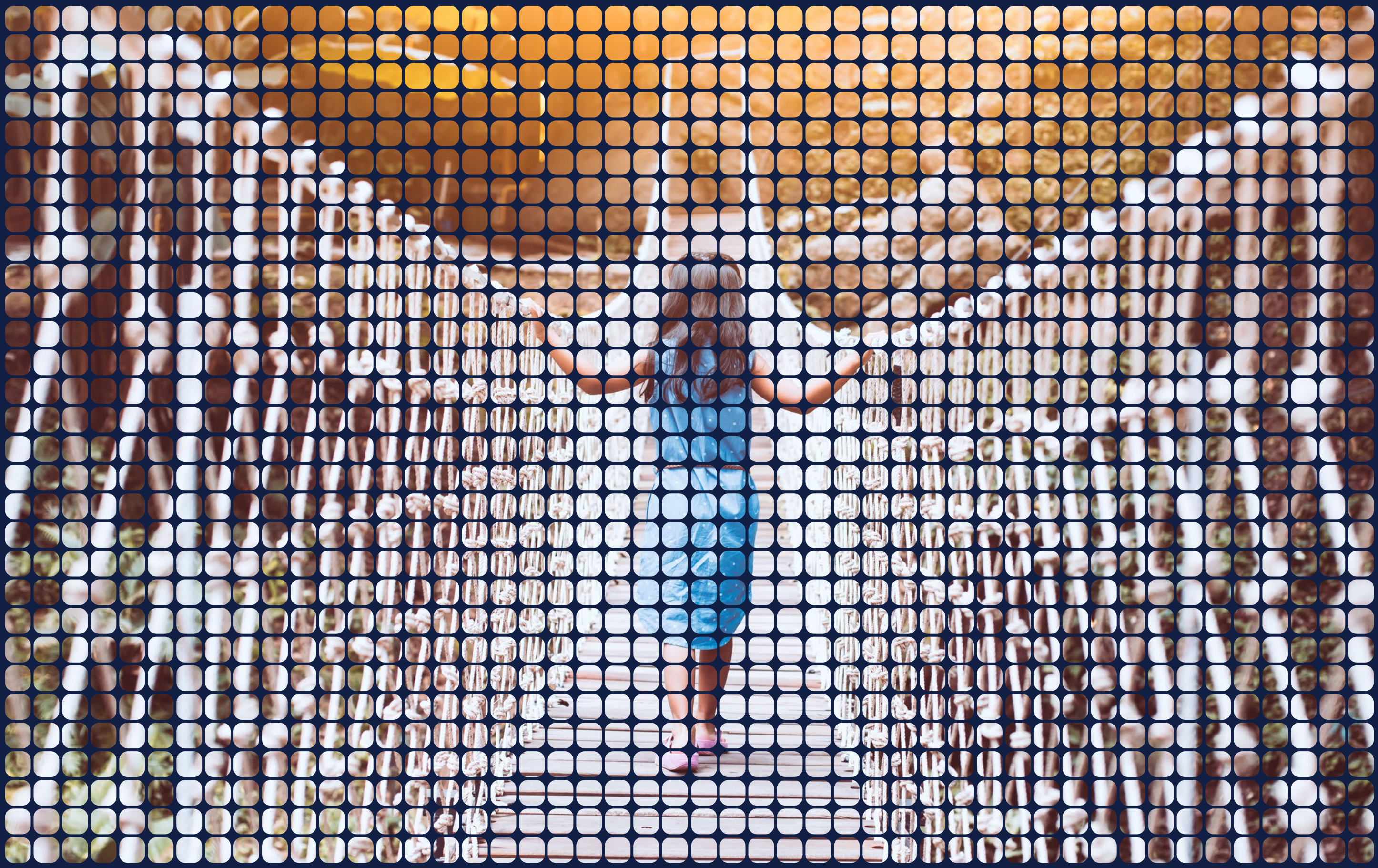

Digital Science Report

The State of Trust & Integrity in Research

Perspectives on data sharing, policies and practices

AUTHORS

Leslie McIntosh, PhD, MPH; Edit Herczog, MS; Lyric Jorgenson, PhD; Gerardo Machnicki, PhD; Taunton Paine, MA; Dina Paltoo, PhD, MPH; August DeVore, MS; Josh Sumner, MS; Cynthia Hudson Vitale, MA

September 2022

DIGITAL science

ripeta

About Digital Science

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

About Ripeta

Ripeta's mission is to assess, design, and disseminate practices and measures to improve the reproducibility of science with minimal burden on scientists, starting with the biomedical sciences. We focus on assessing the quality of the reporting and robustness of the scientific method rather than the quality of the science. Our long-term goal includes developing a suite of tools across the broader spectrum of sciences to understand and measure the key standards and limitations for scientific reproducibility across the research lifecycle and enable an automated approach to their assessment and dissemination. Visit www.ripeta.com

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

Digital Science, 6 Briset Street, London, EC1M 5NR, UK. info@digital-science.com
Ripeta, C/o Digital Science, 625 Massachusetts Avenue 2nd Floor, Cambridge, MA, USA 02139.
info@ripeta.com

DOI: 10.6084/m9.figshare.
xxxxxxxxxx

Contents

Foreword

Hilary Hanahoe, Secretary General, Research Data Alliance (RDA)

In the Research Data Alliance's nascent years and indeed during its conception (2010-2012), terms like "open science", "open research", and "FAIR data" had not yet been coined. Nevertheless, these concepts rested in the forefront of the RDA funders' and founders' minds: the necessity to support and facilitate the digital data era; to address the deluge of data being produced, making it available, accessible and reusable; and the need for cross-pollination and coordination of experts across the globe.

Trust and integrity were very much the ideals and tenets on which these conceivers laid the RDA foundations, and still today they are core to the organization and the community. The vision is that researchers and innovators can openly share and reuse data across disciplines, across technologies and across countries to solve the grand challenges of society. Solving these grand challenges requires excellent science, excellent research, excellent data management. Data are key but without trust and integrity these ambitious goals will not be achieved.

The Research Data Alliance itself is born from the concept that trust is key to collaboration and cooperation. Indeed, its guiding principles of openness, transparency, consensus and community driven, inclusivity and technology neutral were carefully defined and are continuously maintained as a commitment to facilitating trust building.

In the nine years that RDA has been operating, the growth of policies and strategies issued by funders, governments, institutions, publishers, etc. to make research outputs more open and research practices more inclusive and collaborative has been considerable. It is the centrepiece of many institutional, national and global recommendations and strategic thinking. If we demand trust and integrity as cornerstones of these policies and strategies, how do we ensure they are preserved and upheld? What is the reward? What is the incentive? What is the prize? What is the game changer? How can you convince research organizations and researchers and scientists to invest in workflows and practices that contribute to open research and open science? New policies require organizations to have appropriate infrastructure and services, researchers and scientists to change their culture, change their practices and open their results. But that is a huge responsibility and burden for them. And even the most willing and able encounter barriers, especially when the statistics are stacked against them. Sources say that data scientists invest approximately 45% of their time in data preparation. What about the many institutions and organizations that do not have access to data scientists, either in-house or outsourced? Then data wrangling is a task requested of the researchers themselves and hence their time invested in "data housekeeping" is time not invested in the research itself.

The open research ecosystem needs to have a plethora of infrastructures, tools, skills and expertise available to facilitate the implementation of open research and achieve the integrity goal. This means investment in skilled expertise, training and infrastructure to support. This means a new way of assessing research and researchers' careers to encompass the open research ask. This means focusing on quality and reproducibility.

Our actions today and the lessons we can and should learn from the past can and must shape our future and that of many generations to come. This is a collective responsibility of all stakeholders. We are the ones that can make this happen. We have a duty to build on this progress and are all key to advancing the open science reality and building the future. This means ensuring that trust and integrity are the central and permanent foundations upon which this future is constructed.

> "The open research ecosystem needs to have a plethora of infrastructures, tools, skills and expertise available to facilitate the implementation of open research and achieve the integrity goal."

Executive Summary

This inaugural report, ***The State of Trust & Integrity in Research***, discusses key issues of research integrity that relate to funding agencies and their role in fostering an ecosystem of trust in research. Specifically, the report delves into funding agency policies and how they translate into practice through publications.

Produced by Ripeta, a Digital Science company dedicated to supporting and building trust in science, ***The State of Trust & Integrity in Research*** comprises contributions from a range of experienced authors, including from the National Institutes of Health (NIH), other external commentators, and Ripeta itself.

The report addresses :

- fostering public trust in science,
- improving integrity in research,
- policies, data sharing, and open access practices,
- the roles of key stakeholders in fostering research integrity.

Section 1: Improving Research Integrity: Policies and Perspectives

Several key funding agencies have either announced or implemented policies directly aimed at ensuring research integrity is evaluated, supported and improved.

The first section contains two articles. The foundational first piece contextualizes the issues and hinges on three key roles for trust and integrity in research: a Motivator, an Enabler, and an Outcome. Potential benefits for developing countries – particularly in the age of COVID-19 – are examined.

This section also includes a second article from experts based at the National Institutes of Health (NIH), a major US federal funding agency that has made significant commitments towards open science, including requirements for researchers receiving NIH funding. They explain the NIH's organizational-wide policies and how appropriate data stewardship contributes to public trust in science.

Section 2: Funding Agency Data Management and Data Sharing Policies – Analysis and Comparison

The second section considers how policies are put into practice at a range of funding agencies. Exclusive analyses by Ripeta compares the approach taken by five major world funders.

Finally, the report presents a case study and analyses of the practices and policies for open research at the global charitable foundation Wellcome, particularly as a lens for evaluating impact.

Recommendations

While policies can be critical for advancing trust and integrity in research, it is not enough to simply have policies in place.

Together, these articles not only accentuate the critical role funding agencies and other research stakeholders have in improving research integrity, they also highlight discrepancies between the various funding bodies' policies and practices. Examples include: different approaches to use of data management plans, data repositories, data retention, and different levels of research papers published in open access (OA) journals.

Ultimately, the report recommends the need for more standardized, coordinated policies among funding agencies – at least at the national level – to comprehensively address research integrity. Central to these efforts should be a focus on incentives for researchers and institutions for compliance and for training and education. Fine-tuning existing policies may improve both researcher compliance and science.

"In this inaugural report on the State of Trust & Integrity in Research, we delve into funding agency policies as they have been translated through research and into practice. These policies have been developed to increase quality research outputs and thus the trust in scientific research, and this report illuminates the impact of said policies on practice."

State of Trust & Integrity

Research dynamics have changed with technological advancements, greater scholarship transparency, and global events. The walls of science have become more porous, with the flow of conducting and communicating scholarship quickly traversing the scholarly, public, and policy ecosystems. This newfound permeability offers rapid solutions to grand challenges to improve quality of lives. This porosity can also undermine scientific integrity without adequate structural support and quality control.

