frontiers | Research Topics

Trust and infrastructure in scholarly communications

Edited by
Daniel W. Hook, Linda Suzanne O'Brien
and Stephen Pinfield

Published in
Frontiers in Research Metrics and Analytics

ISSN 1664-8714
ISBN 978-2-83251-088-9
DOI 10.3389/978-2-83251-088-9

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Trust and infrastructure in scholarly communications

Topic editors
Daniel W. Hook — Digital Science, United Kingdom
Linda Suzanne O'Brien — Griffith University, Australia
Stephen Pinfield — The University of Sheffield, United Kingdom

Citation
Hook, D. W., O'Brien, L. S., Pinfield, S., eds. (2023). *Trust and infrastructure in scholarly communications*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83251-088-9

Topic Editor Daniel W Hook is Director of Digital Science and Research Solutions Limited, which either owns or is a shareholder in Altmetric, Dimensions, Figshare, GRID, Ripeta, Symplectic and Overleaf, and is a minority shareholder in IFI Claims, Writefull, Scismic and Gigantum. The other Topic Editors declare no competing interests with regard to the Research Topic subject.

Table of contents

frontiers | Frontiers in Research Metrics and Analytics

EDITORIAL
published: 01 June 2022
doi: 10.3389/frma.2022.925500

Editorial: Trust and Infrastructure in Scholarly Communications

Daniel W. Hook[1,2,3]*, *Linda S. O'Brien*[4] *and Stephen Pinfield*[5]

[1] Digital Science, London, United Kingdom, [2] Centre for Complexity Science, Imperial College London, London, United Kingdom, [3] Department of Physics, Washington University in St Louis, St Louis, MO, United States, [4] Griffith Business School, Griffith University, Southport, QLD, Australia, [5] Information School, The University of Sheffield, Sheffield, United Kingdom

Keywords: trust, research infrastructure, scholarly communications, metrics, ranking, cultural norms

Editorial on the Research Topic

Trust and Infrastructure in Scholarly Communications

When Dickens wrote: "It was the best of times, it was the worst of times" in A Tale of Two Cities, he was describing the effects of two separate but linked revolutions—one in the UK where industrialization and technology had changed the social fabric as cities garnered population at the expense of their rural surroundings; and the other in France, were bloody revolution had overthrown the *ancien régime*, only to replace it with a new reign of terror. Today, we stand at the beginning of the first exponential industrial revolution: The technologies that have been born in the world of the Internet have elevated data and the AI that it fuels to the modern equivalents of oil and fire, respectively. We have seen this trend before in other industrial revolutions, but this time there is a difference—the technology that we have created has the capacity to design and build the technology that will supersede itself, leading to a self-fuelling feedback loop. Industrial revolutions have been inextricably linked to disruption—not just of industry, but also of society. The technical, economic, cultural, and societal structures that once seemed so embedded and immutable are changing quickly and the world of research is not immune.

An ever more connected system of global research information and infrastructure is transforming all aspects of the research process from how we discover and consume content to how we communicate it and the impact that it can have. At the same time, Web 2.0 has given the ability to self-publish to anyone—a tremendous freeing of global communication, but the signal-to-noise ratio has become challenging to manage, and fake news and unreliable information abound. In this new world research participation is becoming more global (both from international and intranational perspectives); research communication is more transparent through the open access, open data, and open science movements; it is becoming ever closer to translation and societal impact as we see through the rise in importance of the UN Sustainable Development Goals, a variety of grand challenge agendas, moonshots, and the adoption of impact into the evaluation environment. Data are at the centre of this change—whether it be the need to handle data volumes, the "shape" and format of data, or data as code and as a fuel for AI—the heart of any technology strategy is not just about how we collect, manipulate, consume and deploy data, but also about how it is structured, who can find it, who has access to it, who has sovereignty over it, and what capacity they have to calculate with it. And, it is more critical than ever that we understand the provenance, context, and bias of data. As data becomes part of our most powerful tools, we have to understand the "error bars" more than ever before. With so much change it is unsurprising that the infrastructures and norms of the research world, which were built in a different time, are under stress.

OPEN ACCESS

Edited and reviewed by:
Dietmar Wolfram,
University of Wisconsin–Milwaukee, United States

***Correspondence:**
Daniel W. Hook
daniel@digital-science.com

Specialty section:
This article was submitted to Scholarly Communication, a section of the journal Frontiers in Research Metrics and Analytics

Received: *21 April 2022*
Accepted: *29 April 2022*
Published: *01 June 2022*

Citation:
Hook DW, O'Brien LS and Pinfield S (2022) Editorial: Trust and Infrastructure in Scholarly Communications. Front. Res. Metr. Anal. 7:925500. doi: 10.3389/frma.2022.925500

As a result of this impetus to change, infrastructures that have underpinned the research conversation for 350 years are changing and, as such, are vulnerable. There has never been a more important time to consider how we trust the infrastructures that we rely upon, and how those infrastructures can engender trust in communities. Thus, the collection of articles in this Research Topic reflect a diverse range of different perspectives on how the scholarly communications research infrastructure is changing and the issues of trust in both the best and worst of times.

Research metrics drive evaluations and reputations in many parts of the world. Gadd makes a powerful argument that institutions should challenge the current ranking methodology and introduce healthier approaches to ranking, while Sumner et al. consider a shift away from attention-based ranking and metrics by introducing measures of trust.

Carrying out research is also replete with challenges in a technology-driven age. The rise of essentially Western technologies carries with it implicit assumptions about how the world is structured. Zeitlyn considers issues of trust between researcher and research subject in the context of social anthropology when technology, used unwittingly, can compromise anonymity. Ruckstuhl considers related issues associated with indigenous populations and how Western ways of knowing are anathema to people who experience knowledge in a completely different way.

Flanagan et al. take a practitioners' view on how global norms simultaneously require and create the trust that is required to collaborate. Ignat et al. then question how that trust can be leveraged against us and undermined by surveillance capitalism in the research world. Porter imagines the stakeholders in the research community as citizens and conducts a large-scale analysis of the extent to which ORCID is an empowering and inclusive piece of infrastructure. Altman and Cohen take an expansive view and address the question of what makes a good ecosystem? Finally, Barbour discusses whether the disparate forces of openness, integrity, inclusion, and innovation in scholarly communication compete or complement each other.

We hope that this Research Topic leads to a broader and better-informed discussion on the infrastructures that are being created. It is clear that we are at a critical point in the development of the research world—setting up our infrastructures and cultural norms to respect many different perspectives will be critical in creating an inclusive and open yet trusted and sure foundation on which the future of research can be built.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

ACKNOWLEDGMENTS

We are grateful to all the authors and reviewers who were kind enough to invest their time and efforts into this collection of articles.

Conflict of Interest: DH is the CEO of Digital Science, the provider of software in the research infrastructure space including Altmetric, Dimensions, Figshare, GRID, Overleaf, ReadCube, Ripeta, and Symplectic.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Research Metrics and Analytics

PERSPECTIVE
published: 09 September 2021
doi: 10.3389/frma.2021.680023

Mis-Measuring Our Universities: Why Global University Rankings Don't Add Up

Elizabeth Gadd *

Research & Enterprise Office, Loughborough University, Loughborough, United Kingdom

Draws parallels between the problematic use of GDP to evaluate economic success with the use of global university rankings to evaluate university success. Inspired by Kate Raworth's *Doughnut Economics,* this perspective argues that the pursuit of growth as measured by such indicators creates universities that 'grow' up the rankings rather than those which 'thrive' or 'mature.' Such growth creates academic wealth divides within and between countries, despite the direction of growth as inspired by the rankings not truly reflecting universities' critical purpose or contribution. Highlights the incompatibility between universities' alignment with socially responsible practices and continued engagement with socially irresponsible ranking practices. Proposes four possible ways of engendering change in the university rankings space. Concludes by calling on leaders of 'world-leading' universities to join together to 'lead the world' in challenging global university rankings, and to set their own standards for thriving and maturing universities.

Keywords: global rankings of universities, GDP, global inequities, responsible metrics, higher education institutions

OPEN ACCESS

Edited by:
Stephen Pinfield,
The University of Sheffield,
United Kingdom

Reviewed by:
José Augusto Guimaraes,
São Paulo State University, Brazil
Daniel W. Hook,
Digital Science, United Kingdom

***Correspondence:**
Elizabeth Gadd
e.a.gadd@lboro.ac.uk

Specialty section:
This article was submitted to Scholarly Communication, a section of the journal Frontiers in Research Metrics and Analytics

Received: *12 March 2021*
Accepted: *18 August 2021*
Published: *09 September 2021*

Citation:
Gadd E (2021) Mis-Measuring Our Universities: Why Global University Rankings Don't Add Up. Front. Res. Metr. Anal. 6:680023. doi: 10.3389/frma.2021.680023

INTRODUCTION

In 2008, in the wake of the global financial crisis, President Nicolas Sarkozy commissioned senior economic thinkers Joseph Stiglitz, Amartya Sen and Jean Paul Fitoussi, to investigate how economic and social progress might better be measured. The resulting report, later published as a book called *Mis-Measuring Our Lives: Why GDP Doesn't Add Up* (Stiglitz et al., 2010), concluded that the use of a single indicator to evaluate social progress was causing significant financial, economic, social, and environmental damage. These ideas have been developed more recently by Raworth (2017) in *Doughnut Economics*. In this perspective I argue that the same criticisms can be levelled at the use of global university rankings to assess the performance of higher education (HE) institutions and suggest some ways in which the HE community might seek to engender change.

THE PROBLEM WITH GROWTH

The idea of growth is almost universally seen as a positive. However, as Raworth, Stiglitz, Sen and Fitoussi make clear, whether growth is actually a positive, depends on how you characterise and measure it. The use of a single indicator, Gross Domestic Product (GDP), to measure economic growth is hugely problematic because it largely ignores the means by which growth is achieved (including disregarding environmental consequences) and the impacts of that growth on the human race (such as growing inequality). Escalating numbers are living in poverty while wealth increasingly amasses in the hands of a few. This is the case even though GDP actually fails to capture all the

elements that contribute to a growing economy (such as volunteering and childcare) and also fails to recognise that a growing economy does not represent everything that is important in life (such as fresh air and friendship). Despite this, society is politically locked into a GDP-based growth-at-all-costs mindset, because not to grow would be viewed as a failure, even though such growth might be leading to our own extinction.

Of course, higher education also increasingly operates as a market, much to the concern of many in the sector (Molesworth et al., 2010). If we accept that this is the case, then global university rankings are almost certainly its problematic single indicator of success. I say 'single' but in fact there are three dominant global rankings (Academic Ranking of World Universities (ARWU), Quacarelli-Simonds (QS) and Times Higher Education World University Rankings (THE WUR)), which share many similar features (see Bekhradnia (2016) for an overview) and many others besides (IREG, 2021).

University rankings share many characteristics with GDP. They are both single indicators seeking to signify the success of multi-faceted and hugely complex entities. They are both composite indicators which seek to incorporate different facets of those entities, but simultaneously fail to incorporate some of their vital qualities. Neither seek to normalise for inherited characteristics that give the entities measured an advantage over others (age, wealth and geography) and yet both provide their 'winners' with further advantages (membership of the G8 for example). Despite such criticisms, both established and emerging entities continue to trust GDP and university rankings as a benchmark even though, as the Sarcozy report put it, those attempting to do so "are like pilots trying to steer a course without a reliable compass."

GROW OR THRIVE?

One of the main justifications one hears for university rankings is that they enable HEIs in low and middle-income countries (LMICs) to leverage investment in HE where their governments otherwise might not do so. Indeed, in the last 20 years, Taiwan (Lu, 2004), Russia (Osipian, 2020), China (Anon, 2017) and Japan (Yonezawa, 2006) have all invested in programmes to develop 'world-class' (read 'ranking-topping') universities. Unfortunately, all too often that investment is made to enable institutions to climb the rankings rather than to develop strong universities (Munch, 2014). As Raworth observes about GDP, such indicators put pressure on entities to grow, whether or not they thrive. Whereas what we really need are indicators that cause entities to thrive, whether or not they grow. As newer entrants soon realise, unless they have the natural advantages of already highly ranked institutions (old, large, wealthy, 'white,' 'Western,' English-speaking research-intensives (Marginson and van der Wende, 2007; Salmi, 2009; University Wankings, 2021) their chances of displacing such organisations is very low. Thus, if they are unable to create a comparable university, their only option is to create a similar-looking surrogate.

To this end, we see university marketing budgets soaring as institutions seek to paint themselves as 'world-leading' (Moore et al., 2017; Hall and Weale, 2019). In India, a new class of 'Institutions of Eminence' has been created (India Ministry of Education, 2018), and recently this honour was bestowed on a new university, yet to prove its worth, perhaps in the hope that nominative determinism would do its work.

At the darker end of the spectrum of ranking-climbing activity, there are large numbers of HEIs seeking to either 'game' the rankings (Calderon, 2020) or simply to cheat. Activities might include legitimate attempts to solicit survey respondents that are likely to assess an HEI favourably or illegitimate practices such as paying for an institution's name to appear on highly cited papers (Kehm, 2020), or "industrialised self-citation" activity to boost THE WUR citation scores (Holmes, 2017). Such activity is by no means limited to HEIs from LMICs, however. Morphew and Swanson (2011) report on activities by US universities to present admissions and faculty data in ways that are advantageous to their ranking position.

In some cases the rankings agencies themselves are seen to be complicit in gaming. Chirakov (2021) reports how Russian HEIs that frequently engage with QS Consultancy services seem to "improve their positions in global university rankings over time, regardless of improvements in their institutional quality," observing that in the QS ranking, one's "faculty-student ratio score. . .can be "improved" relatively easily by adjusting the way data is reported." Holmes (2016a) and Holmes (2016b) describe how changes to the calibration of THE WUR methodologies seem to favour the hosts of that year's Times Higher Education Summits.

A last resort for institutions or regions that do not fare well in the existing rankings is to create their own. This was the origin of the ARWU ranking, developed by Shanghai Jao Tong University in an effort to challenge the dominance of Western universities. Recent efforts include Moscow's Three University Missions Ranking (MosIUR, 2020) which puts six Russian universities in the top 200, outperforming the one that appears in the top 200 of the QS and THE WUR thus making headway on their otherwise failed ambition to get five institutions in the top 100 by 2020 (Osipian, 2020).

Of course, all this activity focuses energy and resource on developing universities that 'grow' up the rankings, rather than institutions that truly 'thrive.'

GROW OR MATURE?

It is not only emerging institutions that suffer at the hands of a growth (or climbing) fixation, it's mature institutions too. This is because, as Raworth observes, nothing grows forever. In the natural world there is a growth phase followed by a maturing, fruit-bearing phase. Thus, when an institution matures, it would not be unusual for its income, productivity and other indicators currently assessed by global university rankings such as staff: student ratios to stabilise and with them, the institution's rank.[1]

[1]Of course, one of the failings of rankings is that they count things like Nobel prizes from time immemorial, giving some older institutions an eternal advantage.

Indeed, the same could be observed of the wider academy. With the global rush to invest in research and development (R&D), questions are now being asked as to what is the optimum size of both individual HEIs and a nation's R&D sector, and at what point does the return on investment start to diminish? However, whilst in the natural world, a plateauing of growth would be considered a healthy situation—a sign of a thriving entity—in the current HE economy, where success is measured by rank, this could have a significant negative impact on an institution's long-term viability.

Instead, mature institutions, like mature economies, made anxious by this stasis, start taking drastic and desperate action in order to keep on climbing. Such actions might include merging smaller institutions into larger ones to increase their visibility and impact, as with the creation of France's mega-university, Paris-Saclay (Anon, 2020). They might also involve dismissing researchers that fail to publish in highly-cited journals (Bonyhady and Chrysanthos, 2020) or recruiting only academics on Clarivate's Highly Cited Researcher list (Baker, 2018). Just as unranked universities might develop a new ranking that better showcases their strengths, ranked universities might put a new spin on existing rankings that suggest they are higher than they are. A recent effort to aggregate the already aggregated scores from the three most prominent global rankings into the Aggregate Ranking of Top Universities (ARTU) by Australia's University of New South Wales (2020) is a prime example.

Locke (2011) has observed that the global university rankings run on a deficit model, characterised by anxiety. Institutions are either anxious to be placed, anxious to climb, or anxious to retain their rank. However, it is those mature institutions at the top that have the most to lose; better to be on the rise than on the decline. Luckily for the ranking agencies, fear sells. As such there is no shortage of data and consultancy products available to those who wish to improve their position and the conflict of interests this represents has not gone unobserved (Chikorov, 2021). One such product is an exclusive club called the World 100 Reputation Network (2021) for institutions ranked in the top 200 of one of the big four global rankings to enable them to share strategies for retaining their ranking topping status.

THE GROWING INEQUITY OF GROWING INEQUITY

This club is an excellent example of the Matthew effect (where the rich get richer, and the poor get poorer). Top 200 institutions have special status: funders fund them, talented students and faculty want to work at them and so governments invest in them. However, we know that it is the already wealthy, established, often English-speaking institutions in the global north that dominate the top 200. Their rank elevates their status, attracting greater endowments, enabling further investments in people and facilities, which further increases their lead. The effect of pursuing ranking-related 'growth,' just as with GDP, increasingly concentrates the 'wealth' (reputation and financial) in the hands of a few, leaving others without (Aldred, 2019).

Data from the OECD (2021) plotting global investment in tertiary eduction shows that in 1995 the spend per tertiary education student ranged from 0 to 15K USD, however by 2016 the range had almost doubled to 3.76K–30K USD. Whilst there may be many factors influencing these figures, two things are clear: 1) those at the top have stayed at the top, and 2) the disparity between the "haves" and the "have-nots" is growing, rather than shrinking.

Disparities within countries is as problematic as disparities between countries. As Sarkozy's report points out, the use of averages to depict growth can mask huge inequities in the underlying data. Average income can go up, whilst the actual income of the majority of citizens goes down, obscured by the extremely high incomes of a small number of wealthy individuals inflating the mean.

There have not been many analyses of the growth of reputational or financial wealth of universities over time. However, an investigation by Shen (2013) demonstrated a growing disparity in academic salaries offered by the richer and poorer US universities. He showed that "a full professor at a public US doctoral university in 1970–71 could have expected a salary equal to 91% of what a colleague at a comparable private university earns. But by 2012–13, the proportion for a public university professor's pay had declined to only 65% of his/her peers at private schools."

