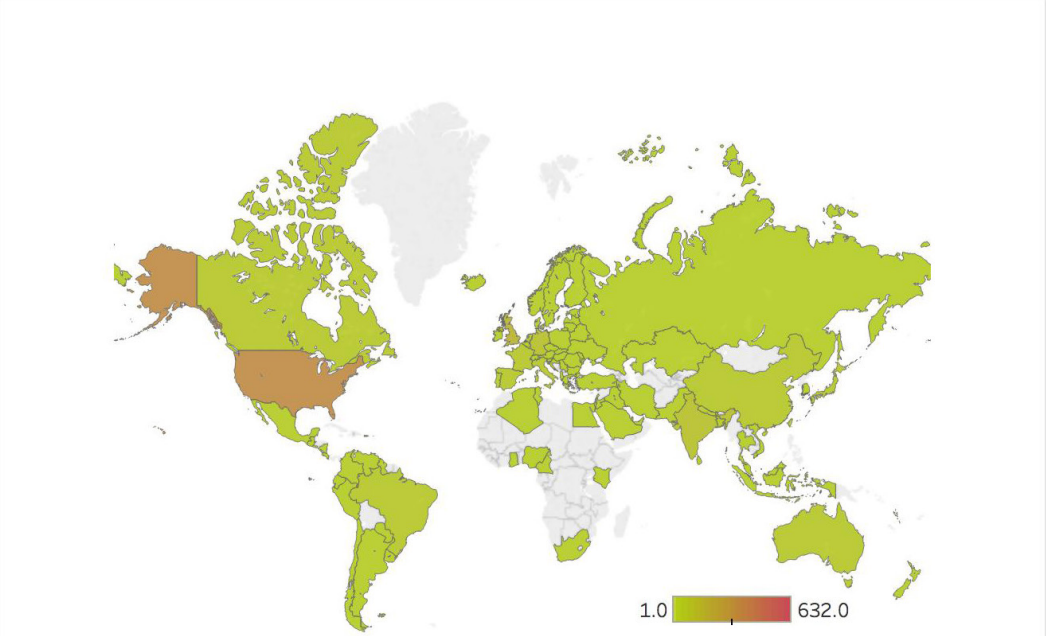
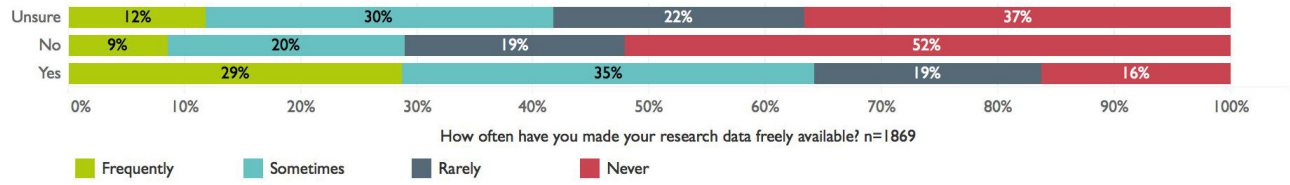


Figure B - Are you aware of freely available research data?



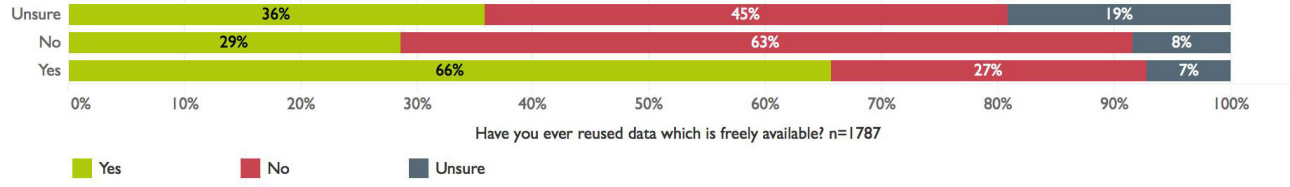
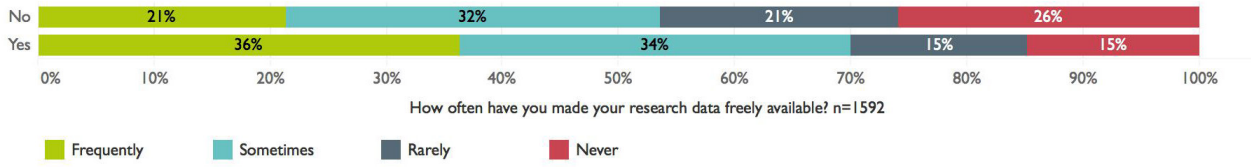


Figure C - Have you previously prepared a Data Management Plan?



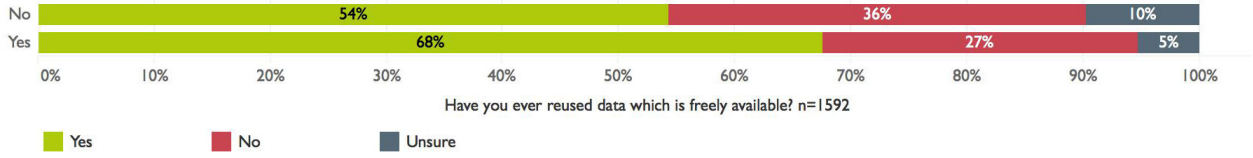
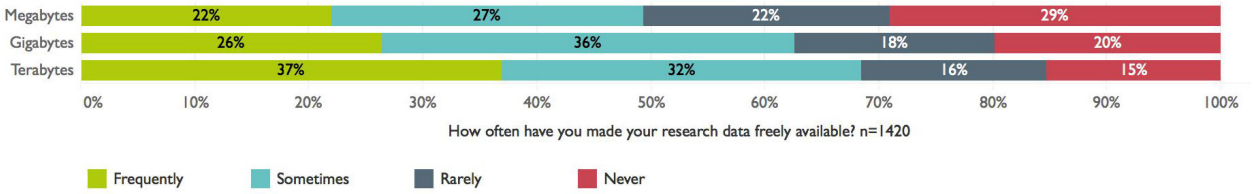


Figure D - How large is the total data associated with your typical research project?