The State of Trust & Integrity in Research report has culminated from rich partnerships and collaborations built by an academic researcher and a data librarian concerned about making science better and better science easier. We have lived through the reproducibility 'crisis' and now delve into research integrity awareness. Key stakeholders – funding agencies, publishers, institutions, and researchers – have grown as a community to course correct the scholarly ecosystem's direction as we rebuild an infrastructure to transport science into the next era. This includes the commitment of transparent practices and the obligation of greater scholarly representation – from researchers to publishers, to institutions, to funders.

In this inaugural report on the State of Trust & Integrity in Research, we delve into funding agency policies as they have been translated through research and into practice. These policies have been developed to increase quality research outputs and thus the trust in scientific research, and this report illuminates the impact of said policies on practice.

The open science movement complements funder policies to achieve reliability and trust of research through transparency, openness, and reproducibility. An initial push for more thorough research reporting using funding statements and data availability statements has changed the requirements for researchers, institutions, and publishers to increase their transparency of not only the science but of the support and potential conflicts with the research. Yet, easily and robustly checking how policies have been put into practice in the landscape of open science has been difficult until now.

We first look to science and policy experts from government and industry to discuss trust and integrity as science moves to open science in two guest pieces:

- *Three Roles For Trust And Integrity In Open Science With Specific Implications For Developing Countries And Other Disadvantaged Agents* — analysis from an industry researcher in Latin America; and,
- *Fostering Public Trust In Science Through Policy And Data Stewardship* — perspectives from professionals at the US National Institutes of Health.

These perspectives provide broad discussions around trust and integrity of science and the challenges faced by funders. To contextualize these thoughts, we focus on two objects of a larger picture: data management and data sharing. We seek to answer two questions through analyses of policies and publications:

1. **What are the established data management and data sharing policy practices for funding agencies?** Experts at Ripeta dove into funding agency policies to understand the requirements of data management and data sharing practices.

2. **How do funder data policies translate into practice?** From more than 60 policies, we take a deep dive into data availability statements and data sharing within the publications from five of the funding agencies – using Ripeta's automated checks to analyse the adherence to policy guidelines and report the adherence to policy guidelines.

Next, we present a case study with the Wellcome Trust. After their leadership in creating open science policies, the Wellcome Trust then assessed the adherence in publications, producing interesting findings particularly on data sharing.

Lastly, we present a third guest piece in the conclusion: *Trust In Open Science Is Necessary But Not Sufficient For Society To Support Science* — reflections from a consultant and former European Union (EU) parliamentarian.

Between expert insights and Ripeta analyses, this report offers a peek into trust and integrity in research through the lens of funder policies. The infrastructure provided from funding agency policies will serve to strengthen the research ecosystem, provided we ensure compliance with these crucial support requirements. They may not be sufficient for complete quality control in this new era of science, but they are necessary for the improvements and fortification of research integrity.

Cheers,
Leslie D. McIntosh, PhD
Founder and CEO, Ripeta

> "Between expert insights and Ripeta analyses, this report offers a peek into trust and integrity in research through the lens of funder policies."

"The two articles that make up this section of the report provide an example of how one US federal agency is approaching enhancing trust in science through policies and requirements, as well as a foundational article on three roles for trust and integrity in research."

Improving Research Integrity: Policies and Perspectives

Internationally, funding agencies have implemented policies and requirements for the responsible management and sharing of research outputs as a mechanism to enhance trust in science. The two articles that make up this section of the report provide an example of how one US federal agency is approaching enhancing trust in science through policies and requirements, as well as a foundational article on three roles for trust and integrity in research.

The first article in this section, written by Gerardo Machnicki, PhD, MSc, ***Three Roles for Trust in Open Science: A Motivator, an Enabler and an Outcome***, contextualizes trust and integrity in open science. Generating and applying meaningful research is fundamental to promote human progress and sustainable development. There is a growing emphasis on improving the quality and impact of research through open science approaches. It is frequent to find concepts such as trust and integrity in considerations about open science. The main message of this contribution is to highlight three roles for trust and integrity: as a motivator, as an enabler, and as an outcome.

The second article in this section, ***Fostering Public Trust in Science through Policy and Data Stewardship***, is authored by researchers and administrators from the US National Institutes of Health (NIH). The authors describe the policies and requirements that NIH has put in place to enhance trust in NIH-funded research.

Three Roles for Trust in Open Science: A Motivator, an Enabler and an Outcome

Implications for Developing Countries and Other Disadvantaged Agents in the Age of COVID-19

Gerardo Machnicki, PhD, MSc

Increasing trust in scientific research is a major objective of open science, reflecting the motivation to overcome “fake news in the post-truth age” (Stracke 2020) and to address the known reproducibility crisis in various scientific disciplines (Ioannidis, 2005; Ioannidis, 2008). This emphasis on integrity and motivation to do research in an utterly honest way is foundational to open science, and key to practices that foster transparent, open and replicable research. However, a diverse and complex set of “openings” is necessary to build open science to its full potential: from conceptualization and planning (“open discourse”), to protocols and analysis plans, to code, data and findings. Progress requires the emergence and positive evolution of initiatives, processes and platforms that support open science, and that champion reliability, integrity and trust in every stage.

Open science is advocated not only because it is expected to produce higher quality science and enhance trust in research, but also on the grounds of democratising knowledge, empowering

"The participation, voice and practical footprint of LMICs in open science at large continues to evolve, and to interrogate the assumptions and operations of open science as understood from the Global North."

disadvantaged geographies and researchers, and putting emphasis on populations in special need. Trust is what enables the engagement of these communities. While the equity-enhancing potential of open science has particular promise in its practice in low and middle income countries (LMICs), researchers and impacted communities often have considerable reservations. In theory, less developed countries have many benefits to derive from research alliances involving findable, accessible, interoperable and reusable (FAIR) data. In practice, there are often ethical and cultural concerns towards this type of open data sharing (Serwadda et al., 2018). Fears associated with open data approaches among LIMC researchers include "...data misuse, violations of patient privacy through participant reidentification, and possible humiliation and exploitation of the researchers themselves" (Serwadda et al., 2018); "fears over free-riding scientists using the data collected by others for their own career advancement" (Serwadda et al., 2018); and concerns over "risks of undermining originating researchers' professional development and challenges in accessing the resources needed to support data sharing" (Jao et al., 2015). Lower willingness to engage in data sharing has been correlated with diminishing shares of investment in research as a fraction of the income of a country (Damalas et al., 2018), although the strength of this association and the direction of causality are uncertain, and may require further study. But despite legitimate concerns, there is hope and optimism towards open science in LMICs that has developed into novel initiatives and numerous positive examples of open science practice.

Research alliances working in LMICs can have a varied record in building trust with collaborators and in generating trustable, impactful research on open science platforms – as exemplified by the WorldWide Antimalarial Resistance Network (WWARN) (Pisani et al., 2016; Humphreys, Tinto and Barnes, 2019). WWARN is an "investigator led network of 260 collaborators, most performing clinical trials related to malaria drug efficacy and resistance in endemic countries" (Pisani et al., 2016), with open science components that provide valuable lessons about both technical enablers (e.g., data standards and curation) and the importance of trust-building. The WWARN collaboration began with an end goal that was not grounded in the needs of impacted communities, and local researchers were therefore wary about the motives for the alliance. Establishing trust was critical to enabling this open science initiative – a role distinct from motivating open science practice, or from trust in science serving as an end in and of itself. As this trust in the WWARN project evolved, so could the governance of the initiative.

Reports from the Open and Collaborative Science in Development Network (OCSDNet) also highlight how open science can benefit research agendas grounded in local needs, and how flexibility in the degree to which certain "openness" dimensions are managed can be key to successful outcomes. For example, in an open science project involving botanic research in Brazil, the concept of open data shifted to managed access in the evolution of the project and the process of trust-building:

> *Perhaps most surprising for the team, however, was around the complex negotiations and cultural shifts that needed to occur, throughout the years, to ensure the project's success. For instance, while preliminary requirements for data providers demanded complete openness, through a series of negotiations, the parameters have since changed to allow data providers the flexibility to decide, on their end, which records are made openly available and how... Communication, transparency and participation, according to the team, were indispensable for building trust, understanding and ownership amongst all actors.* (Hyllier et al., 2017)

A case study on promoting public health data sharing in Kenya also underscores the importance of building trust between researchers, public health practitioners, and the community, and notes in its analysis that "institutional forms of trust are likely to be strengthened by engagement and dialogue with citizens, and governance processes that include openness, solidarity, fairness, and truth-telling" (Jao et al., 2015). Truly collaborative open science requires not merely the development and design of new platforms or tools, but also a series of "complex negotiations around roles and responsibilities; principles and priorities; timelines and resources" and "reflection on how such practices may coincide with existing cultural and institutional norms" (Hillyer et al., 2017). Correspondingly, the principles named within the OSCDNet manifesto include the contribution of open science to sustainable development and that researchers be "mindful of [how] context, power and inequality can condition scientific research" (Albornoz et al., 2017).

The participation, voice and practical footprint of LMICs in open science at large continues to evolve, and to interrogate the assumptions and operations of open science as understood from the Global North. AmeliCA – a Latin American open access initiative – can be considered an innovative precursor to Plan S (Aguado-López and Becerril-Garcia, 2019), and also integrates a set of tools and platforms to foster open access to research findings (AmeliCA, no date). Another initiative within the ecosystem of Latin American open science, "La Referencia" (LA Referencia - Home, no date), was featured in the second UNESCO meeting on open science (United Nations, 2021). Many other initiatives also foster open science and infrastructure in the region, and as is true in other geographies, COVID-19 is recognised as an important motivator for the commitment to "restore and expand the bridges between science and citizens" (Babini and Rovelli, 2020).

Tensions may exist between the role of open science as a pragmatic and methodological endeavour, and its democratising and equity-enhancing potential. However, increased trust in science demands not only improved objective measures of reliability, but also a focus on trust-building that engages researchers and communities – as has been demonstrated in LMICs around the globe. When trust is prioritized as a motivator, enabler and outcome in open science, positive feedback loops evolve. Lessons learned from the COVID-19 pandemic and other global threats (e.g., climate change) demonstrate that the need to promote and actively invest in this dimension of open science is only more urgent than ever.

"Increased trust in science demands not only improved objective measures of reliability, but also a focus on trust-building that engages researchers and communities"

References

Aguado-López, E. and Becerril-Garcia, A. (2019) 'AmeliCA before Plan S – The Latin American Initiative to develop a cooperative, non-commercial, academic led, system of scholarly communication', Impact of Social Sciences, 8 August. Available at: https://blogs.lse.ac.uk/impactofsocialsciences/2019/08/08/amelica-before-plan-s-the-latin-american-initiative-to-develop-a-cooperative-non-commercial-academic-led-system-of-scholarly-communication/ (Accessed: 9 December 2021)

Albornoz, D. et al. (2017) 'Co-Constructing an Open and Collaborative Manifesto to Reclaim the Open Science Narrative', p. 12.