A study exploring the geographical concentration of UK research funding recently showed that 49% of public R&D spend and 71% of capital infrastructure research spend between 2007 and 2014 was in London and the South-East of England—where the United Kingdom's five top-ranked universities are based (Forth and Jones, 2020). A recent assessment of the impact of the COVID-19 pandemic on UK university finances showed that the thirteen in danger of insolvency were mainly less well-ranked universities more likely to be affected by a downturn in student recruitment (Drayton and Waltman, 2020).

Investments made by LMICs to get on the rankings' 'ladder' are similarly concentrated around the small number of institutions where they feel they have the best chance of success. The consequence, as critics of India's Institutions of Eminence point out, is that the rest of that nation's higher education establishments get left behind (Mittal and Tiwari, 2020). Analyses of government-funded university excellence initiatives in other parts of the world such as China (Zong and Zang, 2019), Russia (Lovakov et al., 2021), and Japan (Yonezawa and Shimmi, 2015) all show considerably larger disparities between funded and unfunded institutions at the end of the exercise. These disparities are evident across a range of indicators such as publications in highly cited journals, international collaborations, and the recruitment of talented students and overseas academics.

It is for this reason that Hazelkorn (2017) suggests that governments should invest in world-class HE systems rather than world-class universities. While this still leads to global competitiveness, at least it promotes the funding of a broad range of HEIs that serve a range of local needs, rather than feeding some and starving others.

However, the problem is not just that some get left behind, but ultimately that the rankings they are climbing are not going to get them where they need to go. The pursuit of ranking-related 'growth' is at odds with the ability of universities to mature and thrive. This is because when you look at the behaviours necessary to climb the rankings, they are not behaviours that lead to healthier institutions, but ones that lead to toxic, unhappy institutions with deeply misplaced loyalties. Indeed, the dimensions evaluated by the global university rankings are not always representative of those that lead to a strong university at all.

WHAT DO UNIVERSITIES ACTUALLY DO?

The global rankings seek to assess universities across a number of dimensions: teaching, research, reputation, industry-focus, and collaboration. However, Selten et al. (2019) have demonstrated through principal component analysis and exploratory factor analysis that success in the rankings is essentially a function of an institution's citations and reputation. Unfortunately, citations are a notoriously poor proxy for research quality (Gingras, 2014) and are measured by the rankings using bibliometric sources that significantly disadvantage the global south (Becerril-García, 2020).

Similarly, the use of reputation as a success indicator is hugely problematic. Firstly, reputation is never a true reflection of reality. As Abraham Lincoln once said, "Character is like a tree and reputation like its shadow. The shadow is what we think of it; the tree is the real thing." Secondly, measuring a university's reputation, like measuring shadows, is extremely difficult to do. Again, Selten et al. (2019) found that the opinion surveys used by the rankers to score a university's reputation ultimately measured only brand awareness. Indeed, the THE WUR recently stated that they saw a university's reputation as synonymous with its brand (Ross, 2021).

We can therefore conclude that the qualities ultimately measured by the global university rankings do not map onto the mission statements of most universities. Teaching and learning is a principal aim of all HEIs, and yet has no bearing on an institution's rank. It is, of course, notoriously difficult to measure on a global scale and so rankers rely on very poor proxies such as staff:student ratios, alumni with Nobel prizes or a teaching reputation survey. Unfortunately, Selten et al. (2019) have demonstrated that teaching reputation surveys correlate closely with research reputation surveys, again suggesting that it is brand rather than teaching quality that is being measured.

Universities' so-called 'third missions'—their research impact and enterprise activity—are not measured at all by the mainstream university rankings. Lee et al. (2020) argue that this further discriminates against institutions in the global south that may be more mission-orientated. The THE WUR have recently introduced an Impact Ranking based on the UN Sustainable Development Goals, however, again due to the lack of globally comparable impact data, universities are largely left to supply their own evidence which does not make for an equitable comparison (Curry, 2020). Interestingly, this evidence is supplemented by more bibliographic data from the same globally skewed source as their mainstream ranking which rather mitigates against, rather than for, sustainable development.

However, even if the rankings were able to measure the quality of a university's teaching, research and enterprise, evidence has shown that such successful outputs are largely a product of a university's inputs: their age, wealth and geography. The university that has the wealth and reputation to recruit and resource the most talented academics is likely to get the best outcomes—especially when the world is pre-disposed to overvalue the outcomes of that already well-resourced and well-known university.

Such legacy "variables" should arguably be factored out of any truly responsible evaluation (Gadd et al., 2021). Indeed, what universities need to do to thrive and mature, and where all universities have an equal opportunity to succeed, is to create processes, policies and a culture that successfully convert their 'inputs' into their 'outputs.' The problem is that such things—the things we arguably value most about our universities: academic freedom, equality and diversity, good governance, and a positive teaching and research environment—are all largely unmeasurable.

WHAT TO DO?

Whilst such critiques of global university rankings will not be new to any followers of the debate, what we have yet to see in response to two decades-worth of argument, is any real change in this space. The ranking agencies remain entirely unscathed by repeated criticism and continue to proliferate, whilst end-users seem impervious to their logic and continue to rely on the rankings as a lazy proxy for a university's quality. As such institutions have had to accept global rankings as an established part of the HE landscape (the 'rankings are here to stay' narrative) and to promote their own rank in order to attract students, thus inadvertently lending the rankings legitimacy. In this way, rankings have become an uncontested institutional norm. Given that most HE institutions hold themselves to high standards around data transparency and openness which are not shared by the rankings, this is a particular irony.

It was against this backdrop that the INORMS Research Evaluation Working Group sought to consolidate best practice in the field of 'responsible' university rankings in the form of a set of principles, and to highlight the extent to which the most high-profile rankings met those criteria. They all fell short, and the most high-profile rankings fell even more short than the others. This work has been widely publicised (INORMS, 2020; Gadd et al., 2021), including a piece in Nature (Gadd, 2020), however, to date there has been no response—formal or informal—from the 'big three' global rankings (ARWU, QS and THE WUR). It should be noted that other rankings such as the Leiden Ranking and U-Multirank fared much better against the INORMS principles. However, ironically, whilst not seeking to identify the world's 'top' institutions overall won them higher scores on the INORMS ratings, this diminishes their influence globally as end-users prize quick and easy answers, even if inaccurate.

The question then remains as to how to initiate change in this domain when the key stakeholders are, like those organisations at the top of their rankings, wealthy, powerful, and seemingly impervious to critique. Are there lessons we can learn from the more long-standing and parallel problem posed by the use of GDP to measure economic success?

INDEPENDENT REGULATION

One of the challenges of university rankings is that they are self-appointed and unaccountable. The International Rankings Expert Group (IREG, 2021) claims to be an "independent" body offering ranking audits, however, a large proportion of the seats on their executive committee are occupied by ranking agencies themselves. Were the rankings overseen by a truly independent body, just as the calculation of GDP is overseen by national statistical offices around the world which report into the UN Statistical Authority, this might provide a useful challenge to some of their methodological deficiencies. An obvious choice would be the Royal Statistical Society (RSS), an international organisation whose mission includes campaigning for the effective use of statistics for the public good. The RSS recently turned their attention to the United Kingdom Teaching Excellence Framework on the grounds that it was "likely to mislead students who use TEF to inform their university choices" (Royal Statistical Society, 2019). The global university rankings as currently formulated are clearly subject to the same accusation, and a rigorous investigation by such a prestigious and independent body could be enormously influential.

START A NEW GAME?

Another option for challenging the dominance of an existing unhelpful indicator, as Raworth suggests, is to introduce an alternative. She describes the Human Development Index (UNDP, 2021), a dashboard of alternative indicators to GDP, which measures dimensions such as long life, education and living standards, which can lead to positive societal change. Of course, there is no shortage of challengers to the dominance of the current input/output dominated world rankings. Some are serious, such as the Universitas Indonesia (2020) Green University Rankings, others are less so (Greatrix, 2020).

The problem with new indicators is that all too often they do not displace existing ones but at best complement them and at worst are completely overshadowed by them. However, if the heavy users of such rankings, such as research studentship funders, could collectively agree to focus on indicators that the HE community agree are a better representation of their contribution, then this could be a significant step forward. For just as ranking agencies seek to exploit the marketplace that is Higher Education, they too are subject to the demands of that marketplace. Should the demand for their services change, their influence would change with it. It is this thought that leads to my third suggestion.

LEADERS LEAD

Whilst critiques of global university rankings are not new, what I believe is new, as the appetite for Raworth's *Doughnut Economics* has shown, is our unwillingness to tolerate initiatives that no longer align with our principles and that lead to poor outcomes for our planet and our people. The world has changed from one in which we turn a blind eye to inconvenient truths to one where we seek to tackle them head on.

In the last 10 years we have seen a growth in public statements of commitment to socially responsible practices by corporates, charities and publicly funded organisations alike. In Higher Education there has been a spotlight on Equity, Diversity and Inclusion (EDI), sustainability, improving research culture, Responsible Research & Innovation (RRI), open research and of course responsible research evaluation. Universities have declared their commitment to responsible practices through accreditation with organisations like Athena Swan (Advance HE, 2021), Stonewall (2021), the Race Equality Charter (Advance HE, 2020), the UK Reproducibility Network (UKRN, 2021), and through adopting principles such as those espoused by the Declaration on Research Assessment (DORA, 2021), the Leiden Manifesto (Hicks et al., 2015) or the Hong Kong Principles on Researcher Evaluation (Moher et al., 2020).

When one considers the perverse effects of the global university rankings: their deeply problematic methodologies that lead to a pursuit of "growing" rather than "thriving" or "maturing"; their bias towards already established, wealthy, English-speaking organisations in the global north; and their contribution towards growing academic inequities across and within countries; it is hard to understand how an organisation that is truly committed to responsible research evaluation and other socially responsible practices can legitimately continue to engage with them.

Of course, one has sympathy with divided leaders who are fully cognizant of the rankings' flaws whilst simultaneously having to rely on them to survive in a HE marketplace that is not of their making. However, in a world where leaders are increasingly called upon to make hard value-led choices, we may be approaching a time where these fundamentally incompatible positions cannot be maintained. As Leeds University's Vice Chancellor, Simone Buitendijk, recently observed.

> *"If there was ever a good time to define the moral narrative for global institutions' strategies, whether businesses, NGOs or universities, it is now. COVID has taught us the importance of prioritising human values over competition for profits, or for limited, metricised and quantitative outcomes"* (Buitendijk, 2021).

There is currently an opportunity for HEIs to rethink both participation in, and promulgating the results of, global university rankings that better aligns with institutional values. Indeed, the (European Commission's, 2020) recent report *Towards a 2030 Vision on the Future of Universities in Europe* directly challenged the reliance on university rankings as an "overly simplistic" measure of university success, preferring alternative metrics

that highlight universities' wider contribution. I would suggest that this report may provide a key as to how leaders might operationalise any move to challenge the unwelcome impacts of the global university rankings, namely, as a collective.

COLLECTIVE ACTION

Growth addiction can only be challenged by those who have grown: those institutions well-served by the current system. As Masood (2016) observed about GDP, "Any revision to the index won't pass muster unless the interests of its founder countries are protected. . .permanent members of the UN Security Council will not allow a change to GDP that leads to them slipping down the league table."

Given that the global university rankings make rivals of those entities, the only real way they are going to successfully change the system is if they join forces and agree to challenge it together. We see an example of this in the C40 (2021) network of 80 megacities (representing 25% of world GDP) who are collaborating to tackle climate change.

If senior leaders of so-called 'world-leading' mission- and value-led institutions are serious about delivering on their mission and values, it would seem logical that instead of joining exclusive World 100 reputation networks that keep less advantaged institutions from poorer countries out, they should create open, outward-facing networks that let such institutions in. As Gloria Steinem famously said, "Imagine we are linked, not ranked." Were universities depicted in terms of a network, rather than a ranking, it might reinforce the fact that this is a group of organisations with the same mission, and not a group of organisations in the same competition. Whilst institutions may not have the power to prevent third parties from ranking them, they do have the power to self-characterise themselves, and to act, as a network, and not a network that collaborates only in order to compete, one but that collaborates to do good in the world.

Instead of perpetuating the myth that global university rankings measure those things that create strong, thriving institutions, we need a new breed of principled, connected university leaders to actively call them out for their poor, Matthew effect-inducing methodologies, who commit not to use them as KPIs, provide them with data, mention them in marketing, and to avoid ranking-organised summits that further legitimise them. I am also aware this will require persuading their own governing bodies not to offer bonuses based on their rank (Musselin, 2018). Perhaps in an extension of the 'I am not my h-index' campaign promoted by researchers (Curry, 2018), we need a new campaign for universities called, "So much more than our rank"?

To be clear, this is not about giving up on notions of excellence or quality, it is about university leaders being the ones who get to define what those notions mean. It is also about saying no to the scarcity mindset generated by the global rankings, in a world where there is enough to go round.

I accept that this kind of action is on a different scale to anything previously seen in the responsible research assessment space. It has been relatively painless for institutions to implement DORA or the Leiden Manifesto—some adjustments to internal policy and process were all that was needed. The collective will required to challenge the negative impacts of the reputation-based economy as measured by the current world university rankings, necessitates looking beyond our own institutions and scrutinising their long-term, systemic, global effects. As Roman Krznaric (2021) reminds us in *The Good Ancestor: Long-Term Thinking in a Short-Term World*, we need to make the decisions now that our descendants will thank us for. Such perspectives are not often prioritised by HE administrators. However, the tide might be turning.

Dame Ottoline Leyser, CEO of UK Research & Innovation (UKRI) has started to promote the notion of 'net contribution' in the research arena: a suggestion that we are rewarded not only for the contribution we make, but for the contribution we enable others to make (Leyser, 2020). If this approach is more widely adopted it might encourage a broader definition of university 'success'—because another's success, and your contribution to it, becomes your success.

I am presenting here the moral argument, of course, because these are the claims that universities are starting to make for themselves. However, there is a pragmatic argument too. For just as the logical extension of a GDP-based growth addiction is a society where there is not enough disposable income amongst the general population to purchase the products and services of the wealthy few, so will pitting universities against one another in a global competition where only a few similar-looking institutions survive, eventually impoverish us all. We need a diversity of flourishing higher education institutions that serve the diverse needs and developmental stages of the world we inhabit if we are to thrive as a human race. If the current global crises have taught us anything it is, as the thought-leading Margaret Heffernan (2014) points out, that "no one wins unless everybody wins."

If institutions are genuinely committed to responsible evaluation practice, to equity, diversity, and inclusion, and if they are genuinely committed to delivering on their own mission to positively impact the world with their teaching and research, I would argue that this is incompatible with overlooking the negative impacts of the global university rankings.

As Raworth observed about GDP, it is time to move from "economic thinking" to "economic doing." I would urge the senior leaders of any institution that considers itself to be world-leading to lead the world in this significant and important matter. They can do so by joining forces with other principled leaders to proactively stand against substandard notions of excellence and harmful forms of competition that neither reflect their own contribution nor the contribution of their mission-sharing global network. Instead, I encourage them to work with that network to redefine what a thriving and maturing university does, namely, to develop mission-specific policies, processes and cultures that achieve their important ends, and endorse efforts to evaluate them accordingly.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

I am deeply grateful to Cameron Neylon and Ehsan Masood for their comments on an earlier draft of this piece.

REFERENCES

Advance, H. E. (2021). Athena Swan Charter. URL: https://www.advance-he.ac.uk/equality-charters/athena-swan-charter.

Advance, H. E. (2020). Race Equality Charter. URL: https://www.advance-he.ac.uk/equality-charters/race-equality-charter.

Aldred, J. (2019). *Licence to Be Bad: How Economics Corrupted Us*. London: Penguin.

Anon (2020). Egalité: France's Research Reforms Must Balance Competitiveness with Well-Being. *Nature* 587, 7–8. URL: https://www.nature.com/articles/d41586-020-02853-w. doi:10.1038/d41586-020-02853-w

Anon (2017). 'China to Develop 42 World-Class Universities', *People's Daily*. URL: http://en.people.cn/n3/2017/0921/c90000-9272101.html.

Baker, S. (2018). 'Highly Cited Researchers 2018: Australian Growth "Bucking Trend"', Times Higher Education. Available at: https://www.timeshighereducation.com/news/highly-cited-researchers-2018-australian-growth-bucking-trend.

Becerril-García, A. (2020). 'Funders' Influence on Research Organisations' Assessment Criteria and Processes (Panel Presentation)'. In Global Research Council Responsible Research Assessment Conference, 23–27. November. Available at: https://web-eur.cvent.com/event/7ca86a3d-6e6f-4d11-98e9-f01fe69fdf46/summary.

Bekhradnia, B. (2016). *International university Rankings: For Good or Ill?*. Oxford: HEPI. Available at: www.hepi.ac.uk/2016/12/15/3734/ (Retrieved January 09, 2021).

Bonyhady, N., and Chrysanthos, N. (2020). *'Industrial Umpire Lashes Universities "Obsessed" with Rankings and Reputation'*. The Sydney Morning Herald, Available at: https://www.smh.com.au/national/nsw/industrial-umpire-lashes-universities-obsessed-with-rankings-and-reputation-20200311-p5495e.html (Accessed March 11, 2011).

Buitendijk, S. (2021). Confessions of a Leader in a Time of Crisis. *Medium [blog]*. Available at: https://medium.com/university-of-leeds/confessions-of-a-leader-in-a-time-of-crisis-80d4ba14cdcf.

C40 (2021). C40 Cities. URL: https://www.c40.org/.

Calderon, A. (2020). *New Rankings Results Show How Some Are Gaming the System*. University World News, Available at: https://www.universityworldnews.com/post.php?story=20200612104427336 (Accessed June 12, 2020).

Chirikov, I. (2021). *Does Conflict of Interest Distort Global University Rankings?*. California: UC Berkeley Center for Studies in Higher Education. Retrieved from https://escholarship.org/uc/item/8hk672nh.

Curry, S. (2018). Ready-made Citation Distributions Are a Boost for Responsible Research Assessment. *Occam's Typewriter [blog]*. Available at: https://occamstypewriter.org/scurry/2018/07/01/ready-made-citation-distributions-are-a-boost-for-responsible-research-assessment/.

Curry, S. (2020). The Still Unsustainable Goal of university Ranking, Reciprocal Space Blog. URL: http://occamstypewriter.org/scurry/2020/04/26/still-unsustainable-university-rankings/.

DORA (2021). *Declaration on Research Assessment*. URL: https://sfdora.org/.

Drayton, E., and Waltmann, B. (2020). Will Universities Need a Bailout to Survive the COVID-19 Crisis?, IFS Briefing Note BN300. Available at: https://www.ifs.org.uk/publications/14919.