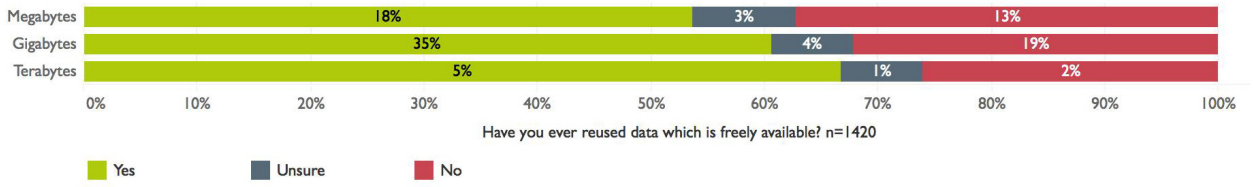
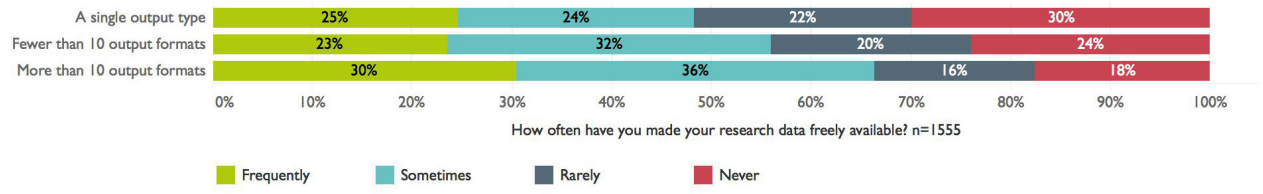


Figure E - How many different file types do you deal with?



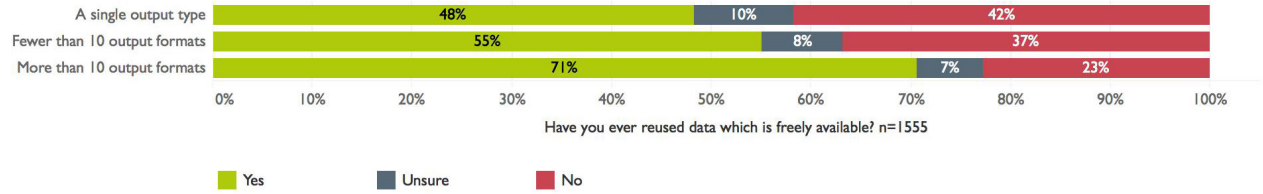


Figure F - How important is file versioning in your research process?

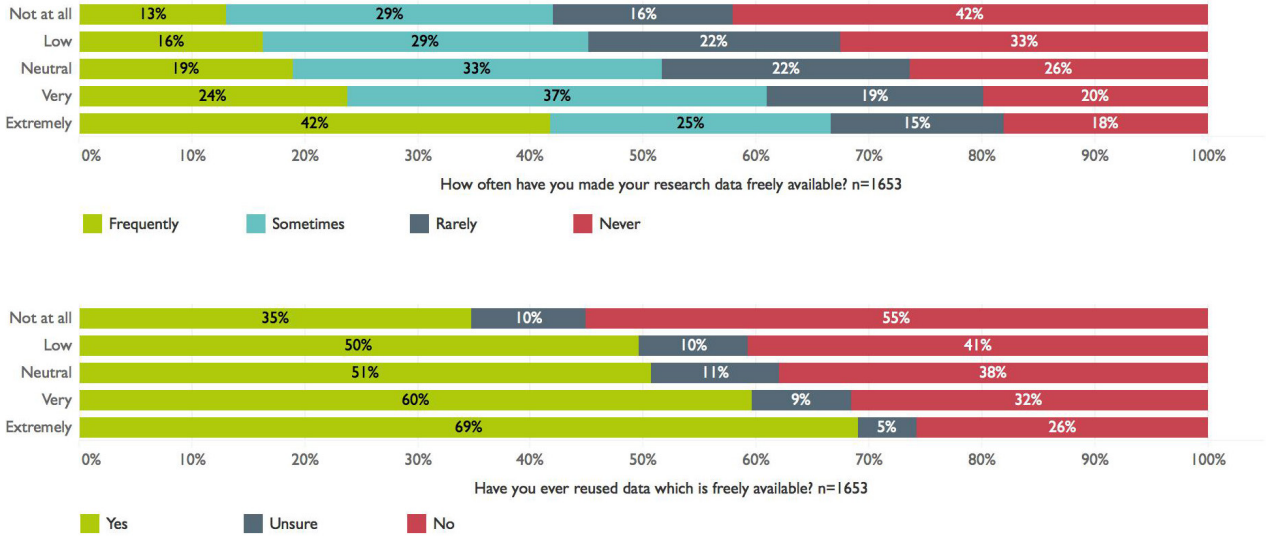


Figure G - How many files do you produce during a typical research project?