AmeliCA (no date). Available at: http://amelica.org/index.php/en/home/ (Accessed: 9 December 2021)

Babini, D.S. de and Rovelli, L. (2020) Tendencias recientes en las políticas científicas de ciencia abierta y acceso abierto en Iberoamérica. Ciudad Autonoma de Buenos AIres, Argentina: Fundación Carolina : CLACSO (Ciencia abierta).

Chakravarty, A. et al. (2021) Generating Actionable Insights from Real World Data - The COVID-19 Evidence Accelerator, FDA. FDA. Available at: https://www.fda.gov/science-research/fda-science-forum/generating-actionable-insights-real-world-data-covid-19-evidence-accelerator (Accessed: 9 December 2021)

Christensen, G.S., Freese, J. and Miguel, E. (2019) Transparent and reproducible social science research: how to do open science. Oakland, California: University of California Press.

Damalas, D. et al. (2018) 'Open data in the life sciences: the "Selfish Scientist Paradox"', Ethics in Science and Environmental Politics, 18, pp. 27–36. doi: https://doi.org/10.3354/esep00182

Davey, M. (2020) 'COVID-19 studies based on flawed Surgisphere data force medical journals to review processes | Coronavirus | The Guardian', 12 June. Available at: https://www.theguardian.com/world/2020/jun/12/covid-19-studies-based-on-flawed-surgisphere-data-force-medical-journals-to-review-processes (Accessed: 9 December 2021)

European Health Data and Evidence Network (EHDEN) (2020) 'EHDEN now working with 28 Data Partners across 11 countries to harmonise clinical data across therapeutic areas including COVID-19', ehden.eu, 28 July. Available at: https://www.ehden.eu/ehden-now-working-with-28-data-partners-across-1-countries-to-harmonise-clinical-data-across-therapeutic-areas-including-covid-19/ (Accessed: 9 December 2021)

Fecher, B. and Friesike, S. (2014) 'Open Science: One Term, Five Schools of Thought', in Bartling, S. and Friesike, S. (eds) Opening Science. Cham: Springer International Publishing, pp. 17–47. doi: https://doi.org/10.1007/978-3-319-00026-8_2

Haendel, M.A. et al. (2021) 'The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment', Journal of the American Medical Informatics Association, 28(3), pp. 427–443. doi: https://doi.org/10.1093/jamia/ocaa196

Hillyer, R. et al. (2017) 'Framing a Situated and Inclusive Open Science: Emerging Lessons from the Open and Collaborative Science in Development Network', p. 16.

Humphreys, G.S., Tinto, H. and Barnes, K.I. (2019) 'Strength in Numbers: The WWARN Case Study of Purpose-Driven Data Sharing', The American Journal of Tropical Medicine and Hygiene, 100(1), pp. 13–15. doi: https://doi.org/10.4269/ajtmh.18-0649

Ioannidis, J.P.A. (2005) 'Why Most Published Research Findings Are False', PLOS Medicine, 2(8), p. e124. doi: https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J.P.A. (2008) 'Why Most Discovered True Associations Are Inflated', Epidemiology, 19(5), pp. 640–648. doi: https://doi.org/10.1097/EDE.0b013e31818131e7

Jao, I. et al. (2015) 'Involving Research Stakeholders in Developing Policy on Sharing Public Health Research Data in Kenya: Views on Fair Process for Informed Consent, Access Oversight, and Community Engagement', Journal of Empirical Research on Human Research Ethics, 10(3), pp. 264–277. doi: https://doi.org/10.1177/1556264615592385

LA Referencia - Home (no date). Available at: https://www.lareferencia.info/en/ (Accessed: 9 December 2021)

Mehra, M.R., Ruschitzka, F. and Patel, A.N. (2020) 'Retraction—Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis', The Lancet, 395(10240), p. 1820. doi: https://doi.org/10.1016/S0140-6736(20)31324-6

Observational Health Data Sciences and Informatics (OHDSI) (2021a) Chapter 3 Open Science | The Book of OHDSI. Available at: https://ohdsi.github.io/TheBookOfOhdsi/ (Accessed: 9 December 2021)

Observational Health Data Sciences and Informatics (OHDSI) (2021b) 'COVID-19 Updates Page – OHDSI', 3 September. Available at: https://www.ohdsi.org/covid-19-updates/ (Accessed: 9 December 2021)

Pisani, E. et al. (2016) 'Beyond open data: realising the health benefits of sharing data', BMJ (Clinical research ed.), 355, p. i5295. doi: https://doi.org/10.1136/bmj.i5295

Pisani, E. and Botchway, S. (2016) 'Learning from the pioneers: lessons about data platforms drawn from the WWARN experience', in. doi: https://doi.org/10.6084/M9.FIGSHARE.4476308.V1

Serwadda, D. et al. (2018) 'Open data sharing and the Global South—Who benefits?', Science, 359(6376), pp. 642–643. doi: https://doi.org/10.1126/science.aap8395

Stracke, C.M. (2020) 'Open Science and Radical Solutions for Diversity, Equity and Quality in Research: A Literature Review of Different Research Schools, Philosophies and Frameworks and Their Potential Impact on Science and Education', in Burgos, D. (ed.) Radical Solutions and Open Science. Singapore: Springer Singapore (Lecture Notes in Educational Technology), pp. 17–37. doi: https://doi.org/10.1007/978-981-15-4276-3_2

United Nations (2020) UNESCO Recommendation on Open Science, UNESCO. Available at: https://en.unesco.org/science-sustainable-future/open-science/recommendation (Accessed: 9 December 2021)

United Nations (2021) Open Science Conference 2021, United Nations. United Nations Available at: https://www.un.org/en/library/OS21 (Accessed: 9 December 2021)

Fostering Public Trust in Science through Policy and Data Stewardship

Lyric Jorgenson, PhD, Taunton Paine, MA, and Dina Paltoo, PhD, MPH

(Authors are listed in alphabetical order. Contributions are described in detail in the Acknowledgements sections.)

As we approach over two years of being gripped by a global pandemic, COVID-19 has elevated science, and trust in science, into a national conversation. To be clear, this is a conversation that is long overdue. Public engagement in science and biomedical research is essential to building and sustaining trust in the enterprise itself. This trust is particularly important when human health hangs in the balance. As the world's largest funder of biomedical research, the United States National Institutes of Health (NIH) has placed public trust at the nexus of all our efforts, from transparency in the science we support through the leveraging of resources to maximise the United States' investment. Accordingly, the NIH promulgates policies to drive a culture of responsible data stewardship to not only advance science but to hold ourselves accountable to the public we serve.

Policies promoting responsible data sharing practices are one lever by which the NIH fosters an ecosystem of trust. First, effective data sharing provides a foundation for ensuring rigour, reproducibility, and reliability in biomedical research studies. Second, sharing data allows for transparency in the products of the research itself, demonstrating accountability for the value of investing in research. Importantly, policies governing how data are stored, shared, and used allow for participants to ensure that their time and investment is done so in a manner respectful of their preferences, but also of the interests of their populations and communities.

For its part, the United States Congress recognised the importance of making research data available by advancing legislation that, for example, requires information from clinical trials to be readily accessible regardless of findings and publications from NIH-supported research should be accessible by all (Food and Drug Administration [FDA], 1997; National Institute of Health [NIH], 2021). At the NIH, specific policies directed towards advancing data sharing and improving trust in research date back as early as 1988, and have been further expanded upon by specific expectations across the agency (U.S. Department of Health and Human Services [USDHHS], 1988; Sorlie, Sholinsky and Lauer, 2015). Over time, the NIH has doubled down on the importance of data sharing expectations, applying them to large awards (i.e., more than $500,000 in direct costs per year), high-value data sets such as research generating human genomic data, and datasets needed to achieve specific scientific and programmatic priorities (NIH, 2003; NIH, 2007; National Library of Medicine [NLM], 2021). Most recently, the NIH issued its Policy for Data Management and Sharing (DMS Policy, 2020), effective from January 2023, to promote sharing across all NIH-funded research, regardless of funding amount or research focus. The DMS Policy underscores NIH's commitment to data stewardship, reinforcing public accountability and transparency. Availability of data is only the first step, however, and as such, as stated in the NIH Strategic Plan for Data Science (2018), it aims to make the data findable, accessible, interoperable, and reusable (FAIR) so that all can benefit from this investment.

As expected, the NIH prioritizes careful stewardship of data derived from human volunteers, particularly in determining how and when to share. NIH's genomic data sharing policies (2007; 2014) mandated consent for generating and sharing human genomic data and created a framework and governance structure for sharing data for secondary research uses, consistent with participants' consent. These principles of respecting research participants' interests and privacy in future research use are critical for avoiding harm to trust in the biomedical research enterprise, and underpinned NIH's historic agreement with the Lacks Family and continue to guide NIH leadership (Hudson and Collins, 2013; Wolinetz and Collins, 2020). Promoting good data stewardship is a theme of both the Strategic Plan for Data Science, which prioritized enhancing data sharing, access, and interoperability while ensuring the security and confidentiality of patient and participant data, and the DMS Policy, which expects researchers to maximise appropriate data sharing, consistent with ethical and legal limitations while encouraging researchers to consider controlling access to human data, even if de-identified. The NIH also invests in infrastructure to facilitate responsible stewardship of data, such as the National Heart, Lung, and Blood Institute's (NHLBI) BioData Catalyst, a platform that facilitates access to and analysis of NHLBI data resources that recently opened to all researchers (NHLBI, 2021).

As science and technology advance with evolving societal views, new considerations about data sharing arise that may impact trust. If earned, trust is not a static arrangement, but must be maintained through continued investment and vigilance, whether through policy, oversight, or effective communications. Central to NIH's—and any—policy-making effort is robust stakeholder engagement and participation in the policy-making process. The NIH takes stakeholder engagement as a critical component of policy development and accordingly utilises numerous mechanisms to achieve these aims, including workshops and requests for public input. For example, given the breadth and scope of the DMS Policy, stakeholder engagement was an iterative, multi-year process, including a focused component devoted to Tribal consultation (Office of Science Policy [OSP], 2021).

As an agency continually on the cusp of pioneering advances, proactive consideration of breakthroughs on the horizon remains a component of our responsible data stewardship. For example, the NIH recently hosted a workshop on the ethics and policy of aggregating or linking data from participants from diverse research and non-research sources through privacy-preserving record linkage technologies (Office of Data Science Strategy [ODSS], 2021). NIH also recently asked the NIH Director's Novel and Exceptional Technology and

Research Advisory Committee (NExTRAC) to consider how emerging technologies are facilitating types of research questions that require increasing granularity and aggregation of data about individuals, and the potential implications for individuals, groups, and society (Tabak, 2021). As data sharing and reuse become increasingly important for fields such as Artificial Intelligence and Machine Learning (AI/ML) it increases the urgency of questions of equitability in access to and use of those data. Accordingly, NIH is investing in efforts to increase the participation and representation of researchers and communities currently underrepresented in the development of AI/ML models through the AIM-AHEAD initiative (ODSS, 2021b).

> "Science that the NIH supports can only improve human health when society trusts in the knowledge and products we generate."