European Commission (2020). *Towards a 2030 Vision on the Future of Universities in Europe*. doi:10.2777/510530

Forth, T., and Jones, R. (2020). *The Missing £4 Billion: Making R&D Work for the Whole UK*. Nesta. Available at: https://www.nesta.org.uk/report/the-missing-4-billion/.

Gadd, E., Holmes, R., and Shearer, J., (2021). Developing a Method for Evaluating Global University Rankings. *Scholarly Assess. Rep.*, 3(1), p.2. doi:10.29024/sar.31

Gadd, E. (2020). University Rankings Need a Rethink. *Nature* 587, 523. doi:10.1038/d41586-020-03312-2

Gingras, Y. (2014). *Bibliometrics and Research Evaluation: Uses and Abuses*. Cambridge, Mass: MIT Press.

Greatrix, P. (2020). The Definitive Ranking of university Rankings 2020. *WonkHE Blog*. URL: https://wonkhe.com/tag/rankings-league-tables/ (Accessed December 8, 2020).

Hall, S., and Weale, S. (2019). Universities Spending Millions on Marketing to Attract Students. *The Guardian*. Available at: https://www.theguardian.com/education/2019/apr/02/universities-spending-millions-on-marketing-to-attract-students (Accessed April 7, 2019).

Hazelkorn, E. (2017). Rankings and Higher Education: Reframing Relationships within and between States. In Centre for Global Higher Education working paper series. Working paper no. 19. Available at: www.researchcghe.org.

Heffernan, M. (2014). *A Bigger Prize: When No-One Wins unless Everybody Wins*. London: Simon & Schuster.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., and Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for Research Metrics. *Nature* 520, 429–431. doi:10.1038/520429a

Holmes, R. (2017). Doing Something about Citations and Affiliations. University Ranking Watch [blog]. https://rankingwatch.blogspot.com/2017/04/doing-something-about-citations-and.html.

Holmes, R. (2016a). More on THE's Bespoke Rankings. University Ranking Watch [blog]. https://rankingwatch.blogspot.com/2016/07/more-on-thes-bespoke-rankings.html.

Holmes, R. (2016b). THE's Bespoke Asian Rankings: the Strange Decline of the University of Tokyo and the Rise of Singapore. University Ranking Watch [blog]. https://rankingwatch.blogspot.com/2016/06/thes-bespoke-asian-rankings-strange.html.

India Ministry of Education (2018). Government Declares 6 Educational 'Institutions of Eminence'; 3 Institutions from Public Sector and 3 from Private Sector Shortlisted. Available at: https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1538188.

INORMS (2020). *Research Evaluation Working Group*. Available at: https://inorms.net/activities/research-evaluation-working-group/.

IREG (2021). *International Rankings Expert Group*. URL: http://ireg-observatory.org/.

Kehm, B. M. (2020). "Global University Rankings: Impacts and Applications," in *Gaming the Metrics Bagilani*. Editor A. Lipman (Massachusetts: MIT Press). doi:10.7551/mitpress/11087.003.0009

Krznaric, R. (2020). The Good Ancestor: How to Think Long Term in a Short-Term World. Random House.

Lee, J. J., Vance, H., Stensaker, B., and Ghosh, S. (2020). Global Rankings at a Local Cost? the Strategic Pursuit of Status and the Third mission. *Comp. Edu.* 56 (2), 236–256. doi:10.1080/03050068.2020.1741195

Leyser, O. (2020). Research Assessment and Research Culture. In Keynote at Global Research Council Responsible Research Assessment conference. 23 November 2020. URL: https://www.globalresearchcouncil.org/news/responsible-research-assessment/ (Accessed February 20, 2020).

Locke, W. (2011). "The Institutionalization of Rankings: Managing Status Anxiety in an Increasingly Marketized Environment," in *Ranking, Reputation and the Quality of Higher Education*. Editors J. C. Shin, R. K. Toutkoushian, and U. Teichler (Dordrecht, The Netherlands: Springer), 201–228. doi:10.1007/978-94-007-1116-7_11

Lovakov, A., Panova, A., Sterligov, I., and Yudkevich, M. (2021). Does Government Support of a Few Leading Universities Have a Broader Impact on the Higher Education System? Evaluation of the Russian University Excellence Initiative. In 'Does government support of a few leading universities have a broader impact on the higher education system? Evaluation of the Russian University Excellence Initiative', Research Evaluation. Oxford University Press. doi:10.1093/reseval/rvab006

Lu, M. L. (2004). The Blueprint and Competitiveness of Taiwan's Higher Education. In paper presented at Cross Strait Seminar on Review and Prospect of the Policy of University Excellence. Taiwan, 25–26.

Marginson, S., and van der Wende, M. (2007). To Rank or to Be Ranked: The Impact of Global Rankings in Higher Education. *J. Stud. Int. Edu.* 11 (3–4), 306–329. doi:10.1177/1028315307303544

Masood, E. (2016). *The Great Invention: The Story of GDP and the Making (And Unmaking) of the Modern World*. New York: Pegasus.

Mittal, S., and Tiwari, S. (2020). Institutions of eminence or Institutions of Exclusion? *The Leaflet*, 12, 2020 . Available at: https://www.theleaflet.in/institutions-of-eminence-or-institutions-of-exclusion/.

Moher, D., Bouter, L., Kleinert, S., Glasziou, P., Sham, M. H., Barbour, V., et al. (2020). The Hong Kong Principles for Assessing Researchers: Fostering Research Integrity. *Plos Biol.* 18 (7), e3000737–14. doi:10.1371/journal.pbio.3000737

Molesworth, M., Scullion, R., and Nixon, E. (2010). *The Marketisation of Higher Education and the Student as Consumer*. London: Routledge.

Moore, S., Neylon, C., Paul Eve, M., Paul O'Donnell, D., and Pattinson, D. (2017). "Excellence R Us": university Research and the Fetishisation of Excellence. *Palgrave Commun.* 3. doi:10.1057/palcomms.2016.105

MosIUR (2020). *Three University Missions Moscow International University Ranking 2020*. Available at: https://mosiur.org/ranking/.

Morphew, C. C., and Swanson, C. (2011). *On the Efficacy of Raising Your University's Rankings*. In University rankings (Dordrecht: Springer), 185–199.

Munch, R. (2014). *Academic Capitalism Universities in the Global Struggle for Excellence*. New York: Routledge.

Musselin, C. (2018). New Forms of Competition in Higher Education1. *Socio-Economic Rev.* 16 (3), 657–683. doi:10.1093/ser/mwy033

OECD (2021). Education Spending (Indicator). Available at: http://dx.doi.org/10.1787/ca274bac-en (Accessed 20 February 2021).

Osipian, A. (2020). *Russia Fails to Achieve International Excellence Target*. University World News. Available at https://www.universityworldnews.com/post.php?story=20201023130100102.

Raworth, K. (2017). *Doughnut Economics: Seven Ways to Think like a 21st century Economist*. London: Random House.

Ross, D. (2021). 'A New View of university Reputation', *Times Higher Education*. Available at: https://www.timeshighereducation.com/world-university-rankings/new-view-university-reputation.

Royal Statistical Society (2019). RSS Identifies 'major Statistical Issues. TEF'. RSS Statistics News. URL: https://rss.org.uk/news-publication/news-publications/2019/general-news-(1)/rss-identifies-major-statistical-issues-in-tef/ (Accessed 6 March 2019).

Salmi, J. (2009). *The Challenge of Establishing World-Class Universities*. Washington: World Bank.

Selten, F., Neylon, C., Huang, C.-K., and Groth, P. (2019). A Longitudinal Analysis of University Rankings. Available at: http://arxiv.org/abs/1908.10632.

Shen, D. (2013). The Cost of Wealth Inequality in Higher Education, LSE Politics and Policy Blog. Available at: https://blogs.lse.ac.uk/politicsandpolicy/in-depth-a-public-school-in-an-age-of-money-and-inequality-in-higher-education/.

Stiglitz, J. E., Sen, A., and Fitoussi, J. (2010). *Commission on the Measurement of Economic Performance and Social Progress (France). Mismeasuring Our Lives: Why GDP Doesn't Add up*. London: New Press.

Stonewall (2021). Stonewall. URL: https://www.stonewall.org.uk/.

UKRN (2021). UK Reproducibility Network. URL: https://www.ukrn.org/.

UNDP (2021). Human Development Index. URL: http://hdr.undp.org/en/content/human-development-index-hdi.

Universitas Indonesia (2020). Green Metric World University Rankings. URL: http://greenmetric.ui.ac.id/overall-rankings-2020/.

University of New South Wales (UNSW) (2020). Aggregate Ranking of Top Universities. Available at: http://research.unsw.edu.au/artu/.

University Wankings (2021). Why Are Our Rankings So White? In Socially Responsible Higher Education. Chapter 5. Leiden, The Netherlands: Brill, 67–79. doi:10.1163/9789004459076_006

World 100 Reputation Network (2021). URL: https://www.theworld100.com/reputation-network/.

Yonezawa, A., and Shimmi, Y. (2015). Transformation of university Governance through Internationalization: Challenges for Top Universities and Government Policies in Japan. *High Educ.* 70 (2), 173–186. doi:10.1007/s10734-015-9863-0

Yonezawa, A. (2006). "'Japanese Flagship Universities at a Crossroads'," in *Final Report of Developing Evaluation Criteria to Assess the Internationalization of Universities*. Editor N. Furushiro (Kwansei: Osaka University), 85–102.

Zong, X., and Zhang, W. (2019). Establishing World-Class Universities in China: Deploying a Quasi-Experimental Design to Evaluate the Net Effects of Project 985. *Stud. Higher Edu.* 44, 417–431. doi:10.1080/03075079.2017.1368475

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Research Metrics and Analytics

OPINION
published: 14 October 2021
doi: 10.3389/frma.2021.736568

For Augustinian Archival Openness and Laggardly Sharing: Trustworthy Archiving and Sharing of Social Science Data From Identifiable Human Subjects

David Zeitlyn *

Institute of Social and Cultural Anthropology, School of Anthropology and Museum Ethnography University of Oxford, Oxford, United Kingdom

Keywords: human subjects, research ethics, embargo, GDPR, archives/ archiving, pseudonymisation, conflicting responsibilities

INTRODUCTION

Saint Augustine used to pray "Lord, make me chaste—but not yet." This finds a parallel in qualitative social science where researchers may endorse data sharing and the calls for open science but with a hesitation that results from conflicting desires and obligations. How can one be open and share research material while respecting moral, ethical and legal requirements not to share? Many of the conflicts can be resolved by adding delay into the process: summarized in an Augustinian inspired archival prayer "to make me open—but not yet." There is warrant for this but existing infrastructures may not easily accommodate it not so much because of technical problems but more because they are not resourced to plan over a long enough timescale.

Social science data involves other sorts of trust than those implicit in the call for contributions to the Frontiers Topic "Trust and Infrastructure in Scholarly Communications" (https://www.frontiersin.org/research-topics/18191/trust-and-infrastructure-in-scholarly-communications?utm_source=F-RTM&utm_medium=CFP_E2&utm_campaign=PRD_CFP_T1_RT-TITLE#overview and in statements promoting the idea of open data, for example, by the Open Data Institute RRID:SCR_021681 https://theodi.org/about-the-odi/). These are relevant to the design, implementation and funding of infrastructures supporting research beyond medical, chemical and physical disciplines. My areas of concern are qualitative social sciences as well as many humanities disciplines including oral history and life history research where, as a referee points out, "the tellers want their stories told" or the participants contribute because they want to contribute to the general development of science (Kuula, 2011). What is distinctive about these disciplines is that as an essential part of the research process, the researchers have human, interpersonal relationships with their interlocutors (which makes terms like informants, interviewees, research subjects or collaborators misleading if not inappropriate).[1] Also much of the importance of the material gathered lies precisely in the aspects that are destroyed by anonymization. Writing as an anthropologist I talk about anthropology but the discussion is relevant to all those undertaking qualitative research with living humans, especially those where more or less structured interviews are not the main form of data collection.[2]

OPEN ACCESS

Edited by:
Daniel. W Hook,
Digital Science, United Kingdom

Reviewed by:
Libby Bishop,
GESIS Leibniz Institute for the Social Sciences, Germany

***Correspondence:**
David Zeitlyn
david.zeitlyn@anthro.ox.ac.uk

Specialty section:
This article was submitted to Scholarly Communication, a section of the journal Frontiers in Research Metrics and Analytics

Received: *05 July 2021*
Accepted: *13 September 2021*
Published: *14 October 2021*

Citation:
Zeitlyn D (2021) For Augustinian Archival Openness and Laggardly Sharing: Trustworthy Archiving and Sharing of Social Science Data From Identifiable Human Subjects. Front. Res. Metr. Anal. 6:736568. doi: 10.3389/frma.2021.736568

[1]This has led one pseudonymous author to suggest they should be coauthors. See Luther Blissett mss.

[2]It may even apply to those in humanities discussing the work of living authors: I suspect few in disciplines such as English etc have considered the relevance of GDPR and data sharing for their research.

Trust type	Gloss	Infrastructural implications
$Trust_1$ (T_1)	Participants trust researcher	Long Term Embargos (census style)
$Trust_2$ (T_2)	Researchers trust participants	Readers can access data (data sharing)
$Trust_3$ (T_3)	Readers trust researcher	Readers can access data (data sharing)

I think it helps to distinguish different types of trust since they need different sorts of archiving infrastructure. Discussions of data sharing address mainly the third of these.

As institutionalized by research ethics boards when a researcher starts talking to an "informant" a first step is to work through a consent process. As has been long pointed out (Neale and Bishop, 2012; Bell, 2014), for qualitative research subjects such as anthropology this is often inadequate and misleading but it continues, driven by forms of institutional inertia and what Wynn and Israel (2018) describe as a *fetishization* of signed consent forms. Commonly, researchers promise to respect the confidentiality of what they have been told by anonymizing the individuals concerned. Until relatively recently this was taken (usually tacitly) to mean anonymizing *in publications* (strictly this is pseudonymisation since versions with identifiers were retained). The GDPR in the EU (and in UK) and its implementation by administrators in research ethics boards has raised the possibility of its application to the actual data collected on which publications are based (see Yuill, 2018) despite its explicit provisions for data archiving. For many researchers the "research material" (a more generous term than data) may be scrappy, poorly structured and may seem inadequate to others without the advantage of having taken part in the research process. For the original researcher they may be more *aide memoire* than hard data but not necessarily the worse for this. Sharing such material may not help assessment of publications based on it. With apologies for a double negative, this is *not* to say they are untrustworthy. Between researcher and reader there is an "ethnographic pact" (paralleling Philippe Lejeune's "autobiographical pact" Lejeune, 1989).[3] For all that sharing such material may have other benefits. It makes the material available for use by other researchers from other disciplines, and to those from the places where it may have originated. Usefulness/comprehensibility is not all or nothing property. Unquestionably the original researcher has hugely privileged access but that does not mean no one else can make any sense at all of fieldnotes etc (another intentional double negative). And in terms of *trust* a refusal to share raises the question of what they might be hiding while an openness to sharing implies of itself an degree of trustworthiness, even if the sharing may be far in the future. In other words, many of the reservations researchers may have about data sharing can be addressed by making very long embargoes the norm.

In publications from the harder sciences, issues of trust revolve around whether the published results and the text discussing them are trustworthy (T_3). To help assess this access to the underlying data is helpful. However, other issues of trust accrue in qualitative social sciences before similar questions about the trustworthiness of publications can be asked. Usually tacit or unspoken, there are different questions of trust underlying consent giving and subsequent interactions: the respondents or interlocutors have to assess the trustworthiness of the researchers (T_1): can these individuals and can the promises they make be trusted? These are different points to those about the trustworthiness of publications. This is fundamental since, of course, if participants really don't trust the researchers then they will not continue and no material will be collected so no publications will result. If they have doubts they may still participate but may provide more or less unhelpful or misleading answers (Kuula, 2011:15). In some cases researchers acting in good faith have made promises they have not been able to keep. The clearest and most notorious instance of this is case of an oral history project about the "troubles" in Northern Ireland which was archived in Boston (United States). The researchers sought to resolve the issue of informant confidentiality by promising former participants in the conflict that their interview data would not be released until after the interviewee's death. But the Northern Irish police used legal subpoenas to break these promises. The researchers had made promises in good faith that in the end legal process meant they were not able (were not allowed) to keep.[4]

In cases where consent is withheld, the potential participant has made a judgment that, for whatever reasons, the researcher cannot be trusted (T_1), and no data is collected. Conversely, once someone has agreed to participate, the researcher has to make judgments about them: is their account to be trusted (T_2)? This may change over time. A claim to have special access and insight may be initially accepted but then later changed as it becomes clear that the person in question does not have access and indeed may be widely regarded by others in the community as a notorious liar. Anthropologists have to manage being simultaneously credulous fools and ultimate cynics. I call this professional bad faith or adopting a position of ironic detachment; Johannes Fabian talks about "the duplicity without which ethnographic research would be impossible—a duplicity which makes us cross borders but not without establishing a record that lets us return to our professional roles and habits" (2008: 6).[5] That said there is a lot that can be learned from liars such as what is deemed to be a plausible alternative version of events and motivations. Disputes in politics (and elsewhere) are revealing about process no matter what

[3]Pursuing this is in danger of taking me off topic so I will not further discuss wider purposes and limits of data sharing here.

[4]See Lowman and Palys (2013) as well as news coverage such as https://www.bbc.co.uk/news/uk-northern-ireland-27238797 and also https://www.irishtimes.com/news/crime-and-law/attempt-to-access-former-ira-man-s-boston-college-tapes-replete-with-errors-court-told-1.3357750 Both accessed 26 Sept 2018.

[5]The ambiguities involved have also been discussed as the part of the dilemmas of 'insider/outside' positionality (Zavella 1996).

actually did occur or who did what to whom. Such material even if considered to be unreliable (T_2) may yet be revealing hence useful and so is worth keeping.

When it comes to data sharing a researcher might well want to anonymize versions (I stress *versions*: not everything) of research material so it can be made accessible relatively soon after the data has been collected–paralleling the ways that the UK Census Office removes personal ID information from some material from recent censuses. Importantly, even though the publicly accessible version may be anonymized, if the key to the anonymization exists, then legally it remains personal data and subject to GDPR rules (and strictly should be called a pseudonymized version).[6] I very much hope researchers retain the keys because the *longue-durée* view is, in my opinion, that the full data should be retained and made available but only in the long term. Perhaps, as with census data, this should default to being 100 years after collection. Even this, in some circumstances, may have to be restricted but for most material it is plausible to suggest that access could be enabled with only a very light administrative touch.[7] The data may still be described as FAIR (meeting principles of findability, accessibility, interoperability, and reusability) only a delay has been introduced into the aspects of accessibility and reusability.