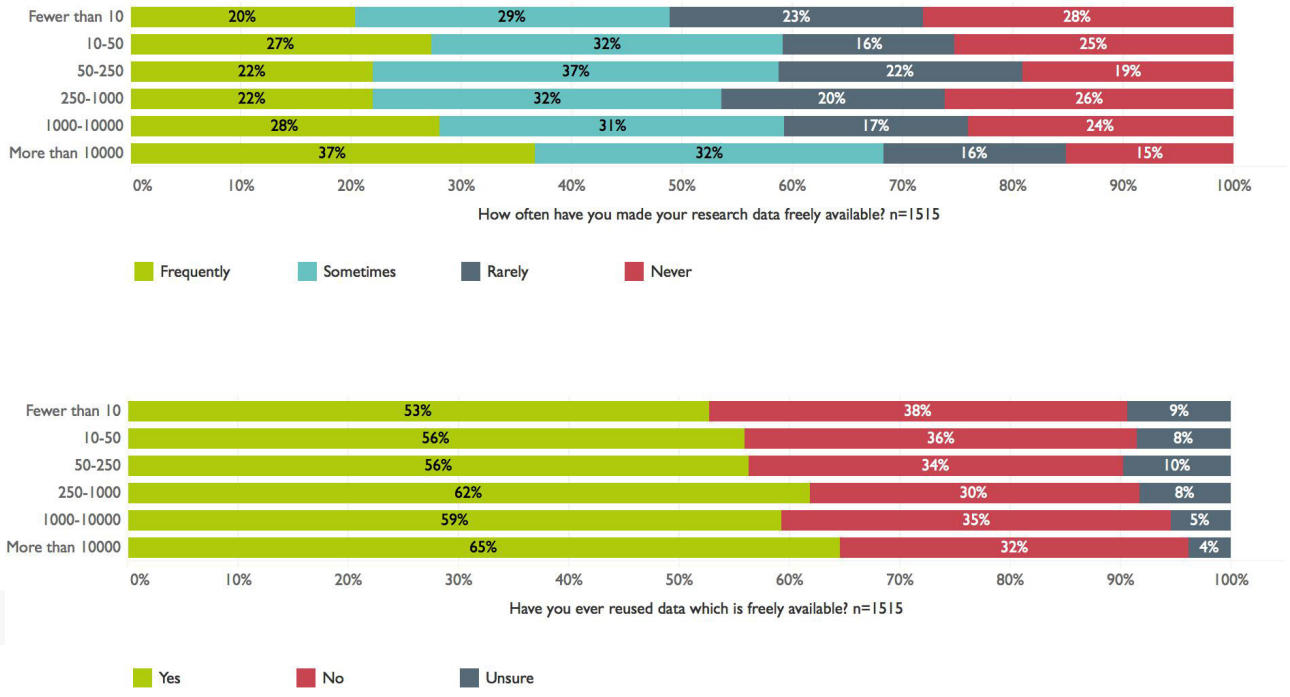


Figure H - Who would meet the costs of making your research data openly available?

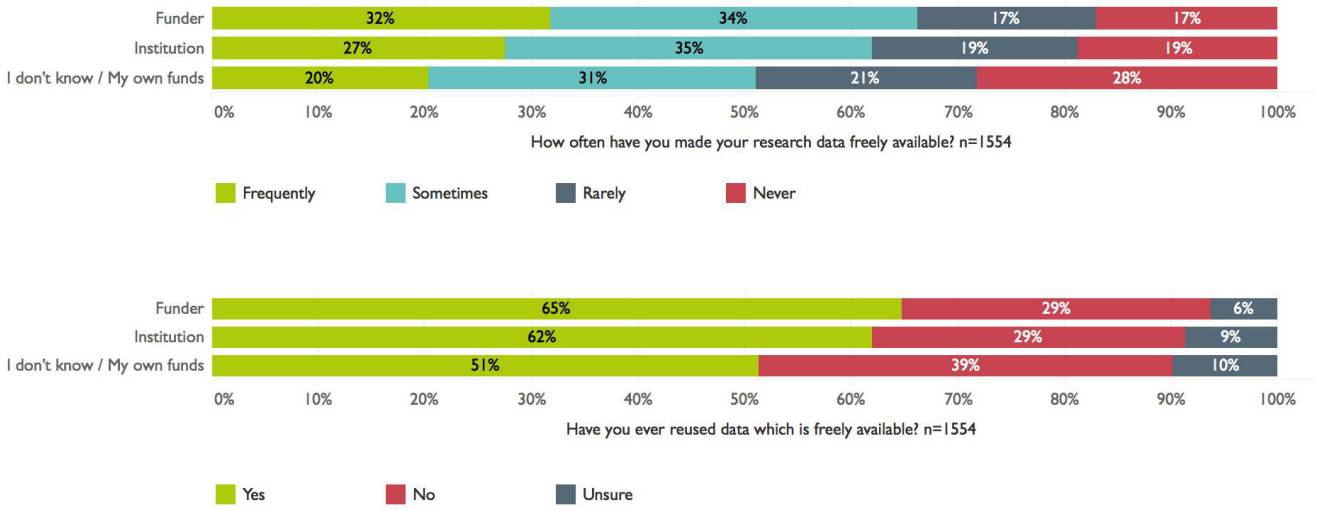


Figure I - Do you annotate your data?