At the end of the day, the science that the NIH supports can only improve human health when society trusts in the knowledge and products we generate. NIH continues to strive towards fostering a system underpinned by integrity and trust, incorporating these principles within everything we do – from public engagement in the development of policies, to the engagement of research participants, to the transparency of research outcomes. While COVID-19 has demonstrated that significant obstacles still remain, NIH will work to achieve the culture of responsible data stewardship envisioned by the DMS Policy to address these obstacles and advance NIH’s mission of transforming knowledge into improved health for all.

Acknowledgements

All authors take equal responsibility for the article in its current form. All authors collaborated in creating the first draft of the manuscript; All authors contributed to review and revision. All authors contributed to the conception of the article (including core ideas, analytical framework, and statement of research questions). All authors contributed to the project administration and to the writing process through direct writing, critical review, and commentary.

References

Food and Drug Administration (1997) FDA Backgrounder on FDAMA [Online]. Available at: https://www.fda.gov/regulatory-information/food-and-drug-administration-modernization-act-fdama-1997/fda-backgrounder-fdama (Accessed: 01 November 2021)

Hudson, K. L. and Collins, F. S. (2013) 'Biospecimen policy: Family matters', Nature, 500(7461), pp 141–142. doi: https://doi.org/10.1038/500141a

National Heart, Lung, and Blood Institute (2021) NHLBI BioData Catalyst welcomes all researchers [Online]. Available at: https://biodatacatalyst.nhlbi.nih.gov/latest-updates/NHLBI-BioData-Catalyst-welcomes-all-researchers (Accessed 01 November 2021)

National Institute of Health (2021) NIH Public Access Policy Details [Online]. Available at: https://publicaccess.nih.gov/policy.htm (Accessed: 01 November 2021)

National Institute of Health (2003) Final NIH Statement on Sharing Research Data [Online]. Available at: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html (Accessed: 30 October 2021)

National Institute of Health (2007) Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS) [Online]. Available at: https://grants.nih.gov/grants/guide/notice-files/not-od-07-088.html (Accessed: 30 October 2021)

National Institute of Health (2020) Final NIH Policy for Data Management and Sharing [Online]. Available at: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html (Accessed: 01 November 2021)

National Institute of Health (2018) NIH Strategic Plan for Data Science [Online]. Available at: https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf (Accessed: 30 October 2021)

National Institute of Health (2007) Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS) [Online]. Available at: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html (Accessed: 30 October 2021)

National Institute of Health (2014) NIH Genomic Data Sharing Policy [Online]. Available at: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html (Accessed: 30 October 2021)

National Library of Medicine (2021) NIH Data Sharing Policies [Online]. Available at: https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html (Accessed: 01 November 2021)

Office of Data Science Strategy (2021) NIH Policy and Ethics of Record Linkage Workshop. Available at: https://datascience.nih.gov/news/nih-policy-and-ethics-of-record-linkage-workshop (Accessed: 01 November 2021).

Office of Science Policy (2021) Engaging Tribal Nations [Online]. Available at: https://osp.od.nih.gov/scientific-sharing/engaging-tribal-nations/ (Accessed 01 November 2021)

Office of Data Science Strategy (2021b) About the Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) Program. Available at: https://datascience.nih.gov/artificial-intelligence/aim-ahead (Accessed 01 December 2021)

Sorlie, P.D., Sholinsky, P.D. and Lauer, M.S. (2015) 'Reinvestment in Government-Funded Research: A Great Way to Share', Circulation, 131(1). doi: https://doi.org/0.1161/CIRCULATIONAHA.114.014204

Tabak, L.A. (2021) New Charge to the NExTRAC. Available at: https://osp.od.nih.gov/wp-content/uploads/Charge_to_the_NExTRAC-Lawrence_Tabak_Jun_2021.pdf (Accessed: 01 November 2021).

U.S. Department of Health and Human Services [USDHHS] (1988) NIH Guide for Grants and Contracts. Bethesda: National Institute of Health. Available at: https://grants.nih.gov/grants/guide/historical/1988_09_16_Vol_17_No_29.pdf (Accessed: 01 November 2021).

Wolinetz C.D. and Collins F.S. (2020) 'Recognition of Research Participants' Need for Autonomy: Remembering the Legacy of Henrietta Lacks', JAMA, 324(11), pp 1027–1028. doi: https://doi.org/10.1001/jama.2020.15936

Funding Agency Data Management and Data Sharing Policies - Analysis and Comparison

The broad sharing and transparency of research outputs provide critical infrastructure for scientific advancement. Funding agencies worldwide and across the private and federal sectors have implemented policies and requirements for sharing research data, underpinning research outputs to further drive the solution to real-world problems,but few comparisons or reviews across policies has been undertaken[1]. To address this, ***Ripeta's project team*** undertook a comparative analysis of 69 funding agencies, which included 33 federal funding agencies and 13 non-profit or private funding agencies. These funding agencies were identified through a search of Sherpa Juliet, search engine queries, and known funding agencies (see methods section for additional details on the search strategy). The project team looked across eleven criteria for review and found that out of the 69 funding agencies, 62 agencies had data management, data sharing, or open science policies from which to conduct this analysis (see methods section). Of these 62 policies, 44 required a data management plan while only 3 required data sharing and 44 indicated data should be shared.

What are the established data management and data sharing policy practices for funding agencies?

In total, 69 funding agencies and related directorates fit the inclusion criteria as defined in the methods section. As seen in Table 1 below, Ripeta's project team identified 62 data policies to compare. There was significant variance in the requirement for data management plans and data sharing policies.

The analysis in this section focuses on 18 standard data management or sharing variables. The variables and their definitions can be found in the appendix.

71% (44/62) of the funding agencies **"required" data management plans**. This is in comparison to the data sharing policies which trended towards being "recommended," or suggesting that it "should" be included[2].

[1] Additional research on funding agency data policies: Joris van Rossum. (2021). Analysis of international funder data policies (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5643352;

[2] Many of the NIH-funded organizations referred back to NIH for the details of DMP and data sharing policies, but specified their own repository location where they would prefer the information be stored.

	Data management plan policy	Data sharing policy
Required	44	4
Should	2	46
None/Not found	13	9
Other	3	3
TOTAL	**62**	**62**

Table 1. Funding agency data management plan and data sharing language comparison.

68% (42/62) of the data sharing policies **allow costs** for implementing the policy as an allowable direct expense to funding. What activities were allowable varied among the data policies analysed. For example, Wellcome's data sharing policy indicates that any justified cost for delivering the plan will be considered. The National Science Foundation directorates were mostly in agreement - stating that the "costs of documenting, preparing, publishing or otherwise making available to others the findings and products of the work conducted under the grant" are allowable. None of the policies provided example costs or budgets for data management and sharing.

66% (41/62) of the data sharing policies specifically **mentioned data repositories** as a mechanism to make data publicly accessible. An even fewer number specifically named a handful of repositories as examples or requirements. The policies varied in terms of when they expected data to be released (available for sharing), but the majority expected it "as soon as possible," or "within a reasonable time." There were some policies, however, which called for very specific time frames (i.e., NHLBI, NIGMS, and CDC).

26% of the policies mentioned **how long data should be retained**. This analysis showed few data policies articulated (16/62) of how long data should be made available or shared - overall 46 data policies made no mention of how long data should be retained. Of those that did, the time varied significantly by funding organization. NIH had the most consistent guidance across their funding directorates, requiring that data be retained for a minimum of three years following the closeout of the grant. Others, such as the Medical Research Council, have retention policies based on the research type. For example, population health and clinical studies have a retention period of 20 years, while basic research data and outputs are expected to be retained for at least 10 years.

What are the challenges in understanding the data management and data sharing policy practices for funding agencies?

There is a clear need for additional guidance on DMPs and data sharing quality and completeness. Three factors continually posed issues for the data collection process and interfered with our ability to accurately show the extent to which the data policies were being implemented, tracked, and updated.

1. **Locating policies** proved to be quite difficult. Data management and data sharing policies were not easily discoverable whether using a search engine or directly searching within a funding agency's website. This suggests that funded researchers attempting to follow up-to-date guidelines on how to manage and deposit their data may be unable to locate the policies that are relevant to them.

2. We also found strong **policy variability** within funding agencies, as some of the funding agencies had different policies by directorate, ICO, or even RFP. These differences ranged from varying deadlines, repository platforms, sharing requirements, and more depending on the research being conducted. These phenomena were particularly apparent with the large number of NIH and NSF directorates. With research frequently funded by multiple agencies, this variability can cause confusion and barriers to compliance for researchers.

3. This is compounded by a persistent theme regarding the **absence or opaqueness of creation or updated dates** on the policies. Dates were equally difficult to manage and evaluate across the policies; sixty percent (n=37) of the policies were missing dates, whether that be effective dates, published dates, or a mix. Twenty policies were missing critical information about when the policy was initially published, and almost half (n=28) did not include details of when the policy was last updated. These dates help clarify when policies were effective and which requirements are most current.

Concluding Thoughts

Given the variability in and across funding agency policies for data management and sharing, harmonisation across agency policies, at least at the national level, would be a significant benefit. Researchers often have multiple grants simultaneously in progress and varying requirements for data management and sharing complicates compliance. Tuning policies to streamline research practices may improve both researcher compliance and science.

From the institutional perspective, meeting these varying requirements can also place a significant burden on institutional infrastructure and services. Many institutions have developed services and technical support, often through cross-institutional collaborations between the university library, the research office, campus IT, and others. While extremely beneficial, this can also result in additional inequities among researchers who have institutional support and infrastructure to meet data management and sharing requirements and those who do not.

Policy into Practice: Data Analyses of Five Funders

Funding agencies had approximately 1.9 million active grants from 2016 to 2020 according to Dimensions data (Digital Science, 2018). We identified and analysed five leading global industry funders. This included one non-governmental funder, the Bill & Melinda Gates Foundation, and four government funding agencies: European Commission (EC), US National Institutes of Health (NIH), National Science Foundation of China (NSFC), and German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF).

The financial investments and priorities of research vary by funder with publications representing only one outcome of their investments.

Funding Statements: How acknowledgments foster transparency

Disclosing the funding sources backing the research fosters a transparent research process. The declaration of funding sources promotes transparency, fulfils funder requirements, and acknowledges the contributions made by the funders.

Typically found within a separate section of the paper, or in the Acknowledgements section:

> *A funding statement indicates whether or not the authors received funding (institutional, private and/or corporate financial support) for the work reported in their manuscript*
> (DeVore, Hudson-Vitale and McIntosh, 2021).

As seen in the figure below, in 2020 the NIH and NSFC have the highest percentage of funding statement inclusion, but all five funding sources show relatively high compliance (Fig. 1). We can see a slight dip in funding statements from the Gates Foundation and BMBF from their original numbers in 2016, but still both remain around 80% inclusion. The EC has made sizable gains in funding acknowledgement since 2016 which is on trend with the Horizon 2020 Programme, a set of rules and guidelines aiming to increase Open Access in Europe.