An example of why it is important to retain complete data records including the names may be found in the title of a book by Michel Foucault based on research in judicial archives.

I, Pierre Rivière, having killed my mother, sister and brother (1835) Published 1973 (filmed in 1976).

I think similar books and films should be possible in 100 years' time with the full names given rather than:

I, Homme03061835, having killed my mother, sister and brother.

In other words I am arguing that any anonymization should be reversible. As already mentioned technically this is pseudonymization. Either a complete version of the material must be kept without pseudonyms, or a key retained to the pseudonyms used so that at a much later date the pseudonymised file can be re-edited to reinsert the actual names and other identifying information. (I suspect that in practical terms it is easier to keep two versions of the file but data managers may take a different view).[8]

ETHICAL CONTRADICTIONS

There are contradictions and conundrums in the ethical position of archiving. In related work Kirsten Bell (Bell, 2018 and Bell and Wynn, 2020) considers conflicting relationships and stances towards consent and participation. She discusses how researchers and those they work with can develop a "procedural ethics framework." Neale and Bishop (2012) describe an example of such a framework for research archives. This addresses the conundrum that prior "informed" consent cannot be given for future research by unknown others asking unknown research questions.[9] Bishop has also discussed other ethical concerns about qualitative data archiving and shown how these can be addressed, hence enabling data archiving and sharing (2009). In the short and medium terms archivists must act as trusted proxies,[10] implementing the processes that the participants consented to, and serve as trusted gate-keepers to the data long into the future (with all the administrative and resource implications that go with this). Russell and Barley (2019) raise questions about "who owns research data" which has a more direct bearing on data archiving.

Consider the possibly conflicting, certainly different, responsibilities that researchers have towards:

- Informants
- Colleagues
- Informants at a later date and their descendants
- Future colleagues (including older versions of themselves)
- Research funders/institutions

These responsibilities and obligations point in different, mutually conflicting, directions. Respect for the privacy of individuals suggests anonymization, closure or not archiving (and data destruction), whereas respect for the descendants of those individuals in the distant future suggests openness and archiving for the long term.

To make concrete the importance of preserving data such as photographs without blurring or other forms of anonymization let me give a clear example of why it would be unethical to destroy or fully anonymise data. Consider the following field photograph that I took in May 1986 (**Figure 1**).

Yanele Blandine, the girl in the white headscarf, is long dead. Her son, Serge Donat, had no photograph of his mother until recently, when I was able to send him a copy. Imagine his response if I had said either:

> I took a photograph of your mother long ago but I cannot send you a copy because I destroyed it on completing my doctorate.
> Or
> I took a photograph of your mother long ago but I cannot send you a copy because I do not have her permission to share it.
> Or

[6]Note that if truly anonymized and there is no key to reverse the process then they are no longer personal data and therefore not subject to GDPR regulations.

[7]Users may still have to register and the archives may wish their role to be acknowledged in any resulting publications.

[8]A Frontiers editor has suggested that cryptographic protocols or infrastructures such as the Blockchain could provide mechanisms for deterministically reversible psuedonymisation (i.e. reversible at a pre-destined point in time based on either computational difficulty to crack anonymity or on a smart contract). I find this an intriguing suggestion but fear it may involve too complex a bet on future infrastructures to achieve wide take up. Keeping two copies, only one of which is pseudonymised seems simpler.

[9]An example might be a linguist studying the phonetics of vowel quality in interview sound recordings. The linguist might be interested because the interviewer had recorded where participant and their grand-parents lived but be entirely uninterested in the content of the discussion that was recorded.

[10]Hence the importance of certification schemes for repositories such as Core Trust Seal (RRID:SCR_021679 https://www.coretrustseal.org/) by which they can be assessed as 'Trusted Digital Repositories'.

FIGURE 1 | Caption: L-R: Barmi (alive 2018), Nde Donat (d 1987?), Ndignoua Salomon, Suzana Thia (alive 2018), Ngon Luise (died), Blandine (died), Dissi (Mougna's child, died) Jacqueline (alive 2018), the two children in the front: the boy is Kounaka Fidèle (Salomon's junior brother (alive 2018), and the girl is Mbitti Josephine (alive 2018). The names were provided in 2018 by Serge Donat, son of Blandine (photo David Zeitlyn Reference: 24_34. jpg 01/05/1986).

> I took a photograph of your mother long ago but I can only send you a copy with her face blurred because I do not have her permission to share it.

I would suggest that all of these possible responses as suggested by the ethos of many ethics panels would be inhumanly rude and indeed would be deeply unethical.

There is a further almost banal point to be made—unnecessary for most of this readership but important nonetheless. Openness and Archiving does *not* mean:

1) giving free access to all comers
2) putting all material online without regulation (even for digital archives)

The UK Data Archive (RRID:SCR_014708 https://www.data-archive.ac.uk) is a digital repository for social science data. It includes quite a lot of anthropological material, and other qualitative data. It has developed ways of working with the sorts of unstructured material that anthropologists tend to produce. Access is not open in the sense of being uncontrolled: users have to register and in some cases *they* have to give a form of consent,[11] so they enter into the same sort of relationships of trust with the informants that the original researcher did. Moreover, material placed into the archive can in some circumstances be embargoed (usually only for a relatively few years). However, the UK Data Archive has also worked with the UK Census to develop protocols (and technical solutions) for "Secure Access Points" which enable remote access to highly sensitive datasets in constrained and controlled fashion. To summarize: Digital Archives can have a wide range of different types of access—ways that have more resemblance to access to physical archives than most digital evangelists might imagine. My suggestion is that a default for qualitative data might be 100 years, then in some unusual cases longer embargoes could still be called for, in many others much shorter ones. Even within the embargo period tightly controlled access could be arranged (again as is currently possible for census data). This is to suggest that what could be called "archival hesitancy," a suspicion about data archiving (sharing) among qualitative social scientists, can be, partly, addressed by shifting to a census-style default from which one could discuss variations. Such an environment would only be possible if research funders allow data to be archived under census-type protocols *and* if archiving services such as the UK Data Archive are given the resources to enable this.

In short anthropology and its fellow subjects can manage to resolve the injunctions of morality and ethical responsibility to our research collaborators with the conflicting pressures to be open and share our data by resolving to be open in ways that parallel how the census is open. A recent example of how openness can be achieved over a century or more is the publication of Alfred Haddon's diaries (Herle and Philp 2021). This was undertaken in collaboration with the descendants of the people Haddon worked with in the Torres Straits in 1888 and 1898. This is the timescale we need to start

[11]Those reusing the data have to sign a User Agreement from the archive with binding terms and conditions that define what they can, and cannot, do with the data they access. This has parallels with the original consent process.

thinking about, this is the timescale that openness in anthropology and cognate disciplines can accept as ethically consistent with the promises we make. The challenge now is to resource infrastructures that can accommodate ethical complexity: to enable us to be open but not quite yet (Parry and Mauthner, 2004; Bishop, 2009; Blisset, unpublished[12]).

ETHICS STATEMENT

Written informed consent was obtained from the individuals and/or minors' legal guardian/next of kin for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

The example cited in this piece and surrounding text has been taken from (Zeitlyn, Forthcoming 2022) with permission of the author. The Author is happy to acknowledge helpful and clarifying discussions at the Colloquium "*Valoriser les archives des ethnologues Usages contemporains des collections*" BNF, Paris, 4-5-6 October 2018. Some of the issues featured in a talk also in October 2018 at All Souls College, Oxford for the *Open Science Seminar* convened by Lisa Lodwick and Jasmine Nirody and at L'école d'été UMC—Marseilles "*Understanding Mediterranean Collections,*" 17 July 2019. Another, shorter, version was presented remotely in November 2018 at a AAA Roundtable "*Approaches to Expanding the Use of Anthropological Archives*" convened by Diana Marsh. The Author is very grateful for all the invitations and for the feedback from these presentations and on-going conversations with Kirsten Bell, who also made extremely helpful comments on a draft.

REFERENCES

Bell, K. (2014). Resisting Commensurability: Against Informed Consent as an Anthropological Virtue. *Am. Anthropologist* 116 (3), a–n. doi:10.1111/aman.12122

Bell, K. (2018). The 'problem' of Undesigned Relationality: Ethnographic Fieldwork, Dual Roles and Research Ethics. *Ethnography* 20 (1), 8–26. doi:10.1177/1466138118807236

Bell, K., and Wynn, L. (2020). Research Ethics Committees, Ethnographers and Imaginations of Risk. *Ethnography* 1466138120983862, 146613812098386. doi:10.1177/1466138120983862

Bishop, L. (2009). Ethical Sharing and Reuse of Qualitative Data. *Aust. J. Soc. Issues* 44 (3), 255–272. doi:10.1002/j.1839-4655.2009.tb00145.x

Herle, A., and Philp, J. (2021). *Recording Kastom: Alfred Haddon's Journals from His Expeditions to Torres Strait and New Guinea, 1888 and 1898*. Sydney: Sydney University Press.

Kuula, A. (2011). Methodological and Ethical Dilemmas of Archiving Qualitative Data. *Iq* 34 (3-4), 12–17. doi:10.29173/iq455

Lejeune, P. (1989). *On Autobiography*. Minneapolis: University of Minnesota Press.

Lowman, J., and Palys, T. (2013). The Betrayal of Research Confidentiality in British Sociology. *Res. Ethics* 10 (2), 97–118. doi:10.1177/1747016113481145

Neale, B., and Bishop, L. (2012). The Ethics of Archiving and Re-using Qualitative Longitudinal Data: a Stakeholder Approach. Timescapes Methods Guides Ser. [Online] 18. 12(1), Available at: http://www.timescapes.leeds.ac.uk/assets/files/methods-guides/timescapes-neale-ethics-archiving.pdf.

Parry, O., and Mauthner, N. S. (2004). Whose Data Are They Anyway? *Sociology* 38 (1), 139–152. doi:10.1177/0038038504039366

Russell, L., and Barley, R. (2019). Ethnography, Ethics and Ownership of Data. *Ethnography* 21 (1), 5–25. doi:10.1177/1466138119859386

Wynn, L. L., and Israel, M. (2018). The Fetishes of Consent: Signatures, Paper, and Writing in Research Ethics Review. *Am. Anthropologist* 120 (4), 795–806. doi:10.1111/aman.13148

Yuill, C. (2018). 'Is Anthropology Legal?' *Anthropol. Action.* 25 (2), 36–41. doi:10.3167/aia.2018.250205

Zavella, P. (1996). "Feminist Insider Dilemmas: Constructing Ethnic Identity with Chicana Informants," in *Feminist Dilemmas in Fieldwork*. Editors L.W. Diane and D. Carmen Diana (London: Routledge), 138–159.

Zeitlyn, D. (Forthcoming 2022). Archiving Ethnography? the Impossibility and the Necessity. Damned if We Do, Damned if We Don't. *Ateliers d'anthropologie* 51.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

[12]Blissett, L. 2006. The Hau of the Paper and Dividual Authors: Reimagining Authorship in Anthropology. Unpublished manuscript.

frontiers in Research Metrics and Analytics

POLICY AND PRACTICE REVIEWS
published: 01 November 2021
doi: 10.3389/frma.2021.746514

Approaching Trust: Case Studies for Developing Global Research Infrastructures

Heather Flanagan[1†], *Laurel L. Haak*[2,3*†] *and Laura Dorival Paglione*[1,4†]

[1]Spherical Cow Consulting, Vashon, WA, United States, [2]Ronin Institute, New York, NY, United States, [3]Mighty Red Barn, Townsend, WI, United States, [4]Laura Paglione, LLC, Rego Park, NY, United States

Trust is a core component of collaboration. Trust is a local phenomenon, and scientific research is a global collaborative, its impact multiplied through open exchange, communication and mobility of people and information. Given the diversity of participants, local policies and cultures, how can trust be established in and between research communities? You need transparent governance processes, thoughtful engagement of stakeholder groups, and open and durable information sharing to build the "stickiness" needed. In this paper we illustrate these concepts through three trust building use cases: ORCID, Global Alliance for Genomics and Health, and SeamlessAccess, platforms sharing an identity and access technical service core, painstaking community building, and transparent governance frameworks.

Keywords: multi stakeholder initiative, trust and reciprocity, stakeholder theory (normative foundations), public infrastructure, IT governance (ITG)

OPEN ACCESS

Edited by:
Stephen Pinfield,
The University of Sheffield,
United Kingdom

Reviewed by:
Houqiang Yu,
Sun Yat-sen University, China
Jiban K. Pal,
Indian Statistical Institute, India

***Correspondence:**
Laurel L. Haak
laure@mightyredbarn.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Scholarly Communication, a section of the journal Frontiers in Research Metrics and Analytics

Received: *24 July 2021*
Accepted: *11 October 2021*
Published: *01 November 2021*

Citation:
Flanagan H, Haak LL and Paglione LD (2021) Approaching Trust: Case Studies for Developing Global Research Infrastructures. Front. Res. Metr. Anal. 6:746514. doi: 10.3389/frma.2021.746514

1 INTRODUCTION

Research is a global endeavor of iteration and collaboration. Research requires trust-building: shared understanding of process, access to source data, and points of validation. A number of trust structures are used by researchers: disciplinary societies cohere practices among researchers, educational degrees and institutional affiliation are proxies of trust, as is publication of research findings in status journals (Haak and Wagner, 2021).

These trust structures require interactions among many stakeholder groups, operating within and across disciplines, institutions, and countries. This is where research infrastructures come into play. These infrastructures support knowledge sharing across stakeholder borders, and at the best of times create a foundation for collaboration (Edwards et al., 2013; Haak et al., 2020). Examples of global-scale research infrastructures include article indexing platforms, researcher profile systems, federated identity systems, data repositories, and global data collection systems. More recently, the research community has started to pay more interest to the governance and sustainability aspects of these infrastructures (Bilder et al., 2020; Skinner, 2019). Organizations such as the Research Data Alliance have fostered cross-disciplinary self-organization of community stakeholders, out of which have come truly amazing consensus rules of behavior—principles of findability and accessibility (Wilkinson et al., 2016), as well as responsibility and ethics (Carroll et al., 2020)—that can be applied to infrastructures to improve research rigor and reproducibility and ultimately improve trust and engagement in the research process.

In this article, we share our "in the trenches" experiences of how these principles, when applied in practice, can drive research infrastructure adoption. Infrastructure is more than a platform, it is a public good, so we need to ensure its accessibility and sustainability. How it is constructed, governed,

and maintained requires intentional engagement and alignment of diverse stakeholders across social and economic factors to maximize trust, utility and impact on public welfare (Dhanshyam and Srivastava, 2021). What we have found is that without alignment and engagement, trust-building suffers. The lower the trust—even for a really strong technology that is desperately needed by the research community—the steeper the uphill push to adopt and implement the infrastructure.

2 APPROACH

Infrastructure adoption depends on how well it serves its intended audience. There are multiple factors involved in building the trust needed for adoption: identification of stakeholders, development of services that respect and meet the needs of these stakeholders, governance (including openness and transparency), communications and marketing, start-up funding, and sustainability (including processes and recurring funding). In this article, we explore stakeholder alignment and engagement as fundamental components of trust-building. We take an ethnographic approach, which involves an emphasis on the "emic" or insider view, rendered as first-person case-study accounts (Eriksson and Kovalainen, 2014). This methodology has been used effectively in studying development of new-technology-based services (Eriksson et al., 2011).

We chose research infrastructure initiatives that share a core of painstaking multi stakeholder community engagement around individual privacy, and also that illustrate the impact of different stakeholder economic and social motivations on alignment and engagement, and, ultimately, on infrastructure adoption. We each have been intensively engaged in the formation of the global-scale research infrastructures SeamlessAccess, Global Alliance for Genomics and Health, and ORCID. Here we share our experiences and impressions of their early-stage development.

2.1 SeamlessAccess

Heather Flanagan was the Pilot Coordinator for the RA21 Academic Pilots and, later, served as Program Director for the follow-on effort, SeamlessAccess. She brought to the table expertise in federated identity and a strong network in the academic identity federation community, supporting both community engagement and technical specifications work. Laure was engaged as a stakeholder, providing input on multi-stakeholder governance principles. Laura was engaged as a subject matter expert on researcher privacy and end-user design. (Note that business ethnography often recommends that researchers embed into their research projects as team members, taking on roles such as project manager, to gain oversight into all aspects of an initiative without necessarily shaping it.) The case study was shared with the SeamlessAccess management team and their comments have been incorporated into the narrative.

2.2 Global Alliance for Genomics and Health

Laura Paglione was engaged with the Data Use and Researcher Identity working group as a volunteer and subject matter expert on researcher identifiers and federated identity. She also served as the Co-chair of the Equity, Inclusion, and Diversity Advisory Group that has a goal to understand, encourage, and support broad participation in the volunteer groups participating in this effort. At the leadership level, she made connections between global standards initiatives and community engagement approaches. The flavor of working group discussions is reflected in the storytelling approach and metaphorical examples used in this case study.

2.3 ORCID

Laurel Haak was the founding Executive Director of ORCID, and Laura Paglione its founding Technical Director, employees 1 and 2, respectively. Both were deeply engaged in starting up organizational operations, establishing participatory co-design culture, engaging stakeholders from pre-launch to implementation and through to specifying initial versions of a certification program, and were principal architects of the ORCID Trust program. Perspectives of ORCID team members and stakeholders are incorporated by reference to blog posts and primary documents.

3 MANUSCRIPT

3.1 Case Study: SeamlessAccess

Trust is local. How do you then approach building a global information technology service? How do you make the work local enough for global participants to trust the outcomes of the collaboration? How do you manage expectations when the technology is too complex for anyone other than experts to understand? A number of studies in public sector organizations show that transparent governance processes and thoughtful invitations to key stakeholder groups are key to the adoption and sustainability of information technology (Wiedenhöft et al., 2020). Trust can generate a certain "stickiness" when individuals feel their community has been heard (AlHogail, 2018), but it depends on ongoing and open engagement within and between stakeholders.

In this case study, we explore how building trust is complicated by poorly understood technology, uncertainty regarding ownership, and loosely aligned stakeholders. We will also look at the additional complexities of rebuilding broken trust between stakeholder communities.