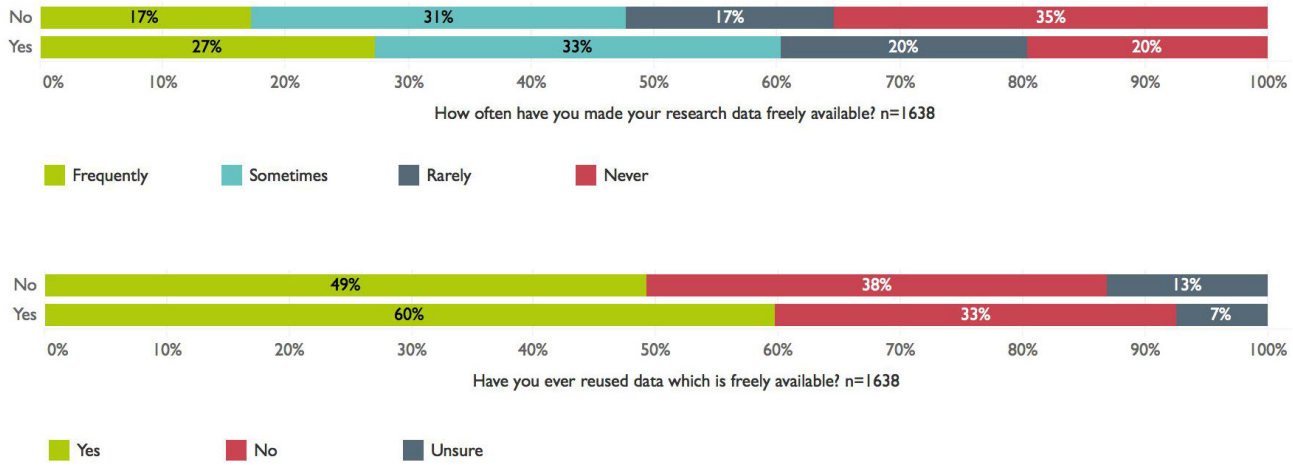
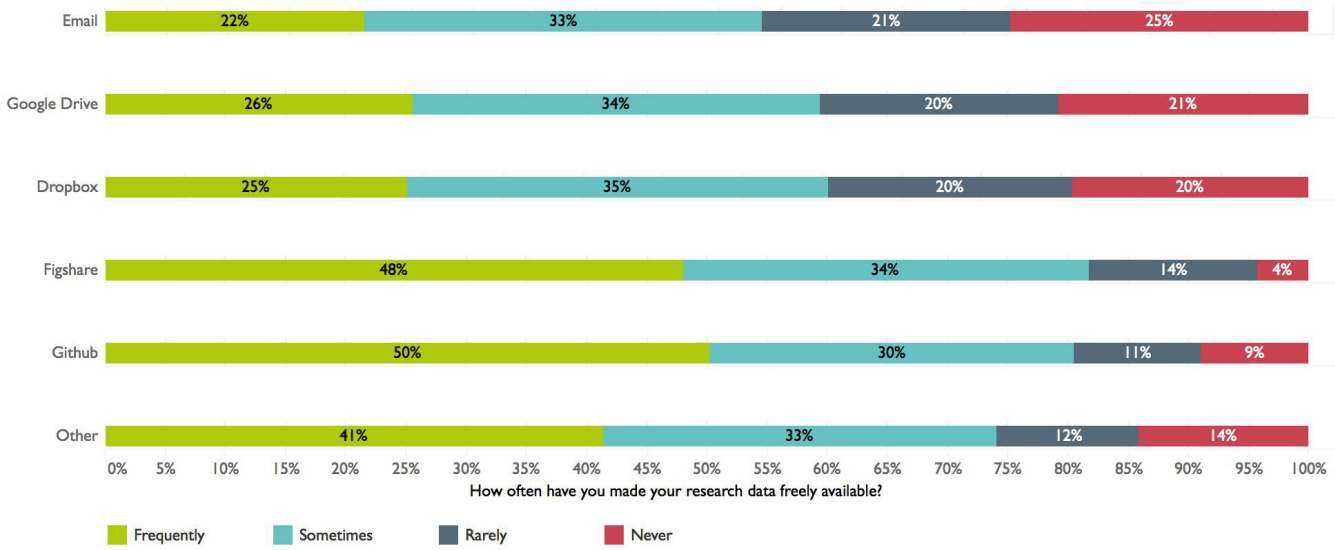


Figure J - What tools do you use to share data? (n=1561, Email n=980, Google n=505, Dropbox n=795, Figshare n=235, Github n=235, Other n=227)



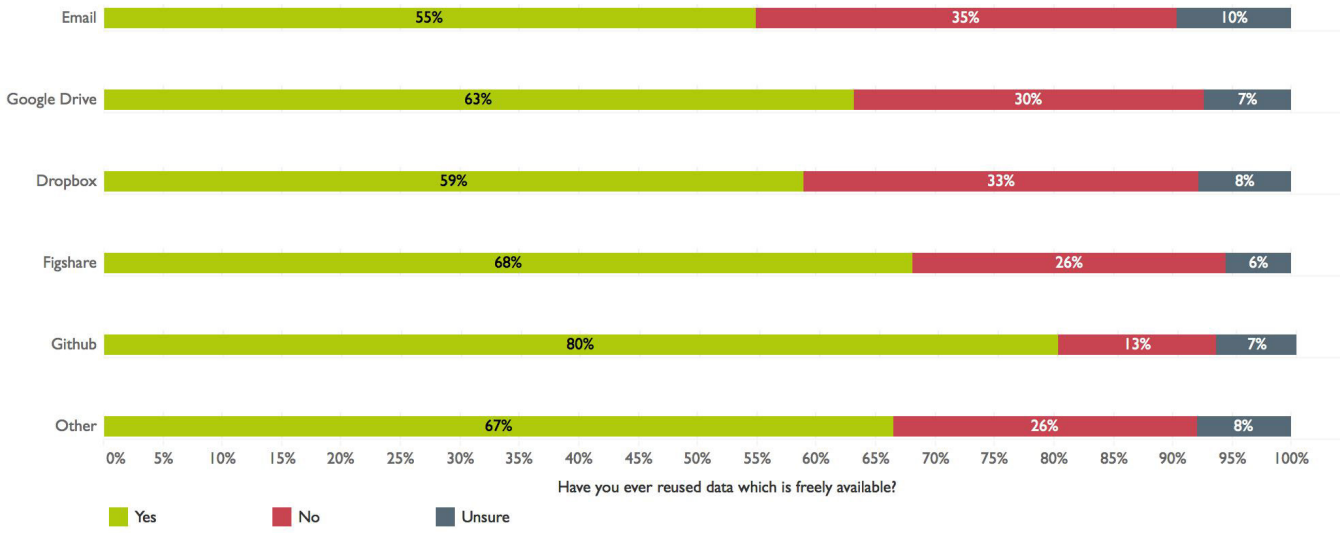
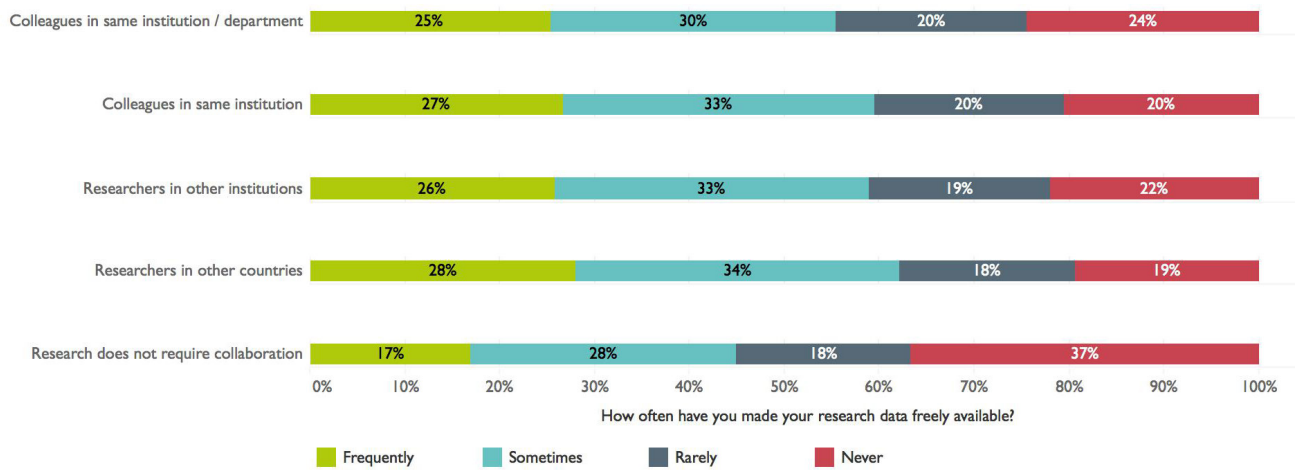