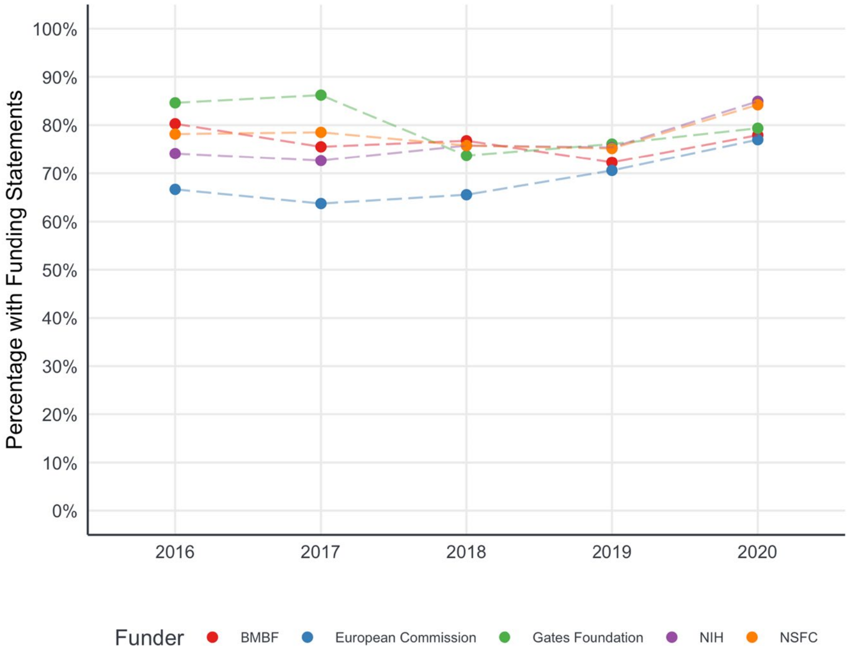

Figure 1. Percentage of funding statements found in publications by funder per year.

What are the primary focal points for funder data policies?

Open access, data management, and data sharing. In practice, the data policies require researchers to cite data sharing through Data Availability Statements (DAS).

Open access publications are over 50% for top funding agencies *Open Access* (OA) is a set of principles and actions which promote free and available publications to all (Suber, 2004). Many classifications for OA publications exist, including CC-0, CC-BY, CC-BY NC, etc. Over the past five years, OA publications have steadily increased by each agency.

For the five funders highlighted in this report, we pulled publications for all OA articles, across licence types. The Gates Foundation had 93% of their publications published as open access as compared to 31% of the NSFC (Fig. 2).

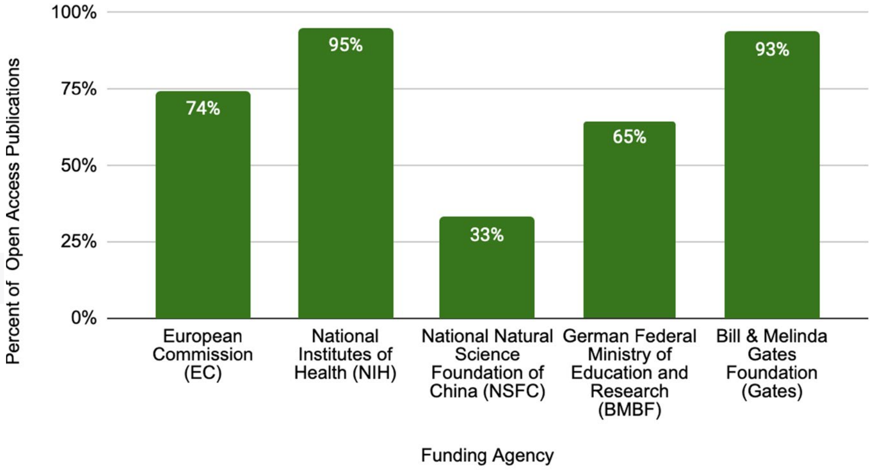

Figure 2. Percentage of open access publications in sample by funding agency.

How does data sharing and reproducibility play into funding agencies policies?

Samples of open access publications funded by each of these funding agencies were assessed for five quality indicators to provide a high level summary of: i) reproducibility-centric measures – data availability statements, data sharing locations, code sharing; and, ii) research reporting transparency – funding sources and ethical approvals (Fig. 3).

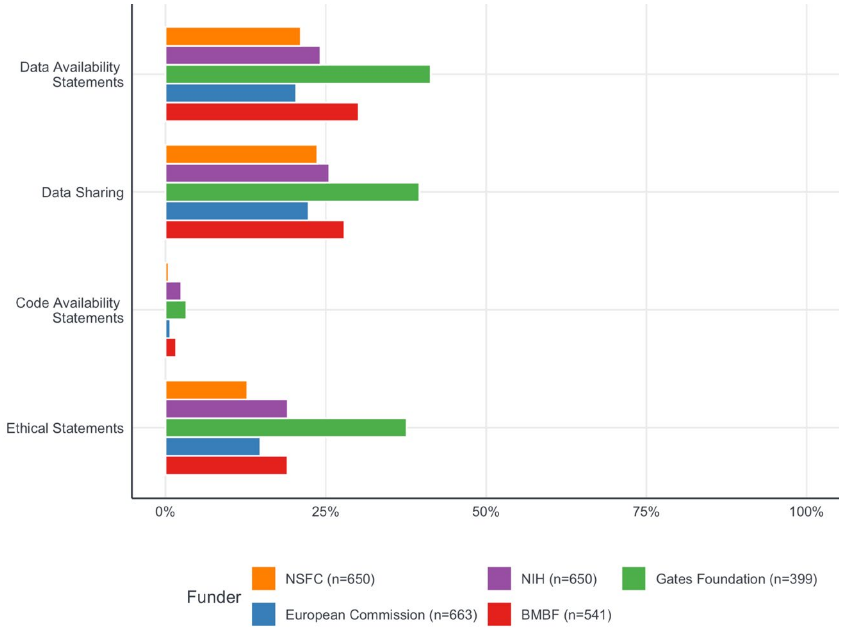

Figure 3. Percentage of sample with quality indicators by funding agency.

Data Availability Statements: A beginning for data sharing

In the United States the Office of Science and Technology Policy (OSTP) was instrumental in setting expectations for federal funding agencies to require the planning and management of research data resulting from extramural research. As a result of this policy, and others worldwide, publishers began requiring Data Availability Statements (DAS) within research papers. These statements are designed to accelerate data sharing.

> *A data availability statement (DAS) is an individual section of a scientific article offset from the main body of text that explains if or how another individual can access a study's research data. Including a DAS in a manuscript helps confirm a study, promotes stronger research transparency, and ultimately improves trust in science. While not required by all journals or funders, the DAS improves the manuscript quality and supports the citability of the data* (DeVore, Hudson-Vitale and McIntosh, 2021b).

Leveraging funding statements to determine funding agency within research articles, we can track the trends of DAS in published literature. As you can see in the figure below in 2020 there is an uptick of publications with a DAS from research from all funding agencies (Fig. 4). While agency policies play a significant role in influencing scientific practice, compliance with these may rely heavily on journal requirements. The journals requiring and checking the inclusion of a DAS (e.g., PLOS) have a near 100% compliance regardless of where the research was funded (data to be shown in follow-up report with a focus on publishers).

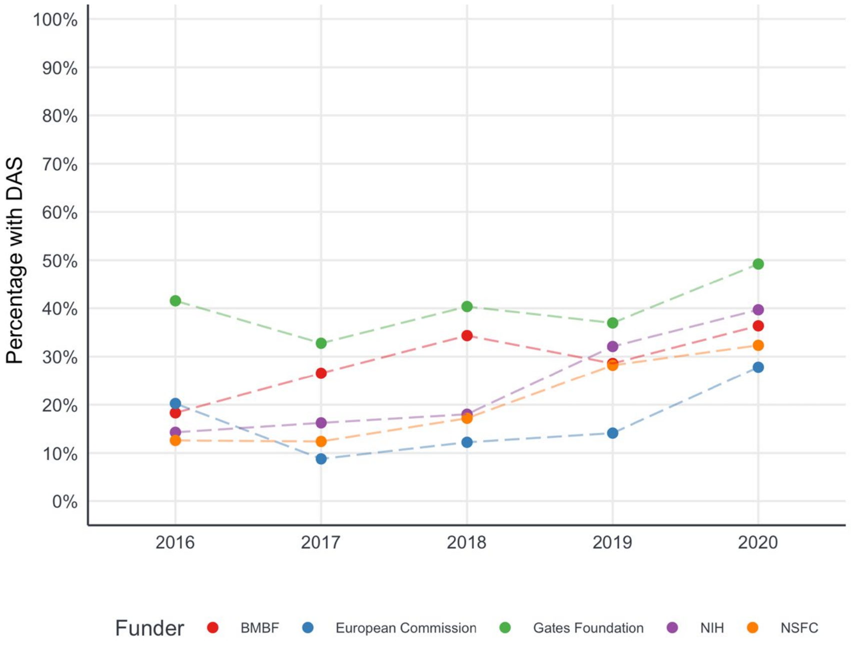

Figure 4. Inclusion of data availability statements per funding agency by year.

Data Location: Making data accessible

Making data accessible through repositories can be a radical act in the move towards FAIR and open data sharing (Wilkinson, Dumontie and Aalbersberg et al., 2016). Data locations like the ones below each have their strengths and weaknesses, but data availability through a repository is the most easily retrievable form of sharing. It allows for researchers to store large amounts of data securely without compromising openness.

> *Research shows that data location can also serve as a proxy for the completeness of the data; for instance, full data sets are more likely to be available when they are shared in external repositories or upon request rather than when they are made available in the article or supplemental files*
> (DeVore, Hudson-Vitale and McIntosh, 2021c).

It is important to note that figure 4 and 5 include a sample size of 2903 papers, only 775 of which included DAS and data locations. While the findings are indicative of some small-scale trends, there will be more to sample from in future reports. From 2016 to 2020 we have noted a steady decline in the number of funders with data available in files.

“Data available upon request” is becoming more prevalent among publications that have also been funded by a federal or private agency. Yet the use of repositories would greatly improve findability and accessibility of the data. While data available upon request still allows for the sharing of data, it is limited in terms of widespread data transparency because of its reliance on an individual or group of individuals to share their data on a case-by-case basis. Recent research has also shown that requesting data has limited success (Tedersoo & et al., 2021, Langille & et al., 2018; Krawczyk & Rueben, 2012).

"Data available upon request” is becoming more prevalent among publications that have also been funded by a federal or private agency. Yet the use of repositories would greatly improve findability and accessibility of the data."