3.1.1 Introducing SeamlessAccess

Technology introduces a wonderful world of online access opportunities. Early adopters, however, often find that the user experience is an afterthought to the technical implementation of an idea. One example of this is the world of federated identity, which offers students and researchers the ability to log in (authenticate) to an online service (such as a library catalog) via the credentials (such as their username and password) managed by their home organization (such as their college, university, or company), also known as an Identity Provider. There is quite a bit of value here: users do not need to remember yet another password, services do not need to maintain user

affiliation records, services can enable sign in from multiple organizations at once through Identity Provider federations, and the impact of account compromise within the service provider is limited. Federated identity has been available, particularly in academia, for over two decades.

Unfortunately, while federated identity offers powerful benefits, the complicated user experience associated with the technology has caused many scholarly communications services—particularly publishers—to avoid implementing it. In 2015, however, publishers decided that the benefits of federated identity were great enough to warrant addressing the user experience. Publishers were feeling pain on multiple levels. A long-established process of maintaining lists of customer IP addresses to indicate which organization the user was from was becoming untenable because they were no longer the stable data they were before computing went mobile. Publishers heard demands directly from users to improve off-campus access to content (away from campus IP addresses). And, publishers were experiencing a business model threat as pirated material became more easily accessed than legally provided content.

These factors led to a community collaboration called "Resource Access in the Twenty-First Century" (RA21), and then later to an operational service called SeamlessAccess (NISO, 2019). RA21 focused on developing "recommendations for using federated identity as an access model and improving the federated authentication user experience." Those recommendations led to the creation of an operational service: SeamlessAccess. SeamlessAccess, in turn, provides a significantly improved user experience for helping users find their Identity Provider in a sea of options when trying to log in to a website that supports federated authentication.

RA21 identified business needs and user demand for a change in accessing scholarly content online. However, the demand for change was not sufficient to encourage trust in the thing that promises that change. The stakeholders involved—the scholarly publishing and the academic library communities in particular—were aligned on neither the problem that needed to be solved nor the solutions possible to solve it.

RA21 and later SeamlessAccess were driven largely (but not entirely) by the scholarly publishing community, in collaboration with researchers and campus IT. These relationships are reflected in the current SeamlessAccess governance body: a coalition of GÉANT, Internet2, the National Information Standards Organization (NISO), and the International Association of STM Publishers. Publishers were hearing from their readership and from the researchers on their editorial boards that access methods needed to change. Publishers engaged with the federated identity and campus library communities, but publishers and campus IT tend to exist in tension with the library community—largely because of differing perspectives on content access. A particular issue for librarians was concern about the impact a move to federated identity might have on a user's right to privacy. In addition, given ongoing budget challenges, librarians had little desire to argue within their own organizations for the resources necessary to support implementation of federated identity. Stakeholders in the scholarly communications ecosystem were not aligned.

3.1.2 Components of Trust

When working with technology, trust comes from more than understanding the technology itself. It also requires trust in how the service that uses the technology is managed and how the service engages with its stakeholders. And of course, each piece - technical understanding, governance, and stakeholder engagement - is interdependent. SeamlessAccess has focused on stakeholder engagement as the core of its trust model, out of which comes technical requirements and governance decisions.

3.1.2.1 Technical Understanding

> "Any sufficiently advanced technology is indistinguishable from magic."—Arthur C. Clarke.

A century ago, business owners probably looked askance at this newfangled thing called an automobile. They likely wondered how they were supposed to let their business depend on this new thing when they had no idea how it even worked, how to maintain it, or how to even pay for it. Eventually, the infrastructure matured enough to make the automobile the ubiquitous thing it is today. Federated identity is also in that early stage where the infrastructure isn't ubiquitous enough that people are just willing to trust that it works.

Like an automobile, federated access is quite complicated under the hood. The trick is to help stakeholders understand what it can do and how it can help solve their problems, without digging too deep into the technology details. There's a caveat to that, however: while you don't want to overwhelm stakeholders with unnecessary detail, that detail must be available for those who want to learn more. The transparency of the technology is critical, while understanding the technology is not.

SeamlessAccess has focused much of its efforts on building an outreach program to educate different stakeholders about federated identity. The Outreach Committee in particular brings in representatives from the primary stakeholder groups to maintain perspective on what kinds of questions people are asking, and determine how to speak to their concerns. The purpose here is to effectively translate the technology for the understanding and benefit of end users.

But there's another component: the technology is not beneficial if it is not implemented. Prior to RA21 and SeamlessAccess, there was no standard way to present federated access to the user. Every service presented the information in different, and often very confusing, ways. On the one hand, SeamlessAccess exists to improve that user experience. On the other hand, the service relies on the organizations integrating it to offer feedback on what's working for their users and what isn't. To this end, SeamlessAccess is helping organizations that want to implement the technology understand what to do and, as importantly, why to do it that way. Allowing for the flexibility for integrators to experiment a bit with what will work best for them is another way to encourage trust in the service. The service has been tagged as a 'beta' product because changes are expected as we iterate on user testing and integrator feedback.

3.1.2.2 Governance

Governance is how community projects make decisions that represent stakeholder needs, wants, and desires. A trusted governance model requires that each stakeholder must see someone in the governance group, either a person or an organization, that they can identify with. They must also understand the motivations and business model behind the service. Is the service for profit? If so, who is making money off of it? Or is the service not-for-profit? If so, how is it being sustained?

In the case of SeamlessAccess, governance happens in layers. There is a core governance team that focuses on legal and financial details, such as developing the by-laws needed to make the project a not-for-profit legal entity. This core group reflects the diverse stakeholder communities that will use the service and the organizations providing the resources to support the operations of the service. Next, there are several committees, including an Advisory Committee, an Outreach Committee, and a Technical Steering Committee. And third, there are working groups to focus on specific challenges, including the Contract Language Working Group and the WAYF Entry Disambiguation Working Group. This layered approach increases the opportunities for people and organizations to engage at a level comfortable to them. The more ways to actively engage stakeholders in governance, the more opportunities to build community trust in the service. The community needs to see that all parties working towards a common goal - building and maintaining a service that benefits all stakeholders.

3.1.2.3 Stakeholder Engagement

The RA21 project polarized the library community, between supporters of federated identity who saw it as a way to enable access to content, and detractors who felt federated identity would lead to a loss of user privacy. SeamlessAccess inherited some, if not all, of that tension. Knowing that trust-building is a core success factor for this project and is a key aspect of its sustainability, the SeamlessAccess governance group worked to engage library stakeholders in design and governance discussions and made it a priority to offer additional education about federated identity.

It is worth noting that stakeholder tension was not restricted to publishers and librarians. Another tension that impacted efforts to promote federated identity as a solution for remote access to content was the tension between campus librarians and campus IT departments. Many librarians have existing infrastructure that lets them manage access to content. Shifting to a federated authentication model would require deeper collaboration between the library and campus IT. These two departments usually exist in entirely different parts of the campus organizational structure, with different priorities, funding, workflows, and users to support. Relations between campus libraries and IT are often weak at best, and antagonistic at worst.

3.1.2.3.1 Getting Stakeholders to the Table. With so many stakeholders, building trust starts with one of the most difficult steps: getting all the stakeholders to the table. If a stakeholder group is not willing to have a conversation, building trust with that group is simply impossible. Of course, once the groups are at the table, there needs to be a concerted effort to build credibility and understanding. Why should anyone at the table trust the convener, much less anyone else participating? NISO and the Research Data Alliance have been particularly effective in this arena (Carpenter and Horton, 2012). A shining example of cross-stakeholder trust building in the scholarly community is the Scholix Initiative, which engages across campus, governmental agencies, identifier infrastructures, data centers and publishers to create an internationally-used method to link research data with the literature (Cousijn et al., 2019).

3.1.2.3.2 Inheriting Stakeholder Dependencies. In the case of SeamlessAccess, to establish trust required bringing together stakeholders who understood the global federated identity infrastructure, people who understood the demands on campus IT, representatives from content providers, and content stewards. All were needed to build the service. However, these stakeholders can sometimes be only indirectly impacted by the value of the service offered. When the service needs the stakeholders more than the stakeholders need the service, it creates lopsided value. This has been a particular issue for SeamlessAccess.

For SeamlessAccess to function, the library itself or its campus needs to be a member of an identity federation. Campus IT and identity federations themselves are not directly impacted by the service SeamlessAccess offers. They have no pain point that this service can directly resolve; they have less motivation to come to the table. And yet, without their cooperation, federated identity is not possible. Without federated identity, there is no SeamlessAccess service.

3.1.2.3.3 But what About the User? As with many technology-based projects, there are many stakeholders who need to have a say in how the project progresses. SeamlessAccess is about creating a better experience for the individual actually trying to access protected content. So, where is the individual? Who represents them? While they are at the forefront of consideration, they do not have a direct seat at the table. The responsibility for representing the individual falls to the other stakeholders involved in the project.

Individuals are the most difficult group to engage with because they cannot be easily described in a way that would allow any one person or organization to represent all their needs. On a single campus, an individual user could be an undergraduate student, a graduate student, a faculty member, a researcher, a staff person, a contractor, a visiting student, a visiting scholar, or even a walk-in patron to the library. They all have different perspectives and may have entirely different needs when it comes to technology.

In the SeamlessAccess scenario, many of the stakeholders laid claim to being the representative for the individual user. Despite that representation, the needs identified are still in conflict; the stakeholders each see one aspect of a very complex ecosystem, resulting in perfectly valid needs that are in conflict with each other. Libraries staunchly defend privacy rights for the user in the face of users regularly giving away information about

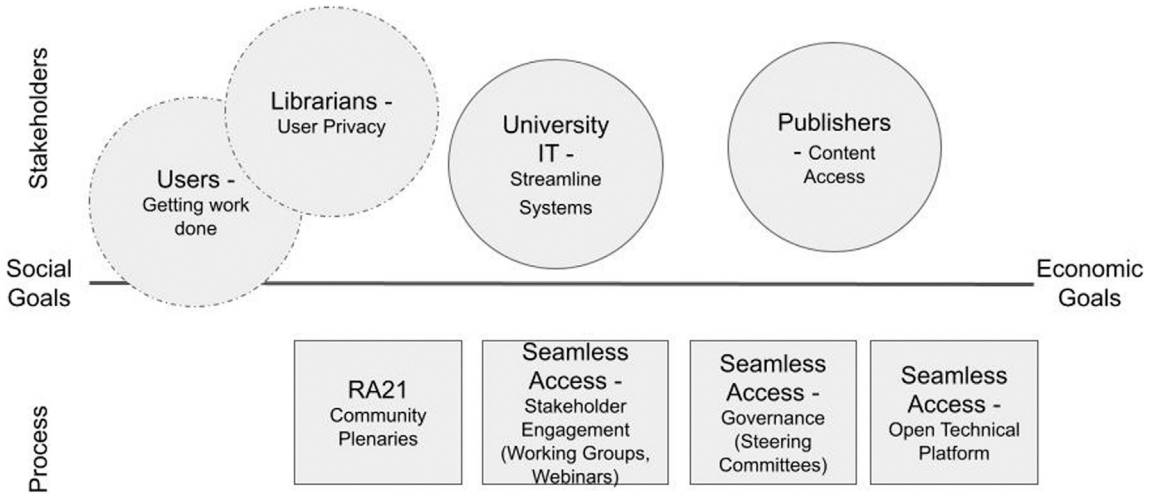


FIGURE 1 | RA21 and Seamless Access were driven by publishers, who have an economic incentive to improve content access. The initiatives employed community outreach tactics and developed a technical platform, but did not have a transparent multi-stakeholder governance process for deciding on requirements.

themselves if that's the easiest way to get to the material they need. Campus IT has technical control but not authoritative control over attributes released by the campus. Publishers interact directly with the user and have an entirely different perspective on personal data collection. They are all accurate, even when they are in conflict.

3.1.2.3.4 Building Trust Through Engagement. Meaningful engagement efforts will slowly erode distrust between stakeholders and build trust in the service offering. In an ideal world, as the stakeholder groups see each other engage in good faith, and users benefit from the service, we will see tension that may exist between groups decrease. By educating and explaining what technology can and cannot do—and by ensuring stakeholders a seat at the decision-making table—it is possible to bridge the fact that sometimes, stakeholders simply do not trust the intent of other stakeholders (**Figure 1**).

Of course, the world is rarely an ideal place, particularly over the last year. As conferences and in-person meetings became a thing of the past, opportunities to let food, drink, and a full view of a person's body language smooth human engagement have not been an option. SeamlessAccess has worked to engage stakeholders at virtual conferences and webinars, through targeted white papers and short videos explaining the service. We also collectively recognized that users were no longer able to live and study on campus and had to have better solutions for remote access. Members of various stakeholder communities that were not bought into the premise of federated access have experienced the need in ways they never have before. Suddenly, there has been a much stronger motivation to find ways to trust the technology, the other stakeholders, and the service itself.

3.1.3 Learnings

SeamlessAccess focused on four challenges when considering how to build trust in the service:

- The service relies on complex technology and yet we need to make it transparent and trustworthy without expecting everyone to understand the details.
- Given the service is still under development, the user value may not be clear to everyone and is subject to change.
- The service needs the stakeholders more than the stakeholders need the service; it creates a lopsided value proposition.
- The service operators do not have direct access to the end-users, resulting in second-hand interpretations from various stakeholders on user needs.

We have addressed these options by focusing on educational outreach opportunities, a layered governance model, and strong stakeholder engagement. Each of these activities will require continued action to maintain the trust we've built and to continue to grow trust, and through it, service adoption.

3.2 Case Study: Global Alliance for Genomic Health

It's 8:00 AM eastern on Wednesday morning, and the Data Use and Researcher Identity (DURI) work stream of the Global Alliance for Genomic Health (GA4GH) is meeting virtually. The group meets every 6 weeks to develop standards for computers to facilitate researcher access to genomics datasets. Today they are working on a standard for computers to understand if the person signing into the system is considered to be a bona fide researcher.

The systems in question are sensitive ones. They house genomics data that has been compiled for research purposes. Access to these data is restricted. Data Access Committees (DACs) review requests for access to the data, weighing information about the person requesting access (the data user)

and the type of study being conducted (the study topic). The information in these datasets is de-identified so that it is not possible to tie it to the specific people (genetic data contributors) from which it was collected or derived. In addition, contributors whose genomic data are (or are not) included in datasets have control over the use of their data. For example, contributors may restrict the types of studies or diseases/conditions for which their data can be used, or if their data are used at all.

There are challenges. The process includes an inefficient process of collecting and exchanging information about the rules associated with each entry into the dataset. It also requires the need for checking each data user's credentials to ensure that only people with appropriate researcher credentials are granted access.

On this morning, the DURI work stream was considering one part of this puzzle: creating a standard to streamline the credential check at the moment when a person attempts to access a dataset. What credentials should this person have to gain access? Can only bona fide researchers access the data or are there other types of experience that can qualify a person as a data user?

The work stream group consists of identity and access experts from around the world, a unique and finely refined group. They hail from technology companies, data repositories and globally-recognized research institutions. They understand the technology involved and its application to sensitive datasets like those in health care settings. They also understand how the technical components are very much an extension of prior work, knowledge obtained from a career-worth of education and experience. Because the group members have compatible and similar backgrounds, they don't have to start from basic principles to develop this standard. Instead they can rely on their shared experiences from doing similar work.

In this case study, we explore how trust is impacted by including a diverse group of people early in the process, and how practical feedback and iteration using work outputs greatly enhances adoption.

3.2.1 The Expert Conundrum

We need experts to create standards, policies, specifications, and research-based findings, but it is exactly this approach that can lead to a lack of trust among those who need to reference, use, or be governed by these outputs. These experts often have similar backgrounds and experiences, and share an understanding of the "prior art" on which they are building.

However, the community for which the experts are creating standards for almost always includes individuals that do not have this underlying understanding or experience. Without it, the work created by the experts can appear disjointed, illogical, and confusing, and these conditions can create a communication gap in explaining and understanding the standard. This gap can lead to mistrust.

3.2.2 Experts Doing Expert-y Things

The participants of the DURI work stream that Wednesday morning were talking about how information (credentials) about the person requesting access should be passed from system to system. Since the members of the group all came from software backgrounds, they were familiar with several protocols for keeping information secure while transferring it to other systems. They also understood existing standards for managing user sign-in to a particular system, and could take for granted that all involved could understand the contents of a written technical standard that described a technical method for ensuring that exchanged information hasn't been tampered with. A simple reference "shorthand" to this tacit knowledge is all that would be included to describe this implicit knowledge in the new standard being developed. There was no need to explicitly describe it because all involved in writing the standard would understand that these additional considerations were included in creating the standard; the "shorthand" (maybe a statement as simple as "using industry-accepted standards") would be enough to assure those with similar background that these items were considered and handled. But how is trust and understanding engendered for those without this background?

At times, the work done by a group of experts is so technically specific that the group doing the work can establish trust based on the "truths of the field." For example, trust in a computer network standard (and those who create it) may be established through an examination of the network's technical capabilities and the feasibility of equipment to accomplish the stated goal. In these situations, the group creating the technical specifications, defining standards, and building policies are subject matter experts in the topic, an elite group that tends to be somewhat homogeneous in background, training, values (at least in this topic area), and approach (which is often rooted in the discipline). This condition can lead to more efficient work, as the norms, approach, and "givens" for the area are often already negotiated and accepted.

But, those using or affected by the standard may be more skeptical. The "shorthand" that is so effectively used among the group of experts may be off-putting and "jargon-y." If trust has not already been established between stakeholders—data users, data contributors, and experts—, the data users and contributors can grow suspicious of whether the experts are acting in the end user's best interest (Petts, 2008).

3.2.3 The Impact on Trust

> "We need to be willing to risk embarrassment, ask silly questions, surround ourselves with people who don't know what we're talking about. We need to leave behind the safety of our expertise."—Jonah Lehrer, Imagine: How Creativity Works.

It is critically important that a technical standard be well understood and trusted by technical experts. These individuals have the background and expertise to be able to determine the quality and effectiveness of the standard. But too often technical standards teams nearly exclusively consider technical experts when gathering input for or communicating outputs about standards, often at the expense or exclusion of other stakeholders. This approach can have a very real impact on the trust that the other audiences put in the standard. This trust gap could ultimately impact the standard's adoption success. In addition, the absence of engagement could provide an information vacuum that may be filled by misinformation, further jeopardizing adoption.