Figure K - Does your research involve collaboration with any of the following groups? (n=1688)



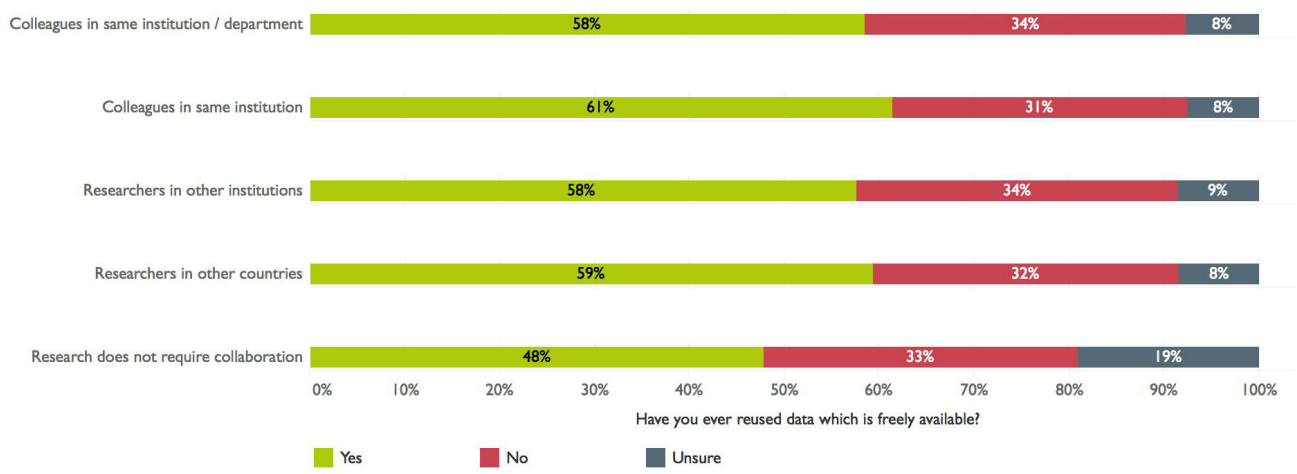
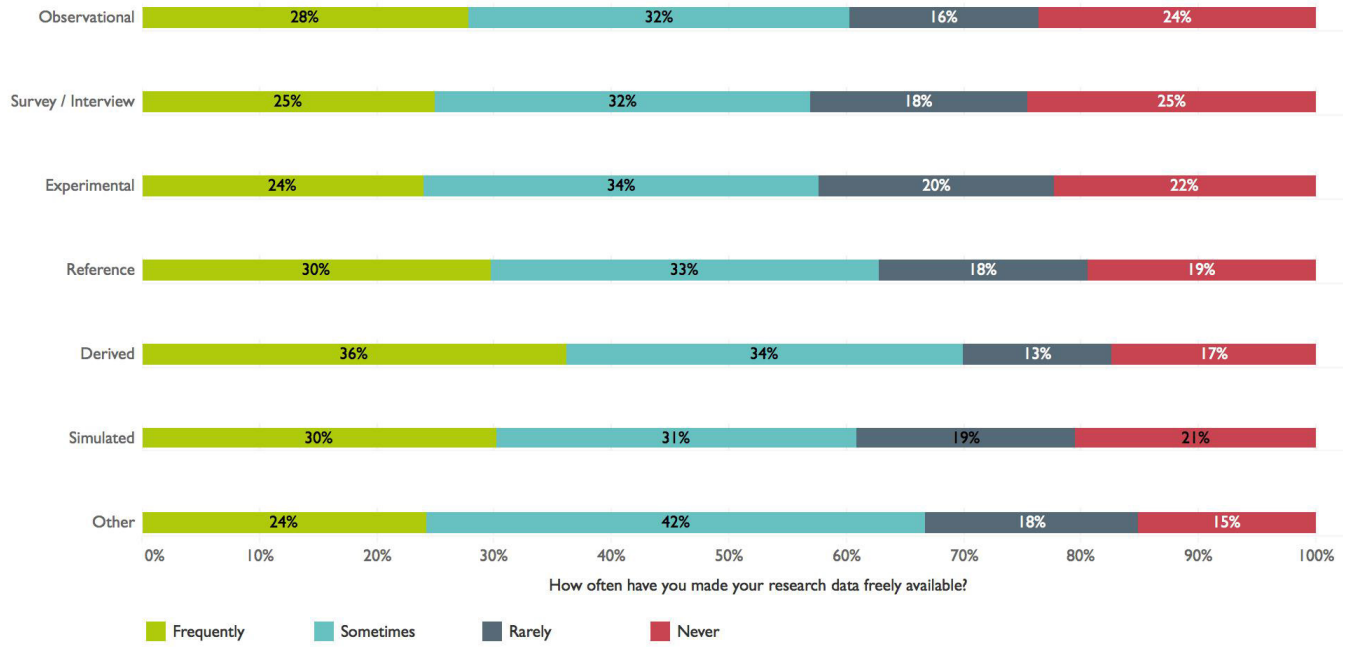