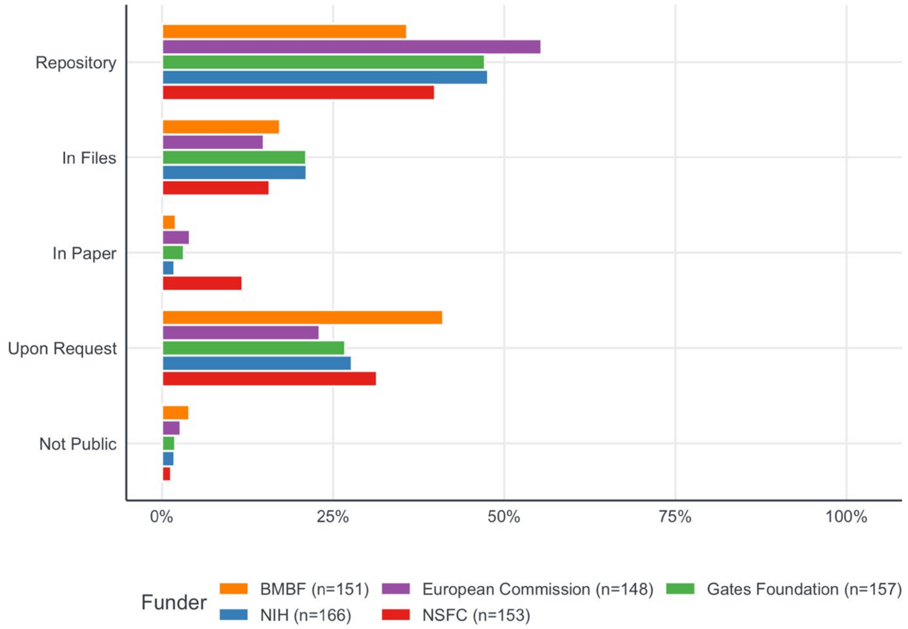

Figure 5. Percentage of samples indicating where data is shared by funding agency.

Ethical Approval Statements: Promoting transparency

Ethical approval statements (EAS) provide crucial information in certain studies with human participants or animal subjects. The EAS ultimately indicates whether or not the author(s) went through the appropriate channels to receive approval for their study.

> *In order to promote transparency and trust in science, researchers should include an ethical statement any time their study is reviewed by and conducted with the approval of an institutional review board (IRB)/research ethics committee (REC), and within the paper, an ethical approval will have its own separate Ethical Approval Statement section. If not, the information is often located in the first paragraph of the methods section*
> (DeVore, Hudson-Vitale and McIntosh, 2021d).

Once again, the Gates Foundation tops the list in the most publications with ethics statements. In 2020 we also saw an increase in the overall number of papers including ethical approval statements from all funders (Fig. 6).

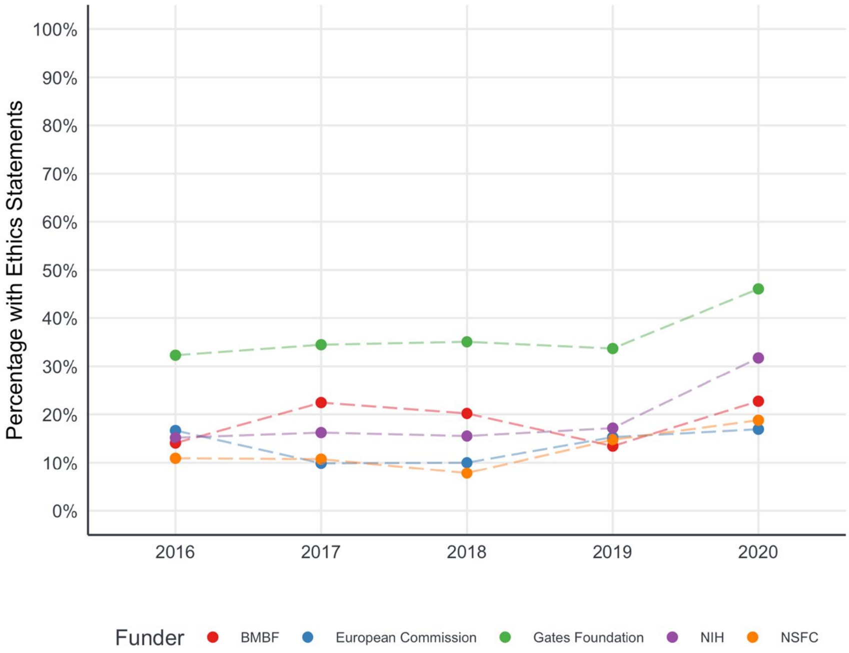

Figure 6. Percentage of sample with ethical statements per funding agency by year.

Code Availability Statements: Taking data transparency a step further

Code availability, or a statement (offset from main text) explaining how or if one can access a study's code, has just begun gaining momentum within open scholarly publishing. It is a way to gain deeper insights into the breadth of data and the methodology for assessing it. When a custom algorithm or code is central to the understanding of a paper's conclusions, researchers are encouraged to include it in the additional materials or a separate section of the paper. This makes it possible to replicate the study and retest conclusions.

Code availability is a relatively new phenomenon, and the percentage of inclusion of sharing code (<5%) is lower than some of the other variables explored in this report. With that said, the NIH and Gates Foundation both saw increases between 2019 to 2020 moving from an average of 3% of publications with code availability to 6%. Both agencies have plans to maximise transparency in publications, so it would make sense to see increases in these areas. The NIH's recently released 2020 Final NIH Policy for Data Management and Sharing, effective 2023, includes code sharing in their Data Documentation section, and the Gates Foundation could also see improvements following their January 2021 Open Access Policy implementation.

Conclusion

Policies are critical pieces of infrastructure and key incentives for advancing trust and integrity in research - when implemented. As the data from the last two sections have shown, it is not enough to simply have policies in place, but essential to check on the adherence and variation to those policies. In viewing the five-year trends, we see adherence to reproducibility and improved transparency practices associated with funding agency policies. Stating where data are shared has increased across funders while code availability statements - not required in most policies - remains rare. An interesting facet not shown, however, is how publishers are key stakeholders in implementing funding agency policies, even though research is also communicated outside of articles. Thus, the responsibilities of improving research integrity falls across the research ecosystem. As the Ripeta team continues to build more automated checks, it will be easier to investigate these trends in the future.

References

DeVore, A., Hudson-Vitale, C. and McIntosh, L.D. (2021) 'Anatomy of a Funding Statement,' Ripeta, 18/October. Available at: https://doi.org/10.6084/m9.figshare.16826455.v1 (Accessed: 25 January 2022)

DeVore, A., Hudson-Vitale, C. and McIntosh, L.D. (2021b) 'Anatomy of a Data Availability Statement,' Ripeta, 18/October. Available at: https://doi.org/10.6084/m9.figshare.14847921.v2 (Accessed: 25 January 2022)

DeVore, A., Hudson-Vitale, C. and McIntosh, L.D. (2021c) 'Anatomy of Data Sharing: Where to Share,' Ripeta, 18/October. Available at: https://doi.org/10.6084/m9.figshare.14923899 (Accessed: 25 January 2022)

DeVore, A., Hudson-Vitale, C. and McIntosh, L.D. (2021d) 'Anatomy of an Ethical Approval Statement,' Ripeta, 18/October. Available at: https://doi.org/10.6084/m9.figshare.16826467.v2 (Accessed: 25 January 2022)

Digital Science (2018) Dimensions [Software]. Available at: https://app.dimensions.ai (Accessed: 22 April 2022 under licence agreement)

The Gates Foundation (2021) Open Access Policy [Online]. Available at: https://openaccess.gatesfoundation.org/open-access-policy/ (Accessed: 20 January 2022)

Krawczyk, M. & Reuben, E. (Un)available upon request: Field experiment on researchers' willingness to share supplementary materials. Account. Res. 19, 175–186 (2012).

Langille, M. G. et al. "Available upon request": not good enough for microbiome data! Microbiome 6, 8 (2018).

National Institute of Health (2020) Final NIH Policy for Data Management and Sharing [Online]. Available at: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html (Accessed: 20 January 2022)

Tedersoo, L., Küngas, R., Oras, E. et al. Data sharing practices and data availability upon request differ across scientific disciplines. Sci Data 8, 192 (2021). https://doi.org/10.1038/s41597-021-00981-0

Wilkinson, M., Dumontier, M., Aalbersberg, I., et al. (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', Sci Data 3(160018). doi: https://doi.org/10.1038/sdata.2016.18

"Integrity-driven practices and accessibility will become increasingly important determinants of trust."

Evaluating Policy in Practice: A Case Study with Wellcome

Wellcome, a UK-based charitable funding agency, operates with the mission to solve global health challenges through innovations in science and technology (Wellcome, 2020). Their research and partnerships span across stakeholders in academia, philanthropy, business, governments, civil society, and the public. In a partnership with Ripeta in 2020, Wellcome evaluated the state of their extensive research portfolios as a mechanism to understand their funded research publications and impact of their open research initiatives. Leveraging Ripeta's quality indicator checks, we compiled a report and offered suggestions for how Wellcome could improve its portfolio and work to further support open science in the future. Below is an abbreviated version with the full report version available on the Wellcome's repository (McIntosh, et al., 2021).

The Problem: How is the Wellcome Open Access policy working?

Open innovations, policies, and coordinated cultural influences have placed the scientific process on a new precipice. The shift to open access (OA) has impacted not just how research is conducted, but the ways in which science is communicated, stewarded, and the subsequent mechanisms for ensuring trust. As we move into the new sphere of open research, integrity-driven practices and accessibility will become increasingly important determinants of trust.

Wellcome commissioned a report as part of its novel move towards evaluating their open access policy. Wellcome was unable to track how and if researchers were adhering to the new policies and requirements to provide access to data and code underlying research findings. Ripeta's report examined the level of implementation of the Wellcome guidelines and gave suggestions for further improvements. In some ways, this analysis was a test of how policy works in practice, and it highlighted key areas of growth in the process of moving from policy to practice.

The Results: Data availability statements increased, data sharing did not.

We summarised the transparency of reporting practices of Wellcome-funded research during the years 2016 and 2019. (For more information on our methods, please see the full report.) After analysing 6,200 articles based on research funded by Wellcome, we found that the number of articles with data availability statements (DAS) increased by 23.7% to a total of 45.5% from 2016-2019. While this is a step in the right direction towards OA, there was little change in the amount of data shared in repositories. Sharing data in repositories is one of the most secure and accessible ways to share data and the lack of repository usage suggests that although there is a sentiment of OA, there is more work to be done in terms of acting on it. Overall, there was an increase from 15 articles in 2016 to 57 articles in 2019 which met all of the quality criteria for reproducibility.

"This analysis was a test of how policy works in practice"

Here are highlighted a number of recommendations for institutions, funders, researchers, or publishers to effectively implement their policy requirements:

- Clearly define how to make research outputs more reproducible. This could mean developing a guide for how scripts, code, data, and more should be managed, shared (when appropriate), and curated.
- Provide training and support.
- Support lower, middle income country (LMIC) researchers who may need supplemental support.
- Offer a way for researchers to check their own work before submission (i.e., ripetaReview)
- Embed research quality checks into the contract proposal review stage (i.e., ripetaReview).

Data Location Results

The figure below (Fig. 7) depicts the most frequently reported data locations in papers that include a DAS, where an overwhelming amount cited a repository as their data location.