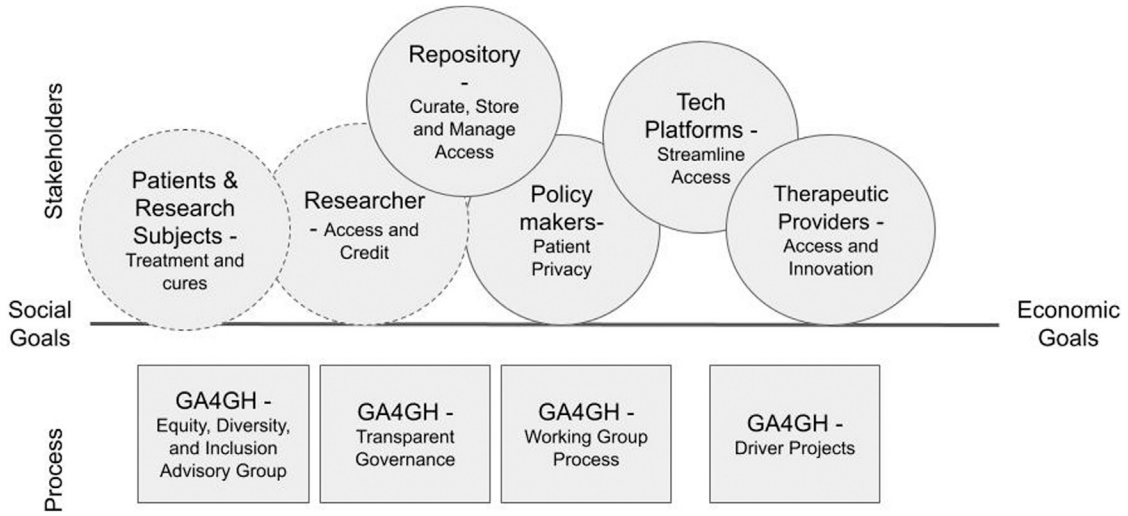


FIGURE 2 | GA4GH has been a multi stakeholder initiative from its founding. It has built a reputation for cross-stakeholder engagement through transparent inclusive processes and is able to attract a broad range of stakeholders to participate in standards development and implementation.

3.2.4 Who Is the Expert?

Who should we consider to be the "expert" in these types of efforts? You, someone who may be reading this article in a published journal, may consider the scholarly community to be the experts—those who have studied the field, researched the impacts, and analyzed the data. But, this lens may not be broad enough to quell doubt, engender trust, and produce acceptance (Noordegraaf, 2020).

In 2021, familiar objections to vaccines (History of Anti-vaccination Movements, 2018) resurfaced as governments and institutions prioritized vaccines as a way to return to pre-pandemic activities after COVID-19. The points of objection to the COVID vaccinations mirrored historic ones often rooted in fear, misunderstanding of the science or source and contents of vaccine raw materials, concerns about reduction of personal liberties, and suspicion of the intentions of those advocating for the vaccines. The net impact of these objections is a lack of trust in the stated vaccine purpose, efficacy, impact, and source. The vaccine expert might find these objections to be dismissable. After all, these concerns might be addressed by simply looking at the impact of past vaccines, scientific studies of efficacy of the current one being promoted, and understanding of how the vaccine was created and how it works. This expert may try to address objections through the lens of their expertise, dismissing objections as being misinformed, illogical, or unreasoned. But, could work have been done at earlier stages to engage these future skeptical audiences? How might things have been different if these individuals and their respective lenses were part of early discussions? Would this have led to different approaches or technical solutions? We could consider the future skeptics to be the experts of their own experience, and this expertise to be a worthy lens to incorporate earlier in the process to ensure trust later on.

GA4GH recognizes the importance of including diverse perspectives and lenses early in the standards-making process (**Figure 2**). In early 2020, the organization created an Equity, Diversity, and Inclusion Advisory Group for the express purpose to "find equitable and inclusive ways to bring diverse ideas into our standards creation process." (GA4GH OmicsXchange, 2020) This group engages "intentional community" principles (Vogl, 2016) to consider who should be included in the standard process and intentionally build a community that includes these parties with the goal of ensuring input and buy-in.

3.2.5 Including Diverse Stakeholders: The Expertise of End-Users

How can the trust built during the process of creating technical standards be transferred to the communities that will use the product but may not be privy to or understand the things unsaid—the norms, background, and implicit knowledge and understanding that something like a standard may contain?

Often with efforts that rely on a high degree of technical and subject matter expertise, the process of discovery and development goes something like this:

1. Identify and describe the problem
2. Develop hypotheses based on prior art, try out solutions, collaborate with other experts
3. Decide on a solution, often getting critiques and feedback from other experts
4. Disseminate more widely, sometimes creating descriptions or versions of the solution that are more accessible to other (non-expert) audiences

With this model, most end-users see only the packaged result. This is similar to back when regulated health and nutritional claims were not included on packaged products. (Institute of Medicine, 2010). Marketers might have provided terms like "healthy" or "good for you" as terms that someone who doesn't produce packaged food should understand. But do

these words help generate trust for the end-user who, say, is diabetic and needs to limit sugar intake, or has a food allergy? The end-user is an expert on how they will use the product and how their body might react to it, but words like "healthy" do not provide enough information for them to have trust that this product will not make them sick. The ingredients alone may not be enough. How it was prepared may be important for religious reasons (for example, keeping kosher), or for cultural reasons (for example, how its creation may have impacted the earth).

Food producers are getting better at engaging with end users to understand and include key factors in design, production, and marketing processes. It is no longer uncommon for packaged food labels to provide details about origin, method of creation, and ingredients, as well as values of those who are creating the product. The creation of things like technical standards, infrastructure, and research-based outcomes have not yet caught up. The omission of details for diverse audiences results from a lack of consideration for diverse stakeholder perspectives and needs. And these omissions impact trust.

GA4GH brings the end user into the standard process through Driver Projects. GA4GH Driver Projects are real-world genomic data initiatives that help guide development efforts and pilot the tools developed as part of the standards-making process. In addition, these projects enable stakeholders around the globe to advocate, mandate, implement and use GA4GH frameworks and standards in their local contexts, thereby building applicability and trust.

3.2.6 Learnings

The Wednesday meeting of the GA4GH Data Use and Researcher Identity (DURI) work stream meeting is coming to a close. Trust has been built into their process in two key ways:

1. Inclusion of diverse voices early in the process: Through the inclusive community programs advocated for by GA4GH's Equity, Diversity and Inclusion Advisory Group, early discussions include diverse voices. This practice enables a broad set of factors to be included for consideration, including religion, culture, relationship to the earth, socioeconomic status, logistics, and many others.
2. Encouraging practical use of standards as feedback input: Mechanisms such as the Driver Projects have been put in place to ensure that the solutions and their related benefits can be described using a broad set of lenses.

Including many broad perspectives from the beginning may slow the standards development process, but it will help pave the way for greater trust, use, and adoption of the standards that are developed.

3.3 Case Study: ORCID

Like Seamless Access and GA4GH's DURI Project, ORCID started with a beastly technical problem, in this case, uniquely identifying each researcher in the world. ORCID took a researcher-centric approach to solving the problem, enshrining individual-level control and privacy into its foundational principles. It also built a multi stakeholder governance group and established bylaws before building any technology, establishing a reputation that was the foundation for creating legitimacy (Zeyen et al., 2016).

> "To build community requires vigilant awareness of the work we must continually do to undermine all the socialization that leads us to behave in ways that perpetuate domination."— bell hooks, Teaching Community: A Pedagogy of Hope.

With vision, principles, and governance established, ORCID could then transition to developing technology requirements in collaboration with its communities. And show by example, over and over again, that the organization adheres to its principles. This was a slower start than either SeamlessAccess or DURI, but it set the stage for cross-stakeholder communication at the outset, rather than trying to bring groups in later on. As ORCID developed its core technology, design decisions were driven by the fundamental ORCID tenets of community governance and researcher control. Articulated in ORCID's Trust Program, it was the iterative experiences between and among stakeholder groups, the ORCID team, and the technology through which legitimacy and then trust emerged. Starting with early adopters, then focusing on publishers, research institutions, funders, and researchers, ORCID has demonstrated that it listens to and respects its stakeholders, and hews to its principles while evolving its offering as it learns more about community needs through working groups, community meetings, and consortia partners (**Figure 3**). Over the 10 years since the ORCID Board was founded, ORCID launched its registry, generated a base of over 10 million users, and on-boarded over 1,000 members in countries around the world, no small feat for a non-profit start-up.

In this section, we explore how ORCID has applied its trust framework in its work with its implementing partners, and how it has evolved with growing adoption.

3.3.1 Launch Partners

As ORCID prepared to launch in 2012, we gathered a group of highly motivated partners to test our APIs and develop integrations that would be available when we launched the Registry (ORCID, 2020; Launch Part, 2020). We looked for recognizable platforms with broad researcher usage that were jazzed by ORCID's mission and could invest resources and turn around a project in a short period of time. We listened, we learned, we launched.

We took an intentionally iterative approach to technology. As a start-up, we knew we couldn't know everything at the beginning, and that we were likely to make mistakes. We needed the flexibility to re-group. We had regular meetings with our launch partners, and we also set up user forums to gather feedback, prioritize features, and address issues.

One of the early decisions we made was to err on the side of rapid integration and reduced user burden. An example of this in practice was the design decision to not require email verification during the ORCID registration process. Yes, I can see you wincing. It seemed like a good idea at the time because it brought more implementers to our door and streamlined the process of using ORCID for researchers in standard workflows. However, we found that when email verification was left out of

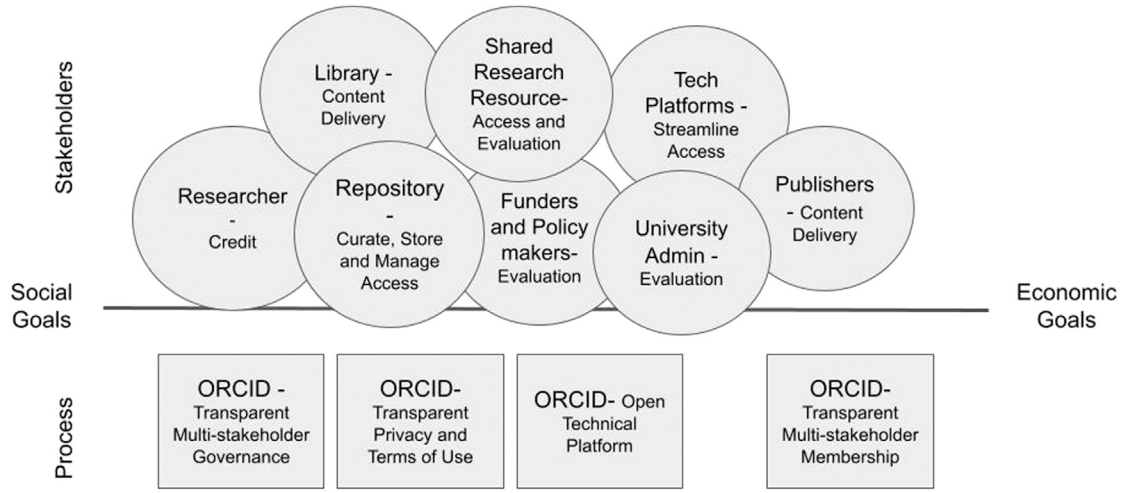


FIGURE 3 | Like GA4GH, ORCID has been a multi stakeholder initiative from its founding. It started with a practical solution, and built adoption by engaging iteratively with stakeholders, developing trust through transparent processes and demonstrating value and mutual social and economic benefit.

the registration process, a substantial proportion of account emails were never verified. This had the unintentional consequence of increasing researcher burden because we had to send out multiple messages to request email verification, it led to numerous help desk tickets filed by researchers requesting access to their accounts, and it hindered our ability to track active records, a key indicator of researcher adoption.

Over the next 4 years, researchers started to preferentially use systems that had integrated ORCID. Stakeholder sentiment helped to encourage implementers to add email verification into their workflows, assuaging fears that the extra steps would cause researchers to use other platforms. In turn, in 2017, ORCID was able to require email verification in all ORCID integrations, as well as require it for researchers to access basic ORCID account features (Demain, 2017). This example of stakeholder alignment and researcher engagement shows how increased trust in a research infrastructure can enable iterative improvements and broaden adoption.

3.3.2 API Versioning

Anyone involved in a technology start-up knows that driving early adoption is key. Without users, who really cares how cool the technology is. In the world of non-profits, the primary way to drive adoption is through mission alignment and mutual benefit. In our eagerness to onboard partners, we customized APIs to specific use cases. After 18 months, we were supporting over 20 API versions, which any developer will know is not sustainable. However, this approach at launch is not horrible, and is in fact quite common. Customization allowed us to work closely with our partners, figure out what worked well for users and what we could abstract across multiple platforms, and ultimately what features to fight for when developing the next API version. However, at some point, the customization needs to end and harmony must be established.

ORCID released its final mock API in March 2012 (ORCID, 2012). In addition to an ongoing API Users Group open to all with interest in the API, we formed a cross-stakeholder technical working group in February 2013 to examine the metadata used for Works in the ORCID Registry. The group helped us review the API and service models, and supported the iteration and socialization of a new API, version 1.2 (Paglione, 2013).

Over the next few years, we iterated on this API backbone model but knew we needed to make a break from initial assumptions to enable scalability. The ORCID record was not a monolithic document. We needed to enable calls and updates to individual sections and items. We launched version 2.0 in 2017, which helped manage hyper-authored publications, reduced confusion for implementers, and also added new functionality to support peer review recognition, improved user notifications, and the ability to support almost any persistent identifier (Peters, 2017).

We sunsetted version 1.2 at about the same time we started developing API v 3.0 (Teresa, 2018), and in 2020 transitioned to API 3.0 (Demain, 2020). Through this evolution, we worked closely with implementers to test new API functionality in early release candidates and also developed a policy for how API versioning would be handled so that enough time was given implementers to update and to ensure that priorities of implementers, members, and researchers were all considered in the versioning process (Blackburn, 2021).

API versioning is not unique to ORCID. However, how ORCID has handled versioning is an example of how to evolve technology while building trust across stakeholder groups. That ORCID was able to sustain, develop, and launch 3 substantially different APIs in its first 9 years is testament to the strength of the ORCID team and its commitment to serving ORCID communities.

3.3.3 Implementation Documentation

As the ORCID user base grew, more stakeholders and more platforms began to implement ORCID features. Here again,

ORCID took a broad approach, encouraging multiple approaches. We captured use cases specific to countries, community sectors, and workflows. The challenge came in sense-making for our implementers and users—and for our Help Desk. Just like the API, it is not sustainable to support custom documentation for every use case. We needed to consolidate messaging and documentation. It took several iterations to figure out which workflows worked best, decide what to prune, and then how to group use cases and information in a meaningful way.

Our Help Desk was launched in 2012, with live support, online documentation, and a User Forum with voting for new features. In 2014, we launched our first major documentation update, focused on implementers (Bryant, 2014). With more feedback from our growing member base, we created our Member Support Center in 2015, organizing technical documentation into sector-specific workflow guides, augmented by planning and communication resources (Paglione, 2015). We outgrew our user help desk system, and in 2018 with help from community translators transitioned to a new platform that enabled better local language support; along with this we also released more standardized help documentation and videos and were able to capture better statistics on how well we were serving our users (Cardoso, 2018).

The latest and for sure not the last documentation iteration was a massive upgrade to the ORCID Website in 2020, which updated how content was organized, based on usage patterns and community consultation. (Petro, 2019). In turn, these changes have both enabled and supported a pro-active product approach that more seamlessly integrates user feedback and key statistics to track ORCID adoption and impact. (Demain et al., 2021). Again, the iterative approach has allowed ORCID to engage stakeholders, test new approaches, integrate what works, and continue to innovate and build trusted services to meet the evolving needs of the ORCID community.

3.3.4 Certification of ORCID Service Providers

Throughout these product iterations, ORCID sought to ensure stakeholder alignment by keeping track of its members and implementers: who is using what API version for which use case(s). Similarly, ORCID continues working to engage researchers by streamlining the user experience and clarifying the benefits of using ORCID in research workflows.

With more platforms integrated, this community management work becomes more difficult, but also easier. Compared to when it launched in 2012, ORCID in 2021 is much more visible in the community. With this visibility comes opportunity.

We learned from our initial attempt at creating an implementer community, Collect and Connect, what worked and what did not (Mejias, 2018). We brought together our core principles of community governance and researcher control to develop an enticing change management program for implementers. We engaged our community of implementers in developing a certification program (ORCID Senior Team, 2020), something we had talked about at launch but decided was not the right time.

Now, within 1 year of its launch, the certification program has certified 15 platforms, with global reach (ORCID). One of the goals of this program is to drive a common researcher-centric user experience, another is to recognize certified implementers. The program meets ORCID principles, provides a mutual benefit to our stakeholders, and strengthens ORCID engagement with implementers, providing a dedicated channel for updates on new developments and listening to implementer experiences and feedback. And, the certification program increases trust among those using ORCID by extending the principles and values of ORCID to integrations that build on the ORCID platform.

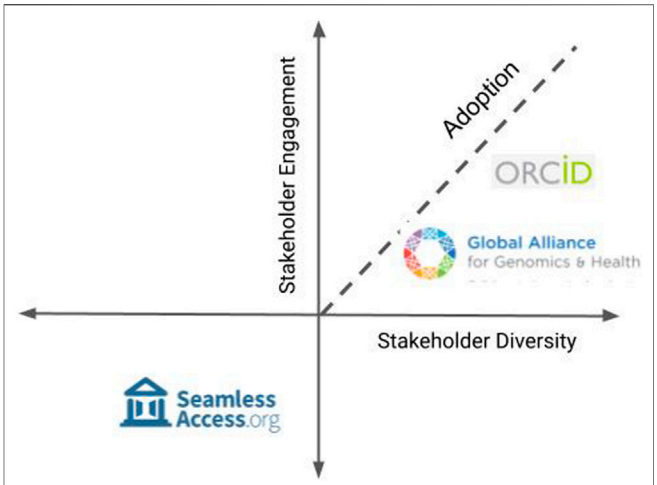


FIGURE 4 | Adoption is driven by engagement of all stakeholders. Successful multi stakeholder initiatives motivate community participation through collective social and economic incentives.

3.3.5 Learnings

Since 2012, ORCID has evolved from a nimble and high-energy organization to one that is established, sustainable, and influential. It has done this through enthusiastic stakeholder engagement and patient community development. By respecting researchers. By partnering with its stakeholders to try many approaches and learning from mistakes. By always centering on its core values and principles and keeping its activities mission focused, even when it makes some stakeholders feel their needs are not being met. This is what makes up ORCID's trust network.