Figure L - Which of the following methodologies do you use to gather data? n=1676



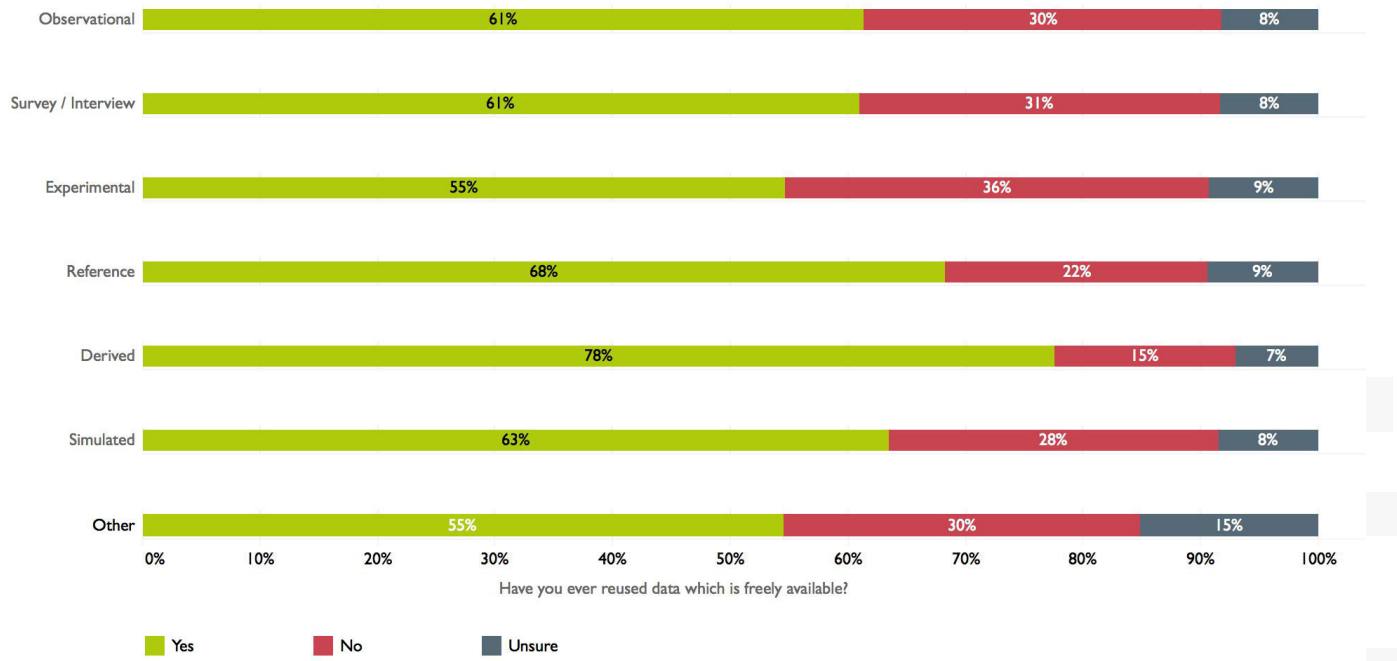
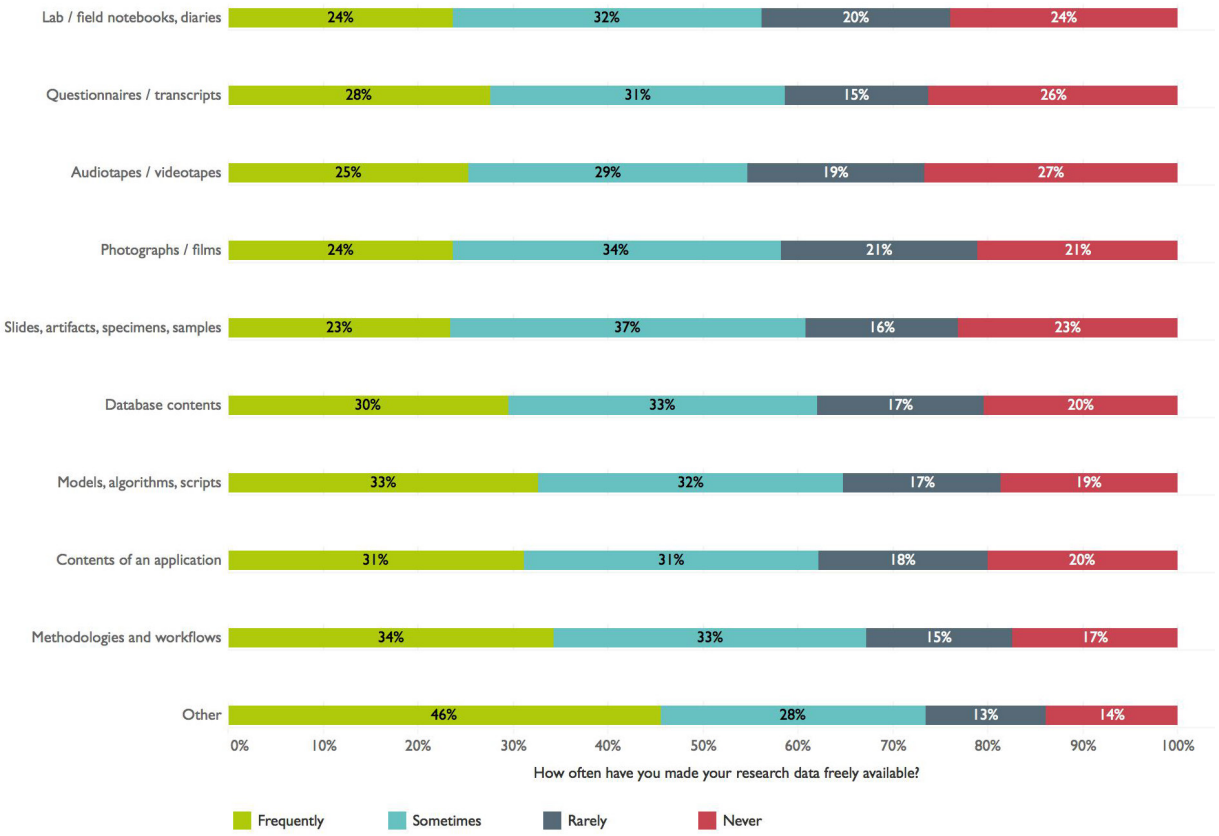
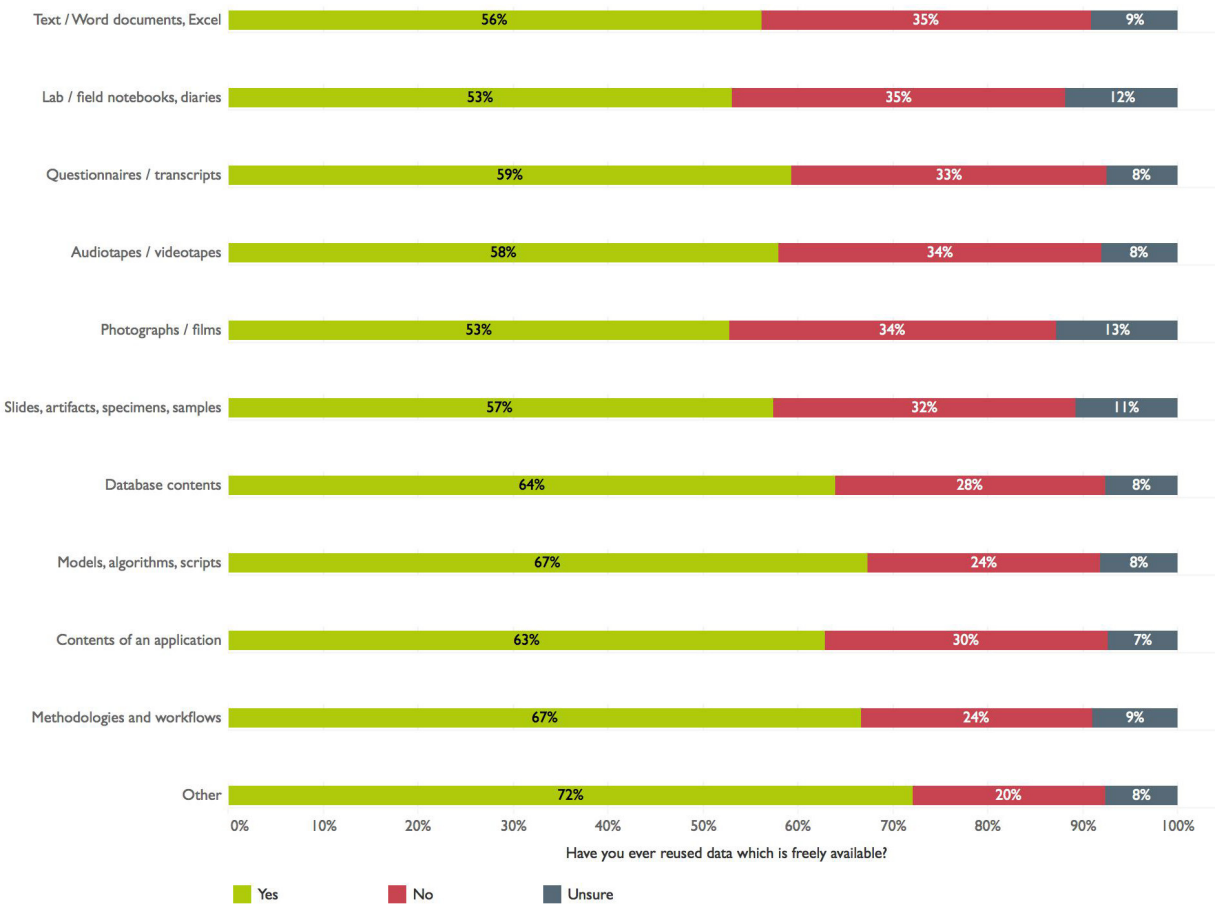


Figure M - Which of the type of data do you produce? n=1667





Contributor Biographies:

Sir Nigel Shadbolt is Professor of Computer Science at the University of Oxford and Principal of Jesus College. He is also the Chairman and Co-Founder of the Open Data Institute (ODI). In 2009, Sir Nigel was appointed Information Adviser to the UK Government, helping transform public access to Government information, including the widely acclaimed data.gov.uk site. In 2006 he was one of three founding Directors of Garlik Ltd, which was awarded Technology Pioneer status by the Davos World Economic Forum and won the prestigious UK national BT Flagship Award. In 2013 he was awarded a Knighthood for services to science and engineering. He is a member of the GDS (Government Digital Services) Advisory Board. In 2015 the Chancellor asked him to Co-Chair the UK French Data Taskforce.