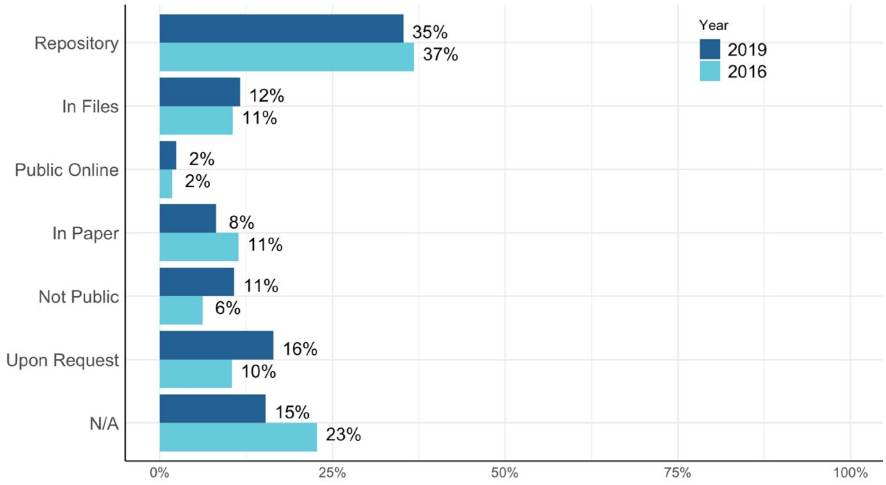

Figure 7. Data location frequency by type in papers with a DAS (sampled from 6200 papers).

The most common repositories used by researchers who included a DAS in their paper were a mix of private and disciplinary repositories: Github, Figshare, OSF, GEO, and Genbank (Fig. 8). We found an increase in the usership of almost all between 2016-2019. Genbank was the only repository to show a decrease, but it was minimal.

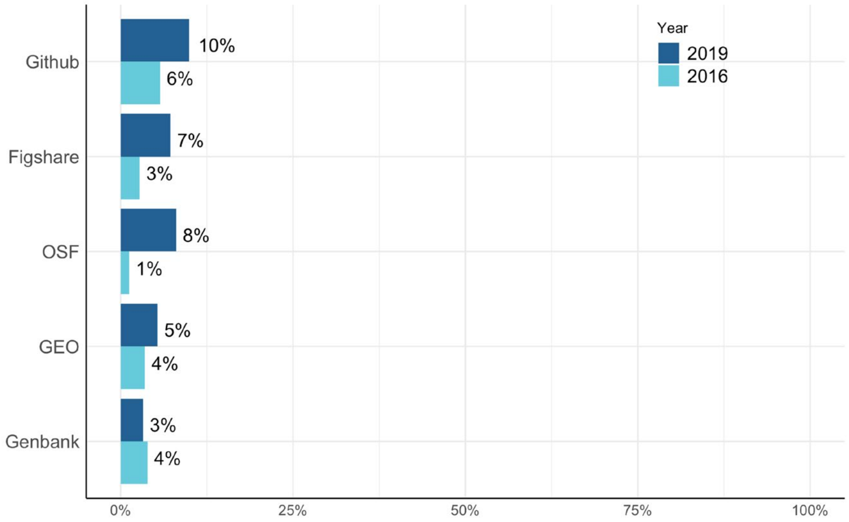

Figure 8. Most frequently used repository types from papers.

Analysis Software Results

Analysis software is shared to ensure that researchers aiming to replicate the study will have all the necessary tools to do so (Fig. 9). R, a free open-source programming language used to conduct statistical analyses and plot data, was overwhelmingly the most widely used analysis software, and was unchanged between 2016 and 2019.

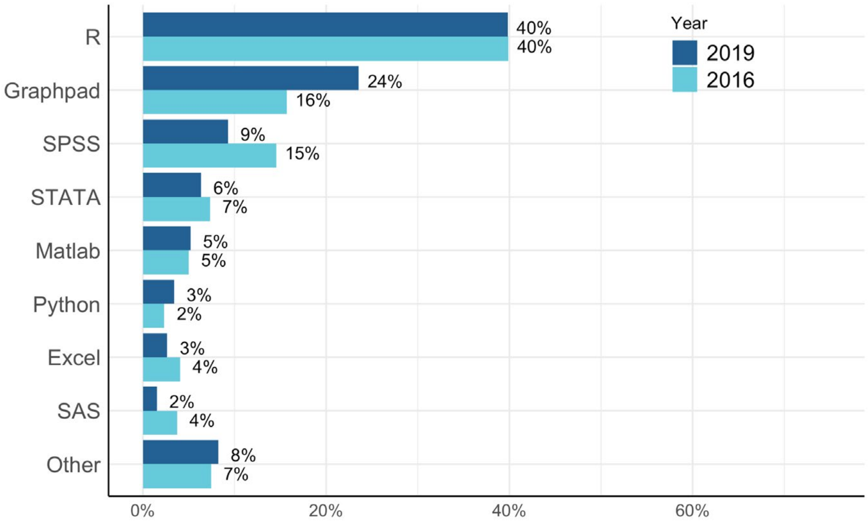

Figure 9. Common software types (denominator 3100 for each year).

Further insight into software, data, and code sharing trends came from the comparisons between subject areas (Fig. 10). Environmental science, for example, had one of the largest changes in DAS inclusion. Biomedical sciences and biology also showed a significant increase in this area. Technology and societal studies both revealed minor decreases in software, DAS, and code availability, but both fields saw higher numbers in other areas. When reviewing the corpus of papers in its entirety, there were increases in most indicators of trust, but we would like to see greater code and software availability.

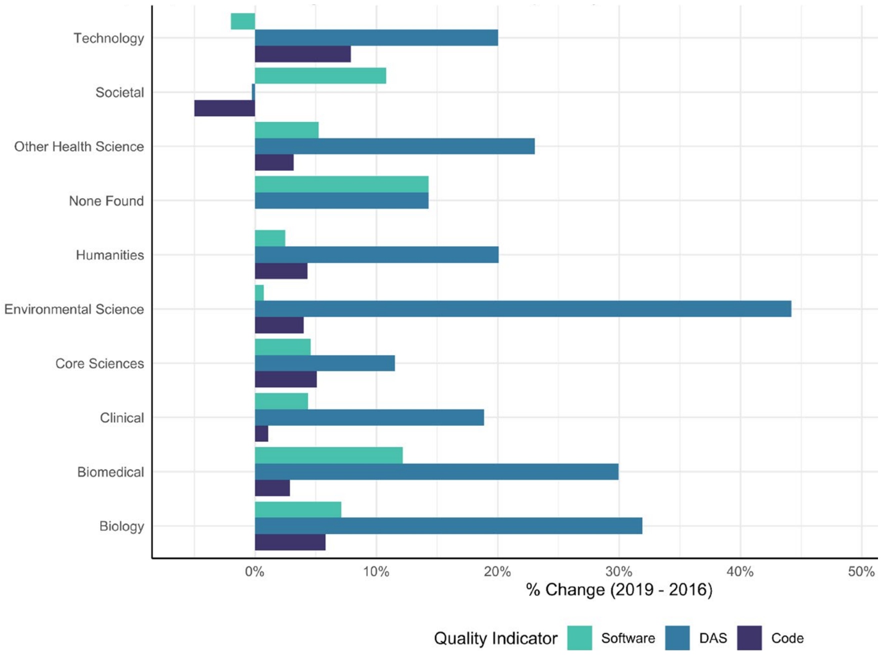

Figure 10. Software, data, and code sharing trends from 2019 to 2016 per subject area.

While each of the quality indicators alone support a degree of transparency, ideally they would all be used in accordance with one another to ensure papers are reaching the highest degree of transparency. With that in mind, papers were evaluated on their intersections between the indicators: i.e., do they include more than one indicator? If so, how has that changed over the three year period?

Overall, we saw an increase in the number of papers which included more than one indicator. Particularly, there was a strong relationship between the existence of code availability statements and data availability statements.

Acknowledgements

The team who compiled the 2020 report, “Transparently Reported Research: An analysis of Wellcome-funding publications in 2016 and 2019” consisted of Leslie D. McIntosh, Cynthia Hudson-Vitale, and Josh Sumner.

References

McIntosh, Leslie D.; Sumner, Josh; Vitale, Cynthia (2021): Transparently Reported Research: An analysis of Wellcome-funding publications in 2016 and 2019. Wellcome Trust. Online resource. https://doi.org/10.6084/m9.figshare.13810220.v1

Wellcome (2020) Who We Are. Wellcome. Available at: https://wellcome.org/who-we-are (Accessed: 26 January 2022)

Trust in Open Science is Necessary but not Sufficient for Society’s Support

Edit Herczog, MS

The 21st century seems to be the age of uncertainty and complexity, where trust in science is a critical component in the balancing act between society, economy, and the environment. While open science offers a necessary, vital component to support trust and integrity of science as a whole, open science alone is not sufficient.

Research is traditionally a trusted lighthouse in cultures and societies worldwide. Compared to the pre-internet age where information was regarded as a privilege, now there is an information flood. With a few simple keywords used to search within a database or search engine, millions of related (and sometimes unrelated) results may be returned and available for review and analysis. Various societies or countries and their leadership have sought clarity and certainty from scientific lighthouses, to select the appropriate information from the noise. Yet, this approach now falls short due to too much information and too little trust in the science.

The complex and urgent challenges facing us requires a change in scientific research. Collaboration, transparency and open science have become a necessity, recognised by funding agencies worldwide. As a result of the policies and mandates from funding agencies, the speed of transition towards open science has accelerated and become a new norm.

In the last decade, researchers have pushed for technical solutions fostering Open Access (intentionally or not) and consensus across scientific domains, such as with the FAIR principles or persistent identifiers (PID) requirement policies. While many funding agencies incorporated these principles into their guidelines, policy implementation and compliance along with researcher cultural change has lagged. Certain regions and domains trail behind, often lacking a passionate leadership at research performing organizations (i.e., private, public, non-profit organizations conducting research) to champion these practices. While there is a growing number of open science practitioners, there are three main factors that hinder the cultural shift to open science that will require continuous mitigation in the decades to come.

First, there is a need for **change management** in research organizations and in the research value chain on how the added value of open science practices can be measured. Commonly agreed-on and standardised key performance indicators are needed for institutional faculty affairs and promotion and tenure best practices that are inclusive of new curricula and appraisal systems, and throughout data generation and (re)use (from laboratory to publication).

Second, open science **requires investment** and transparency that in the short term comes with surrendering other priorities. With a fixed budget, choices must be made between finding solutions to imminent problems (e.g., SARS-Cov-2 vaccine), investigating research questions, and supporting the scientific infrastructure (e.g., workforce development, repositories). It is easy to agree on open science, but it is a quasi-mission impossible to set aside the budget to support practising it. Being a researcher and voluntarily adhering to open science practices is not sustainable. Even less so is expecting researchers to choose an improved digital infrastructure instead of new equipment to perform state-of-the-art research. It is up to the funding agencies to set strategic priorities and build trust among the stakeholders to implement their open science policies.