4 CONCLUSION

Infrastructure is a critical component of research, whether it is manifested as technical standards, services, or community norms. Research infrastructure requires community trust for its adoption. As these case studies illustrate, we must take care to draw a wide circle when including stakeholders and interests in the design of infrastructure. We need to consider social and economic motivations and work to develop infrastructures are mutually beneficial. We also must ensure that there are transparent processes in place to support ongoing stakeholder engagement. Plotting the three infrastructures on axes of

stakeholder diversity and engagement, we see these two factors can predict community adoption (**Figure 4**).

We must find ways to listen outside of our expertise and comfort zones, and build open, ethical, and socially responsible infrastructure through iterative community consultation. The concept of "connected professionalism"—where expert groups are more porous and consider societal perspectives—is relevant here (Noordegraaf, 2015). Infrastructure principles must be transparent; this is the foundation for open governance (Wiedenhoft et al., 2017). We must also consider how the infrastructure will be supported over time so that it may be adopted, adapted, and accessed. And we need to ensure that infrastructures are designed so that researchers—communities of experts, contributors, and users—can use and benefit from them. This finding is in line with work on sustainability of "platform as commons" through participatory design, such as we see for open source software (Poderi, 2019).

A compelling theoretical framework for multi-stakeholder initiatives combines club theory with institutional theory, and posits company interest in joining an initiative is largely based on reputational risk and reward (Zeyen et al., 2016). However, this framework is largely based on economic incentives for participation. Open infrastructure initiatives have an equal or stronger social good component; how then to drive participation and adoption? Here, stakeholder theory provides a means for both justifying and assessing engagement across economic, social, and other factors (Jamali, 2008). Normative stakeholder theory is rooted in the view that customers and firms share an environment, and holds that all stakeholders are intrinsically valuable and deserve consideration, whether or not they have a direct economic stake. In this view, individual researchers are not just targets for marketing initiatives to grow market share, they become as important as firms for the views and experiences they bring to design and implementation. Similarly, game theory shows that reciprocity (adoption) is driven by trust, which is in turn dependent upon the beliefs stakeholders have about other initiative participants (Cox, 2004). How engagement is constructed is critically important; bringing together individual "experts" rather than representatives can lead to stronger trust, suggesting that working groups may be a key component of infrastructure trust (Song, 2009). Multi stakeholder initiatives thus provide the venue to drive institutional change and create mutual benefit through inclusive participation by a range of stakeholders.

The case studies we present align with and support this theoretical framework, showing that successful multi stakeholder initiatives (success measured in terms of infrastructure or standards adoption) engage diverse individual actors as well as institutions. We argue that it is in balancing social and economic incentives that initiatives can attract both the institutions that can effect structural change, and the people who can drive this change through their participation and advocacy.

Trust is not that easy, and once it's built, it's not guaranteed to continue. Authority to adopt and control rests in a community. This means that ongoing and contextually meaningful outreach and engagement has to happen for infrastructures to maintain trust and provide community benefit. Items may seem out of scope to one stakeholder group, but we must be prepared to listen and address issues across a range of diverse perspectives. Ongoing working groups and a transparent governance structure are necessary for initiative evolution and sustainability.

We find intriguing parallels with co-production and community welfare initiatives, where the concept of "care" is paramount. The difference between "caring about" and "caring for" can have deep implications for stakeholder support and infrastructure sustainability (Light and Seravalli, 2019). Those initiatives that are designed to engender reciprocal accountability and mutual commitment also encourage reflexive engagement among stakeholders.

Infrastructures that succeed do so because the communities they serve care deeply about their success. Care deeply enough to take the time to take part in developing standards, building practice communities, and, in so doing, build the interpersonal and inter-stakeholder trust needed to implement global research infrastructures that can support broad participation, adoption, and benefits for public welfare.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

REFERENCES

AlHogail, A. (2018). Improving IoT Technology Adoption Through Improving Consumer Trust. *Technologies* 6, 64. doi:10.3390/technologies6030064

Bilder, G., Lin, J., and Neylon, C. (2020). The Principles for Open Scholarly Infrastructure, retrieved [1 Sept 2021]. doi:10.24343/C34W2H

Blackburn, R. (2021). ORCID's API Version Policy. ORCID Blog, Available at: https://info.orcid.org/orcids-api-version-policy/(Accessed June 7, 2021).

Bryant, R. (2014). Resources to Support Integration and Outreach. ORCID Blog, Available at: https://info.orcid.org/resources-to-support-integration-and-outreach/(Accessed June 7, 2021).

Cardoso, A. (2018). Announcing Our New, Improved Support System! ORCID Blog, Available at: https://info.orcid.org/announcing-our-new-improved-support-system/(Accessed June 7, 2021).

Carpenter, T., and Horton, V. 2012). NISO and Collaboration: A Place at the Table for All Players. *Collaborative Librarianship* 4 (1), 31, 33. doi:10.29087/2012.4.1.03

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., et al. (2020). The CARE Principles for Indigenous Data Governance. *Data Sci. J.* 19 (43), 12pp. doi:10.5334/dsj-2020-043

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., and Simons, N. (2019). Bringing Citations and Usage Metrics Together to Make Data Count. *Data Sci. J.* 18 (1), 9. doi:10.5334/dsj-2019-009

Cox, J. C. (2004). How to Identify Trust and Reciprocity. *Games Econ. Behav.* 46, 260–281. doi:10.1016/S0899-8256(03)00119-2

Demain, P. (2017). New Feature Alert: Verify Your Email. ORCID Blog. Available at: https://info.orcid.org/new-feature-alert-verify-your-email/(Accessed June 7, 2021).

Demain, P. (2020). Sunsetting API Version 2. ORCID Blog, Available at: https://info.orcid.org/faq/sunsetting-api-version-2/(Accessed June 7, 2021).

Demain, P., Dineen, D., and Demeranville, T. (2021). Product: Our Progress to Date and Future Plans. ORCID Blog, Available at: https://info.orcid.org/2021-release-plan-updates/(Accessed June 7, 2021).

Dhanshyam, M., and Srivastava, S. K. (2021). Governance Structures for Public Infrastructure Projects: Public-Private Management Regimes, Contractual

Forms and Innovation. *Construction Manage. Econ.* 39 (8), 652–668. doi:10.1080/01446193.2021.1938162

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., et al. (2013). *Knowledge Infrastructures: Intellectual Frameworks and Research*. Michigan, US: University of Michigan School of Information. Challenges. Ann Arbor: Deep Blue, Available at: http://hdl.handle.net/2027.42/97552.

Eriksson, P., and Kovalainen, A. (2014). Ethnographic Research. In *Qualitative Methods in Business Research* (Washington DC: SAGE Publications Ltd), pp. 149–163. ISBN 978-1-4462-7338-8

Eriksson, P., Henttonen, E., and Meriläinen, S. (2011). Managing Client Contacts of Small KIBS Companies. *Int. J. Innovation Digital Economy (Ijide)* 2 (3), 1–10. doi:10.4018/jide.2011070101

GA4GH OmicsXchange. (2020) The Importance of Diverse Perspectives in Standards Development: An Interview with Laura Paglione [Audio podcast], Available at: https://www.ga4gh.org/news/omicsxchange-podcast-episode-9-the-importance-of-diverse-perspectives-in-standards-development-an-interview-with-laura-paglione/

Haak, L., GreeneS, S., and Ratan, K. (2020). A New Research Economy: Socio-Technical Framework to Open up Lines of Credit in the Academic Community. *Rio* 6, e60477. doi:10.3897/rio.6.e60477

See review in Haak, L. L., and Wagner, C. (2021). "Virtual Trust Networks in Transnational Research," in *The Future of International Exchanges in a Post-Pandemic World* (Washington DC: American Council on Education), 35–44. Available at: https://www.acenet.edu/Research-Insights/Pages/Internationalization/International-Exchanges-Beyond-COVID-19.aspx.

History of Anti-vaccination Movements (2018). The History of Vaccines: An Educational Resource by the College of Physicians of Philadelphia. Available at: https://www.historyofvaccines.org/content/articles/history-anti-vaccination-movements (Accessed on July 18, 2021).

Institute of Medicine (2010). *Committee on Examination of Front-Of-Package Nutrition Rating Systems and Symbols*, Editors EA Wartella, AH. Lichtenstein, and CS Boon (Washington DC: National Academies Press). Chapter 2, Available at: https://www.ncbi.nlm.nih.gov/books/NBK209859/(Accessed on Aug 20, 2021).

Jamali, D. (2008). A Stakeholder Approach to Corporate Social Responsibility: A Fresh Perspective into Theory and PracticeLaczniak GR, Murphy PE (2012). Stakeholder Theory and Marketing: Moving from a Firm-Centric to a Societal Perspective. *J. Bus Ethicsjournal Public Pol. Marketing* 8231 (2), 213284–223192. doi:10.1007/s10551-007-9572-410.1509/jppm.10.106

Light, A., and Seravalli, A. (2019). The Breakdown of the Municipality as Caring Platform: Lessons for Co-design and Co-learning in the Age of Platform Capitalism. *CoDesign* 15 (3), 192–211. doi:10.1080/15710882.2019.1631354

Mejias, G. (2018). Collect & Connect – Improved and Updated! ORCID Blog, Available at: https://info.orcid.org/collect-connect-improved-and-updated/ (Accessed June 7, 2021).

NISO (2019). *NISO RP-27-2019, Recommended Practices for Improved Access to Institutionally-Provided Information Resources: Results from the Resource Access in the 21st Century (RA21) Project*. Maryland, US, NISO. Available at: https://www.niso.org/publications/rp-27-2019-ra21.

Noordegraaf, M. (2015). Hybrid Professionalism and beyond: (New) Forms of Public Professionalism in Changing Organizational and Societal Contexts. *J. Professions Organ.* 2 (2), 187–206. doi:10.1093/jpo/jov002

Noordegraaf, M. (2020). Protective or Connective Professionalism? How Connected Professionals Can (Still) Act as Autonomous and Authoritative Experts. *J. Professions Organ.* 7 (2), 205–223. doi:10.1093/jpo/joaa011

ORCID (2020). Launch Partners Beta Group (Historical). *ORCID. Online Resource*. doi:10.23640/07243.12824030.v1

ORCID (2021). ORCID Certified Providers List. Available at: https://info.orcid.org/certified-service-providers/ (Accessed June 7, 2021).

ORCID Senior Team (2020). Announcing ORCID's New Service Provider Certification Program. ORCID Blog, Available at: https://info.orcid.org/announcing-orcids-new-service-provider-certification-program/ (Accessed June 7, 2021).

ORCID (2012). Version 1.0.4 of the ORCID Mock API Released. ORCID Blog. Available at: https://info.orcid.org/version-1-0-4-of-the-orcid-mock-api-released/(Accessed June 7, 2021).

Paglione, L. (2015). ORCID Member Support: Improving on Excellence. ORCID Blog, Available at: https://info.orcid.org/orcid-member-support-improving-on-excellence/(Accessed June 7, 2021).

Paglione, L. (2013). Works Metadata: Recommendations from Our Working Group. ORCID Blog, Available at: https://info.orcid.org/works-metadata-recommendations-from-our-working-group/(Accessed June 7, 2021).

Peters, R. (2017). All about Our New API: An Interview with Rob Peters, Director of Technology. ORCID Blog, Available at: https://info.orcid.org/all-about-our-new-api-an-interview-with-rob-peters-director-of-technology/(Accessed June 7, 2021).

Petro, J. (2019). Time for a Website Refresh! ORCID Blog, Available at: https://info.orcid.org/time-for-a-website-refresh/(Accessed June 7, 2021).

Petts, J. (2008). Public Engagement to Build Trust: False Hopes? *J. Risk Res.* 11 (6), 821–835. doi:10.1080/13669870701715592

Poderi, G. (2019). Sustaining Platforms as Commons: Perspectives on Participation, Infrastructure, and Governance. *CoDesign* 15 (3), 243–255. doi:10.1080/15710882.2019.1631351

Skinner, K. (2019). Why Are So Many Scholarly Communication Infrastructure Providers Running a Red Queen's Race? Available at: https://educopia.org/red-queens-race/.

Song, F. (2009). Intergroup Trust and Reciprocity in Strategic Interactions: Effects of Group Decision-Making Mechanisms. *Organizational Behav. Hum. Decis. Process.* 108, 164–173. doi:10.1016/j.obhdp.2008.06.005

Teresa, A. (2018). Sunset Date Set: Upgrade to ORCID API 2.0+ by August. ORCID Blog, Available at: https://info.orcid.org/sunset-date-set-upgrade-to-orcid-api-2-0-by-august/(Accessed *June 7*, *2021*).

Vogl, C. H. (2016). *The Art of Community: Seven Principles for Belonging*, Oakland, CA, Berrett Koehler Publishers, 978-1-62656-8419.

Wiedenhoft, G. C., Luciano, E. M., Luciano, E. M., and Magnagnagno, O. A. (2017). Information Technology Governance in Public Organizations: Identifying Mechanisms that Meet its Goals while Respecting Principles. *Jistem* 14 (1), 69–87. doi:10.4301/S1807-17752017000100004

See review in Wiedenhöft, G. C., Luciano, E. M., and Pereira, G. V. (2020). Information Technology Governance Institutionalization and the Behavior of Individuals in the Context of Public Organizations. *Inf. Syst. Front.* 22 (6), 1487–1504. doi:10.1007/s10796-019-09945-7

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* 3, 160018. doi:10.1038/sdata.2016.18

Zeyen, A., Beckmann, M., and Wolters, S. (2016). Actor and Institutional Dynamics in the Development of Multi-Stakeholder Initiatives. *J. Bus Ethics* 135, 341–360. doi:10.1007/s10551-014-2468-1

Conflict of Interest: Author HF is affiliated with Spherical Cow Consulting. Author LH is affiliated with Mighty Red Barn. Author LP is affiliated with Laura Paglione, LLC and Spherical Cow Group.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Research Metrics and Analytics

PERSPECTIVE
published: 23 December 2021
doi: 10.3389/frma.2021.748095

Perspectives on Open Science and The Future of Scholarly Communication: Internet Trackers and Algorithmic Persuasion

Tiberius Ignat[1], Paul Ayris[2], Beatrice Gini[3], Olga Stepankova[4], Deniz Özdemir[4], Damla Bal[1] and Yordanka Deyanova[1]*

[1]Scientific Knowledge Services, Munich, Germany, [2]LCCOS - Library, Culture, Collections, Open Science, University College London, London, United Kingdom, [3]Cambridge University Library (CUL), University of Cambridge, Cambridge, United Kingdom, [4]CIIRC (Czech Institute of Informatics and Robotics and Cybernetics), BEAT (Biomedical Engineering and Assited Technologies) Department, Czech Technical University in Prague, Prague, Czechia

The current digital content industry is heavily oriented towards building platforms that track users' behaviour and seek to convince them to stay longer and come back sooner onto the platform. Similarly, authors are incentivised to publish more and to become champions of dissemination. Arguably, these incentive systems are built around public reputation supported by a system of metrics, hard to be assessed. Generally, the digital content industry is permeable to non-human contributors (algorithms that are able to generate content and reactions), anonymity and identity fraud. It is pertinent to present a perspective paper about early signs of track and persuasion in scholarly communication. Building our views, we have run a pilot study to determine the opportunity for conducting research about the use of "track and persuade" technologies in scholarly communication. We collected observations on a sample of 148 relevant websites and we interviewed 15 that are experts related to the field. Through this work, we tried to identify 1) the essential questions that could inspire proper research, 2) good practices to be recommended for future research, and 3) whether citizen science is a suitable approach to further research in this field. The findings could contribute to determining a broader solution for building trust and infrastructure in scholarly communication. The principles of Open Science will be used as a framework to see if they offer insights into this work going forward.

Keywords: scholarly communication, track, persuade, readers, authors, open science, trust, infrastructure

OPEN ACCESS

Edited by:
Daniel W. Hook,
Digital Science, United Kingdom

Reviewed by:
Thanasis Vergoulis,
Athena Research Center, Greece
Diana Marsh,
University of Maryland, United States

****Correspondence:***
Tiberius Ignat
tiberius@
scientificknowledgeservices.com

Specialty section:
This article was submitted to Scholarly Communication, a section of the journal Frontiers in Research Metrics and Analytics

Received: *27 July 2021*
Accepted: *26 November 2021*
Published: *23 December 2021*

Citation:
Ignat T, Ayris P, Gini B, Stepankova O, Özdemir D, Bal D and Deyanova Y (2021) Perspectives on Open Science and The Future of Scholarly Communication: Internet Trackers and Algorithmic Persuasion. Front. Res. Metr. Anal. 6:748095. doi: 10.3389/frma.2021.748095

INTRODUCTION

Open Science is part of the "new normal" as the world emerges from the covid-19 pandemic. Open Access to publications is now a well-developed phenomenon for research outputs.

In Europe, there are eight themes which are commonly seen to be part of Open Science principle and practice, including *Research Integrity* and *The Future of Scholarly Communication*, both being the subject of our perspective paper.

These are: 1) Rewards and Incentives, 2) Indicators and Next-Generation Metrics, 3) Future of Scholarly Communication, 4) European Open Science Cloud (EOSC), 5) FAIR data, 6) Research Integrity, 7) Skills and Education, 8) Citizen Science (Open Science EU, 2020).

Research Integrity comprises a set of principles which should underpin research practice. As the 23 research-intensive universities of LERU concluded in their report Open Science and its role in universities: a roadmap for cultural change (Ayris et al., 2018a), a move to Open Science represents a fundamental cultural shift for researchers. The ALLEA code on Research Integrity states that good research practices are based on fundamental principles of research integrity, these being: Reliability, Honesty, Respect, Accountability (ALLEA, 2017a).

ALLEA (ALLEA, 2017b) has produced the European Code of Conduct for Research Integrity that addresses challenges emanating from technological developments and social media, among other areas. For example, it says that "Researchers, research institutions and organisations [should] provide transparency about how to access or make use of their data and research materials." As such, it is recognised by the European Commission as the reference document for research integrity for all EU-funded research projects and as a model for organisations and researchers across Europe.

Web trackers enable profitable business models for organisations that develop web-based applications, especially for those that interfere with people's behaviour. In some cases, governmental agencies use such models, too. Some tech companies consider these trackers fundamental for "the free and open" Internet as we know it (BBC News, 2021). We disagree with this model for developing the Internet and its role in society. Furthermore, we consider this an inappropriate model for the field of scholarly communication.

While allowing ourselves to be surveilled by unknown organisations in exchange for free or underpriced services (Barbu, 2014), we develop a new culture in which our society is trading hard-won freedom for questionable prosperity. That culture will be inherited by future generations, who will be challenged to change it when this trade-off will no longer be bearable.