Malick Tapsoba is an Open Data Specialist. He culminates seven years of leadership experience at ANPTIC (www.anptic.gov.bf) where he works currently as Deputy Manager of the Burkina Open Data Initiative (BODI) since 2013. He successfully engaged Government bodies and Civil Society Organizations in BODI. Under his leadership, BODI achieved the "Open Election" project which strongly contributed to the November 2015 peaceful presidential election in Burkina Faso after 27 years of dictatorship. Malick is also the project manager of ISOC-BF and a member of NextGen@ICANN. He holds a masters degree in international e-services.

Daniel O'Donnell is Professor of English at the University of Lethbridge. His main research interests include Digital Humanities, Scholarly Communication, Old English language and literature, the history of the book, editorial and textual scholarship, and reception-oriented criticism.

David Groenewegen is the Director, Research at Monash University, Australia. David is responsible for Library client services to the science, technology, engineering and medicine disciplines at the University, as well as the contribution the Library makes to the University's research activity. This includes oversight and development of the institutional repository and Monash University Publishing. He is also the University's research data management strategy lead. David spent four years as a Director of the Australian National Data Service, where he was involved with the development and implementation of data management solutions across the Australian university sector.

Natalie Meyers is a Partnerships and Collaborations manager at the Center for Open Science (cos.io) during a part-time leave from her faculty role as an E-Research librarian at the University of Notre Dame's Digital Initiatives and Scholarship unit in the Hesburgh Libraries. Natalie devotes a significant part of her time as an embedded data librarian and served as the Vector-Borne Disease Network digital librarian for the past three years. She is a member of senior personnel for the NSF funded Data and Software Preservation for Open Science (daspos.org) project.

Dr Sabina Leonelli is an Associate Professor in the College of Social Science and International Studies at Exeter University. Sabina is the Co-Director of the Exeter Centre for the Study of the Life Sciences (Egenis), where she leads the Data Studies research strand. Her research spans the fields of history and philosophy of biology, science and technology studies and general philosophy of science. Her current focus is on the philosophy, history and sociology of data-intensive science, especially the research processes, scientific outputs and social embedding of Open Science, Open Data and Big Data.

Nobuko Miyairi is Regional Director, Asia Pacific for ORCID (http://orcid.org), based in Tokyo, Japan. Nobuko builds relationships with stakeholders across Asia Pacific to expand ORCID adoption and awareness in the research community. Prior to joining ORCID, Nobuko held positions at Thomson Reuters and Nature Publishing Group, where she worked closely with research organizations, government policy makers and funding bodies to provide research management solutions. A librarian by training, Nobuko earned an MLIS from the University of Hawaii at Manoa.

Dr Kazuhiro Hayashi is Senior Research Fellow at the National Institute of Science and Technology Policy, Japan. Kazuhiro has been involved in scholarly publishing and communication, in a wide variety of roles, for more than 20 years. At the Chemical Society of Japan (CSJ), he worked successively as Editor, Production Manager, E-journal Manager, and Promotions Manager. Throughout his broad range of roles in publishing he has focused on scholarly communication through E-journals, and he has used his IT skills to reconstruct and improve the way publishing is managed. In 2012 he moved from CSJ to the National Institute of Science and Technology Policy (NISTEP), where he is engaged in a study to provide evidence to develop a Science and Technology policy for administrators and policy makers, now focusing on the future of Scholarly Communication including Open Science and alternative impact assessments.

Dr Mark Hahnel is founder and CEO of Figshare, London. He is passionate about open science and the potential it has to revolutionise 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.

Heather Joseph is the Executive Director of SPARC, (Scholarly Publishing and Academic Resources Coalition), a coalition of academic and research libraries expanding the open communication of scholarship. SPARC supports new models for sharing digital articles, data and educational resources, and is widely recognized as the leading force for Open Access advocacy. Heather is an active participant on committees and projects at several U.S. federal agencies. In 2015, she was appointed to the newly formed Commerce Data Advisory Council and tasked with providing input to the Secretary of Commerce on issues surrounding open data.

Dr Till Bruckner is the AllTrials campaign manager and **Beth Ellis** is a PhD student currently doing an internship at Sense about Science, the charity that runs the AllTrials campaign. Readers interested in the campaign can visit the AllTrials website. For a highly readable in-depth discussion of missing trials and evidence distortion in medicine, we recommend the book "Bad Pharma" by AllTrials co-founder Dr Ben Goldacre.

Dr Daniel Hook has been Managing Director of Digital Science since July 2015. He has been involved in research information management, open access, open data and software development for more than a decade, holding positions as Director of Research Metrics at Digital Science, Founder and CEO of Symplectic and COO of Figshare. By training he is a mathematical physicist specialising in quantum theory. Daniel holds visiting positions at Imperial College London and Washington University in St Louis and is a Fellow of the Institute of Physics.

Jon Treadway is Director, Operational Strategy, Digital Science, London, UK

Dr Briony Fane is Metrics Researcher, Digital Science, London, UK

Dan Penny is Head of Market Intelligence: Researchers and Audience, Springer Nature, London, UK

Anna Gallagher is Research Analyst, Strategy & Market Intelligence, Springer Nature, London, UK

Laura Wheeler is Community Manager at Digital Science, **Lisa Hulme** is Consulting Director of Communications at Digital Science, **Julia Giddings** is Global Marketing Manager at Digital Science and **Alan Hyndman** is Marketing Manager at Figshare.

DIGITAL science

Work smart. Discover more.

Part of the **Digital Science** family

figshare

GRID

Dimensions

SYMPLECTIC Elements

Altmetric

readcube

labguru

BIORAFT

Overleaf

TETRASCIENCE

digital-science.com