Third, resources and infrastructure must be **balanced** between the people championing the cause of open science and the bottlenecks slowing the adoption of open science practices. At a high level, there needs to be a robust view and support of the needs for the open science ecosystem to function and be trusted. Further, communities with less access to research resources and digital infrastructures need to be a clear priority for the allocation of these resources. Open Science should not be built on a 'first-come, first-served' approach, but based on deliberate choices for the good of science and equity.

The pandemic continues to teach numerous lessons. Once more, it has proven that collaborative data exchange in real-time significantly reduces the response time to tackle a wicked public health emergency. However, we have also seen how data nationalism has arisen. Although the best accredited laboratories produced and published transparent and reproducible results, some cultures and countries conducted independent reviews, which resulted in outcomes that conflicted with these expert judgements. This shows us open science practices are necessary for research of integrity - but not necessarily sufficient for society to trust science.

Contributor Bios

Authors

August DeVore, MS

August Devore is a Volunteer and Outreach Coordinator at Syrian Community Network, and previously held a position as the Scientific Communications Specialist at Ripeta. She completed her Masters in Environmental Science and Policy with a concentration in Climate Change and Forced Migration from Clark University. Her studies at Clark were aimed at bridging the information gap between science, policy, and humanitarianism. Her role at Ripeta varied from developing blog posts and Ripeta publications to writing press releases, maintaining social media pages, and editing outside publications.

https://orcid.org/0000-0002-5401-4550

Edit Herczog, MS

Edit Herczog was a Member of European Parliament (MEP) from 2004-2014, and is currently the owner and founder of Vision & Values SPRL, a consultancy company that provides strategic advice to top decision-makers in government and business on complex issues related to data, research, ICT, and energy. She is a Board member of the Transatlantic Policy Network to build bridges between EU and US Parliamentarians, civil servants, and businesses, and is a Council member of the international Research Data Alliance. She holds a MSc Degree in Food Engineering with specialisation in Viticulturist Engineering from the University of Horticultural Science (now part of CORVINUS University of Budapest, Hungary), and previously worked for Unilever Chemicals and ICI Group.

https://orcid.org/0000-0002-2930-5401

Cynthia Hudson Vitale, MA

Cynthia Hudson Vitale is the Director of Scholars and Scholarship at the Association of Research Libraries where she leads the association's portfolio focused on university-based publishing, distinctive collections, and research and scholarship. In 2017, she co-founded Ripeta along with Dr. Leslie McIntosh, where she currently advises on the science of research integrity and open science. Prior to joining ARL, Cynthia built and led computational research and publishing services at Penn State University Libraries and Washington University in St. Louis Libraries over the span of 15 years.

https://orcid.org/0000-0001-5581-5678

Lyric Jorgenson, PhD

Lyric Jorgenson, PhD, is the Acting Associate Director for Science Policy and the Acting Director of the Office of Science Policy at the NIH, providing senior leadership in the development and oversight of high priority and cross-cutting biomedical research policies and programs. Prior to this role, she served in numerous roles across the agency, including Deputy Director of the Office of Science Policy, and has led the development of numerous high impact science and policy initiatives such as the Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative and the National Center for Advancing Translational Sciences (NCATS). Dr. Jorgenson also served as the Deputy Executive Director of the White House Cancer Moonshot Task Force in the Office of the Vice President in the Obama administration, where she directed and coordinated cancer-related activities across the Federal government and worked to leverage investments across sectors to dramatically accelerate progress in cancer prevention. She earned a doctorate degree from the Graduate Program for Neuroscience at the University of Minnesota-Twin Cities where she conducted research in neurodevelopment with a focus on learning and memory systems. She earned a Bachelor's degree in Psychology from Denison University.

iD https://orcid.org/0000-0002-1454-7539

Gerardo Machnicki, PhD

Gerardo Machnicki has twenty years of experience in generating health economics and epidemiological studies developed in local and international positions in the pharmaceutical industry. He is currently an independent professional and leads the 2018-21 Methods Awards group of the International Society for Pharmacoeconomics and Health Outcomes Research (ISPOR). He is the author of twenty-seven peer-reviewed publications and has been a visiting lecturer in several health economics and health systems management programmes. He has led, designed and co-instructed the first ISPOR Latin America real world evidence course (September 2019) and is a member of the Observational Health Data Science and Informatics (OHDSI) consortium. He holds a BS in Economics (UCA), an MSc in Health Economics (University of York, UK), and a PhD in Public Health with a focus on Data Analysis (Saint Louis University, USA).

iD https://orcid.org/0000-0002-4696-605X

Leslie McIntosh, PhD, MPH

Dr. McIntosh has diligently worked to improve science. Since 2014, this has focused on highlighting the need for reproducible science, then on transparently reporting science, and now on the need to build trust in science. She has led diverse teams to develop and deliver meaningful data to improve scientific decisions. Dr. McIntosh is an accomplished biomedical informatician and data scientist as well as an internationally known consultant, speaker, and trainer who is passionate about mentoring the next generation of data scientists. She holds a Masters and PhD in Public Health with concentrations in Biostatistics and Epidemiology from Saint Louis University and a Certificate in Women's Leadership Forum from Washington University Olin's School of Business.

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

Taunton Paine, MA

Taunton Paine is the Director of the Scientific Data Sharing Policy Division in the Office of Science Policy in the Office of the NIH Director. Taunton has been with the Office of Science Policy since 2011. His division is responsible for issues relating to data sharing policy, including issuance of the recent NIH Data Management and Sharing Policy, oversight of the NIH Genomic Data Sharing Policy, and management of the Data Science Policy Council. Previously, he led the Clinical Research Policy team as a senior policy analyst and advised on matters related to the Common Rule, Certificates of Confidentiality, HIPAA, and other privacy and human participant protections issues. Before that, he worked on issues relating to dual-use research. He holds a dual master's degree from Columbia University and London School of Economics and Political Science, where he studied science and technology in the history of international relations.

https://orcid.org/0000-0001-9037-4556

Dina Paltoo, PhD, MPH

Dina N. Paltoo, PhD, MPH is the Assistant Director, Scientific Strategy and Innovation in the Immediate Office of the Director (IOD) of the National Heart, Lung, and Blood Institute (NHLBI), part of the NIH. In this role, she serves as a senior advisor to the NHLBI Director and provides leadership and strategic direction to complex scientific initiatives and programs related to the NHLBI mission. Dr. Paltoo came to NHLBI from the Office of the Director, National Library of Medicine (NLM) at NIH, where she served as the Assistant Director for Policy Development and led NLM's policy and legislative activities that promoted responsible stewardship and access to scientific and clinical data and information, as well as for health information technology. Prior to joining NLM, Dr. Paltoo was the Director of the Division of Scientific Data Sharing Policy and the Director of the Genetics, Health, and Society Program within the NIH Director's Office of Science Policy (OSP) and was responsible for NIH policy efforts and ethical considerations in scientific data sharing and management, open science, and genomics and health.

Dr. Paltoo previously served as a Program Director at NHLBI, where she maintained a scientific portfolio in genetics, pharmacogenetics, and personalised medicine. In her various roles at NIH, she has partnered across the NIH, Department of Health and Human Services, and Federal agencies on initiatives and activities relevant to open science, data science, and public access. Dr. Paltoo received her BS in Microbiology and PhD in Physiology and Biophysics from Howard University and her MPH from the Johns Hopkins Bloomberg School of Public Health.

https://orcid.org/0000-0002-5378-0894

Josh Sumner, MS

Josh Sumner is a computational scientist at the Donald Danforth Plant Science Center in St. Louis, Missouri, and formerly was a data scientist at Ripeta. He graduated from Appalachian State University in May of 2019 with a BS in Strength and Conditioning and a minor in Statistics, and completed a MS in Biostatistics at Washington University in St. Louis in December of 2021. He is particularly interested in statistics, its applications in health and plant science, and its role in sports science research.

https://orcid.org/0000-0002-3399-9063

Acknowledgements

Many people have helped with the preparation, editing, and formatting of this report. Our deep appreciation goes out to everyone at Digital Science including Briony Fane, Simon Linacre, David Ellis, Steve Scott, Daniel Hook, and the entire Ripeta team past and present. Finally, we'd like to express sincere gratitude to Mary Uhlmansiek who has kept this report moving forward and who's project management and expertise were critical for publishing it.

Glossary

Table 2. Glossary of terms.

Term	Definition
Data Availability Statement (DAS)	A statement, offset from the main text of a scientific paper, detailing the access to a study's data. If there is data availability information in a "Supplementary/supporting information/ materials" section or similar, it is not a DAS though it may relate to "Data Location."
Data Sharing	Research data sharing is the act of making your research data available to others for reuse. Source
Data Sharing Locations	Location that gives access to data (raw or processes).
Data Sharing Policies	Policies developed or supported by research stakeholders (e.g., government, funding agencies, publishers, institutions) to support/ require research data sharing to the greatest extent possible while protecting human subject and sensitive information.
Dimensions	Dimensions, or Dimensions.ai, is a comprehensive, aggregated data resource and tool that contains millions of research publications connected by more than 1.6 billion citations, supporting grants, datasets, clinical trials, patents and policy documents.
Funder Policies	Policies developed or supported by research funding agencies to support/require research data sharing to the greatest extent possible while protecting human subject and sensitive information.
Research Integrity	• The use of honest and verifiable methods in proposing, performing, and evaluating research; • Reporting research results with particular attention to adherence to rules, regulations, guidelines, and; • Following commonly accepted professional codes or norms. Source
Reproducibility	Reproducibility is centred around the elements of a paper which may facilitate a future researcher's ability to achieve the same results when replicating the original study.

Appendix

Table 3. Variables for comparison across funding agencies.

Variable	Variable response type	Definition
Agency Name	Free text	Name of the agency or organization
URL of policy	Free text	URL of policy on agency website
Date reviewed	Date	Date at which the data policy was reviewed by project team
Agency type	Validated list	Type of agency (Federal, Non-profit, Unknown)
Last updated	Date	The date the policy was last updated
Effective date	Date	The effective date of the policy
Timing of data management and sharing	Free text	When is the data management and sharing plan due to funding agency
Data sharing	Validated list	To what extent is data sharing required?
Data management	Validated list	To what extent is data management required?
Length of plan	Free text	Number of pages of data management or sharing plan
Data Repository	Validated text	Are data repositories mentioned in the plan?
Data Repository Location	Free text	Which specific repositories are mentioned?
Definition of data	Validated text	Does the policy define research data?
Time for data release	Free text	What is the time release at which data will be made available to the public?
Length of data retention	Free text	Data retention schedule; how long data should be kept for
Access criteria	Free text	Are there specific requirements for making the data available?
Costings	Validated list	Are data management and sharing costs allowed?
Costings activities	Free text	What types of data management and sharing activities are allowed?

Part of **DIGITAL**science

Altmetric

Dimensions

figshare

ifi CLAIMS

Overleaf

readcube

ripeta

scismic

SYMPLECTIC
Elements

SYMPLECTIC
Grant Tracker

writefull

DIGITALscience
Consultancy

digital-science.com