This paper presents a set of recommendations and the authors' perspective on using modern technologies in scholarly communication processes. To build our views, we studied 148 web pages related to the field and we collected 15 expert opinions.

OBSERVATIONS AND DISCUSSION

Modern technologies based on tracking (in Internet and mobile applications), including Artificial Intelligence (AI), digital persuasive technologies and—to an extent—Robotic Process Automation (RPA), are common elements in the new landscape of content creation, content management and information. Scientific knowledge and scholarly communication could become the new territory to be infested by these tracking-related technologies.

While some trackers are less invasive and are placed to support basic functionalities for websites and apps, most trackers are used to expose our behaviour and personal data, for the benefit of a small group of organisations. They are used in prediction models that fuel the business of recommendation engines (Beckers, 2021). They contribute to a surveillance economy and are used to create individual psychographic profiles (Gibney, 2018).

Both desktop and mobile versions of web tracking are implemented by utilizing a plethora of tracking technologies, including cookies, JavaScript components, local shared objects, iframes, and relying on the technology of third-party trackers (Mittal, 2010). The most common way to prevent cookie tracing is to configure the internet browser configuration in order to block third-party cookies. Browser extensions on the other hand could be of assistance in this case, and Incognito mode (which is also referred to as private browsing) can additionally offer protection as well, though not disabling third-party cookies completely (Bielova, 2017). Consequently, a privacy scoring model for each website to evaluate the privacy risks could give detailed insights for detection (Hamed and Ayed, 2015).

The future of tracking shall be evaluated in accordance with the new Internet protocols, passive network traffic monitoring and Developers' technical blogs since a variety of information can be gathered by the analysis of new protocols and extensions covering different web standards and their functionalities respectively (Bujlow et al., 2017). Forrester's data security and privacy playbook provides the tools, information and analysis to aid with the protection of data privacy abuse with a framework that has a three-step process (Balaouras, 2019): ensuring the necessities for better data security and privacy, implementing a road map to brace the business and enhance data security and privacy and carrying out security and privacy solutions, thus affirming the execution of the privacy of data (Abdullah, 2020).

To investigate the frequency of tracking in scholarly communication, we analysed 148 web pages related to scholarly communication. They represent a mix of publishers (55), technology companies (35), preprint servers (27), content aggregators (24), libraries and others (7). They answered 9 questions (**Figure 1**).

Quantitative Observations

The answer to the first question "Does the website inform you whether it uses cookies?" divides the original dataset into 2 disjunctive subsets—the **CC** dataset with web pages openly confirming use of cookies (n = 94) and the **NC** dataset with web pages that don't mention cookies (n = 54). To verify the use of cookies in the **NC** dataset, we checked it with cookie management applications, revealing cookies' presence in most of them. This suggests needed improvement.

Questions 2–8 are not relevant for websites in the **NC** dataset, hence the detailed analysis is focussed to the **CC** dataset containing 94 web pages.

The results highlight surprising observations:

- 60% of webpages in **CC** subset offer no option to manage cookies (Q3). Even if this option is presented, the "Accept All" button is promoted more visibly often (37,2%, see Q2).
- Questions 2 and 4–8 are answered "I can't determine" in more than 50% of cases – suggesting that managing cookies is far from intuitive.

While of the original set of 148 webpages, 68,9% offer the possibility to create user accounts, this percentage is 71,3% for the **CC** dataset. In psychographic profiling, data collected through

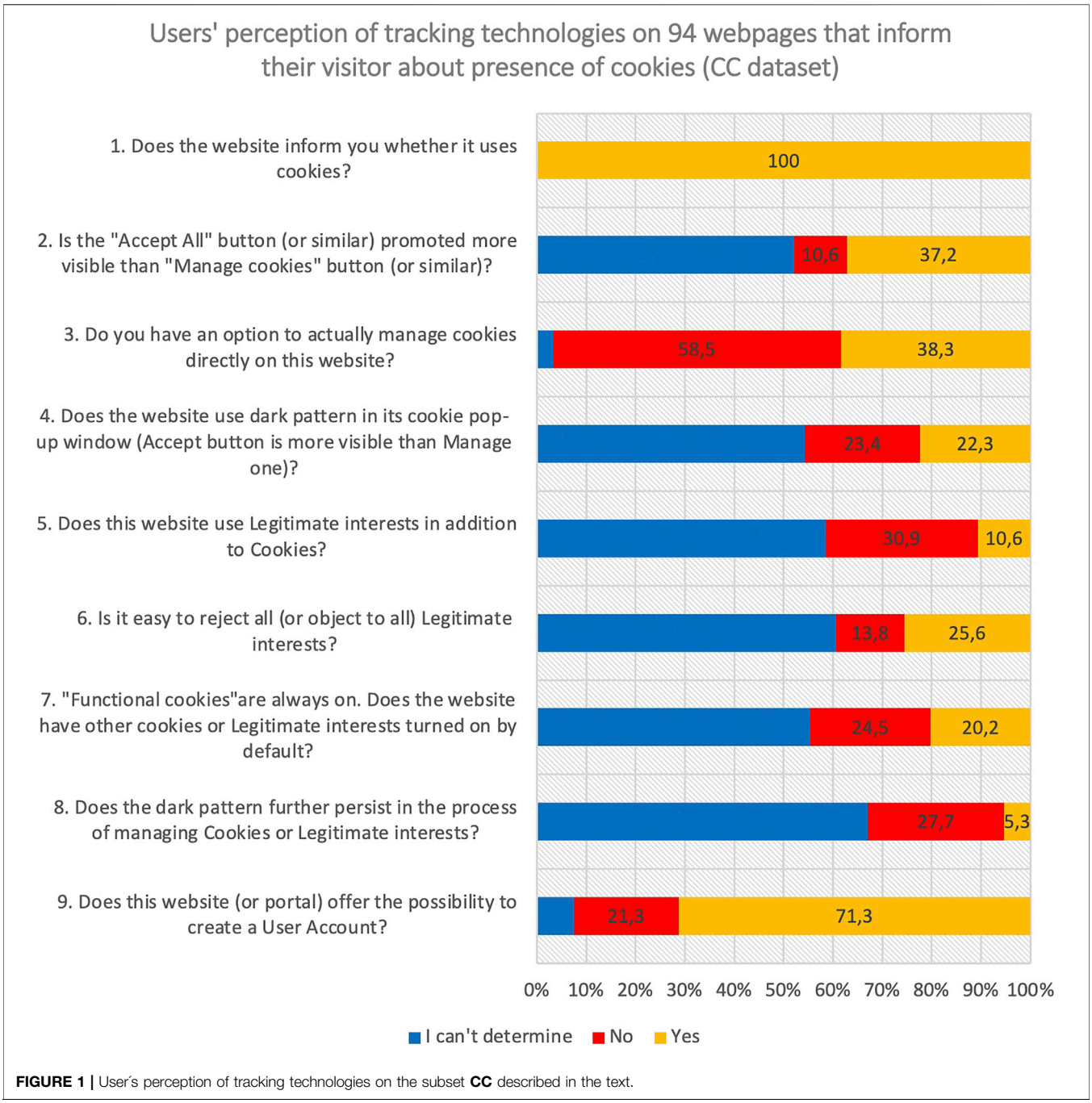


FIGURE 1 | User´s perception of tracking technologies on the subset **CC** described in the text.

user accounts is usually complementary to the data collected through trackers (Aries Systems, 2020; art. 2 and art. 7), with potential for the de-anonymization of the datasets.

The dataset could be downloaded from here: https://doi.org/10.5281/zenodo.5139523.

Expert Interviews

To understand subjective experiences of trackers in scholarly communication, we conducted written interviews with fifteen experts in the fields of scholarly communication. They were selected based on the authors' professional networks. While this cannot be considered a representative sample, it can provide an initial insight into the community's perceptions of these issues. A summary of their answers to 13 questions is presented in **Figure 2** with both an overview of responses and selected quotations.

The overview bar chart interprets the answers using 3 values: NO, YES or "Ambiguous" corresponding to "I do not know" (questions 2,3 6 and 9), "Rarely" (questions 4 and 5), no answer (questions 7 and 8) or "I am not sure—I do not know enough about the topic to give a clear answer" (questions 10–13). This rough simplification of answers is used here to highlight the

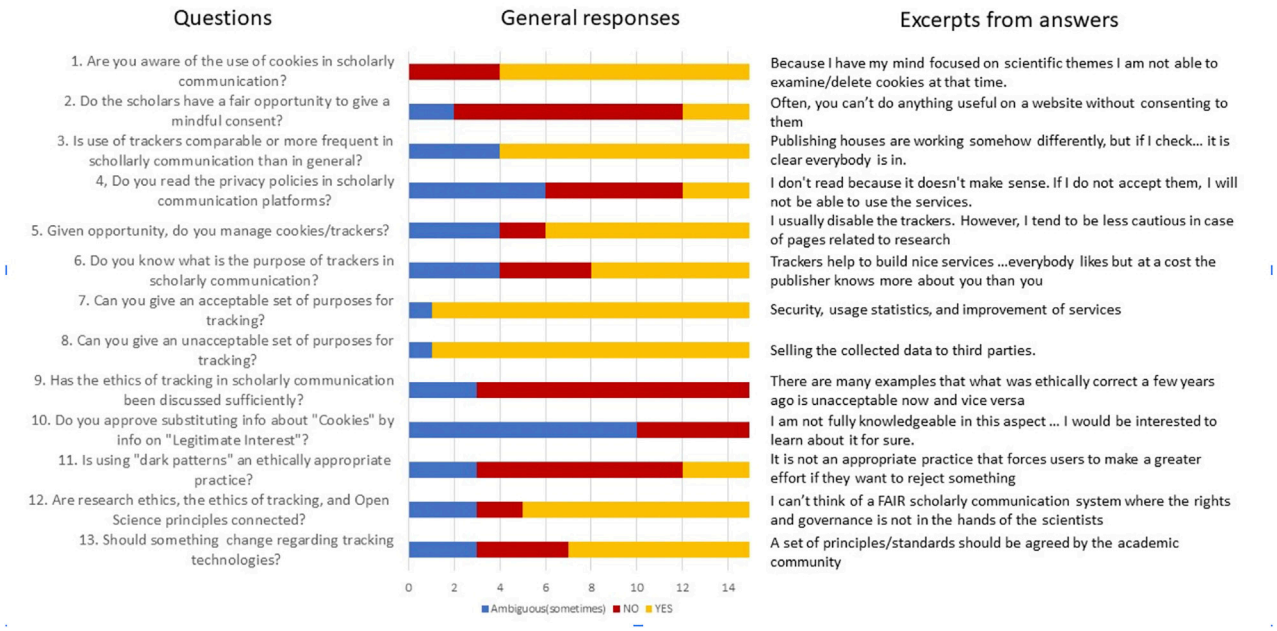


FIGURE 2 | The result of interviewing 15 experts of scholarly communication.

extent of interest in the topic among the academic community. The graph confirms that scholars believe they do not have fair opportunities to give mindful consent for tracking (question 2) and that the ethics of tracking should be discussed more extensively in the academic community (question 9). Moreover, differing opinions about acceptable and unacceptable purposes of the technology suggest that fruitful debates could be organised if the right forums were created.

Essential Questions for the Scholarly Communication Community

Scholarly communication community might be tempted to use trackers and persuasive technologies. Some might serve the interest of readers, authors, peer-reviewers, and research organisations.

Hence, it is important to identify what questions need answers before these technologies become the norm.

We believe these questions are essential for the members of scholarly communication community when considering to use modern technologies like web tracking, AI and RPA:

1. What are the highest ethical paths for a field of communication that needs to build trust and communicate evidence and knowledge?
2. What vulnerabilities are brought to the research community?
3. What are the real opportunities for researchers and for society?
4. What is realistic and what is utopic in these technologies? Which are the demonstrated positive effects of those technologies?
5. How can we ensure that those technologies develop human-centric?
6. Who is governing the development of those technologies?
7. What system could guard the researchers from being manipulated by such technologies (Michael, 2019)?
8. What is the impact of such technologies on educating next generations of curious minds?

AUTHORS' PERSPECTIVE

Our analysis showed that only 64% of websites inform users of their use of cookies, despite this being a legal requirement in the EU, where we accessed them. Even worse, cookies appear in most of 54 websites from the NC dataset consisting of those websites that don't provide the visitor with any information about their cookies policy. The option to manage cookies was either lacking or disguised with "dark patterns" in the majority of sites, contrary to our expectations for transparency and freedom in internet use. Moreover, 69% of all studied websites offered the option to create an account, even though the benefits to users were not always evident. User accounts can store large amounts of information and could be combined with cookie data to track and manipulate behaviour. This paints a troubling picture of the state of tracking in scholarly communication: there is little transparency and significant potential for persuasive technologies to become commonplace.

The experts' interviews corroborated this lack of transparency: most interviewees assumed that large amounts of data were being collected, but admitted to having a poor understanding of what

the process and aims were. They also indicated that, although the option to manage cookies exists in principle, in reality most cookies are accepted unquestioningly due to difficulties and time required to manage them manually. Most concerning was the fact that several interviewees instinctively trust scholarly communication platforms, saying for instance: "I usually disable the trackers. However, I tend to be less cautious in the case of pages related to research as I hope there is a smaller risk of misuse of this data. Of course, I have no hard data supporting this assumption." Thus scholarly communication platforms may be benefitting from a greater degree of trust from their users, but not setting higher standards for themselves, compared to other websites.

Interviewees identified some beneficial uses of tracking, namely personalised recommendations for reading materials, conferences and job opportunities, and the collection of anonymised data to improve website design or report usage statistics. On the other hand, the selling of personal data was overwhelmingly cited as an unacceptable use of tracking. Other unacceptable uses included the profiling of users based on protected characteristics such as ethnicity or political affiliation, advertising (although not unanimously) and the concentration of market power in the hands of a few platforms. Lastly, interviewees agreed that there is an urgent need for dialogue across the scholarly communication community to agree standards of behaviour in this area.

The 2017 ALLEA code says "Authors [should] ensure that their work is made available to colleagues in a timely, open, transparent, and accurate manner ... and are honest in their communication to the general public and in traditional and social media." The problem, however, is that this is an instruction to the author and not to the publisher or any third party host/disseminator of the work. In the section on "Research Misconduct or other Unacceptable Practices," the code identifies as bad practice "Establishing or supporting journals that undermine the quality control of research." However, it defines the scope of this bad practice as simply "predatory journals."

The ALLEA code certainly attempts to bring within scope many areas of Open Science, but treats these subjects as issues pertaining to the author(s). This is an omission and, as this article has identified, a dangerous one if many users implicitly trust scholarly communication platforms. Standards which are expected of researcher(s) therefore do not explicitly cover publishers, hosters and disseminators of that research in the principal European code for research integrity.

Scholarly communication is an essential element of research: it supports rigorous professional conversation between researchers, with independent, critical thought at its core. Tracking the researchers' interactions and persuading them to take certain actions will significantly diminish their genuine contribution to society. Research needs intuition, anticipation, hard work and designed serendipity. Being able to influence these elements, in both a transparent or covert manner, has the potential to control even further the course of human progress (in addition to the funding mechanisms). We need to avoid the unquestioning legitimisation of libertarian paternalism in scholarly communication (Thaler and Cass, 2003).

First of all, tracking and persuasive technologies could change the readership of a journal in a manner completely different than traditional editorial practices. Academic texts without proper editorial work could thrive based on the application of such technologies, instead of the quality of their conversation. Second, surveillance technologies used to build psychographic profiles, persuade algorithmically and pass as humans, pose the potential risk of influencing authors' contributions, including research conclusions and recommendations. Even hypothesis generation could be influenced by the aforementioned technologies: for years there has been a quest to automate the identification of "hot" topics. This approach didn't prove beneficial to research diversity or contribute to the development of generations of curious minds. Using AI and RPA for hypothesis refinement may represent an effective and efficient solution for researchers (The Royal Society and Alan Turing Institute, 2019), but not before defining what represents an ethical use of these technologies. Such systems "provide predictions, but no real insight. The "deep" learners are shallow indeed" (Carey, 2020).

Those we interviewed would welcome more evidence about tracking and persuasive technologies in scholarly communication. To produce such evidence, proper, well-resourced research is needed. This research needs to identify the actual use of those technologies, anticipate their potential use, but also determine which are the best approaches to engage with scholarly communication stakeholders in order to build a safe roadmap for the future. Early engagement is essential for steering a community in a smooth manner towards ethical developments.

The low number of expert opinions and the answers we received is another reasonable indication that we are acting at a frontier of human knowledge. These technologies are largely unknown and it is hard to determine how much priority they deserve.

We believe that in-depth research in this area would support practical approaches for Open Science. Such new understanding is key for at least two pillars of the new research culture: *The Future of Scholarly Communication* and *Research Integrity* (Lawrence and Mendez, 2020).

We believe that this is the best time to research the use of algorithmic technologies and their particular impact in scholarly communication. Furthermore, an advocacy and engagement programme is needed to connect stakeholders and agree on paths forward. The solution will be less about mandates; it will be about creating trust, encouraging transparency and building consensus.

RECOMMENDATIONS FOR THE SCHOLARLY COMMUNICATION COMMUNITY

Both open science and scholarly communication communities need to widen their remit to include guidance and best practice on the use of tracking and persuading technologies. Research integrity codes such as the ALLEA code need significant revision to embrace these new areas. As the LERU Open Science Roadmap makes clear: "To

embrace Open Science, universities and researchers need to embrace cultural change in the way they work, plan and operate. The result will infuse a culture of Open Science throughout the academic organisation and may support other evolutions in academic practice." (Ayris et al., 2018b). The scope of such change needs to be as wide as possible, covering all players in the scholarly communications landscape.

Researchers need to be aware of the dangers associated with cookies. In this article, some of those questioned appreciated the benefits of tracking technologies. However, the findings of the quantitative and qualitative studies paint a concerning picture. There is little transparency and a significant potential for persuasive technologies to become commonplace. There is a need for education to enable researchers to understand the results of using dissemination and syndication platforms (including social media). Research funders, universities, publishers and tech companies should consider co-creating ethical requirements for such platforms. There also needs to be a global advocacy and awareness campaign to open up the issues around the use of cookies and trackers, highlighting the dangers as well as the benefits. This will help re-shape research culture at both national and international levels.

Open Science has also led to the unprecedented sharing of research data. While generally a positive change, this opens opportunities for the detrimental use of technology. An example is using data from a research project on human fears to train an algorithm that persuades people to buy insurance policies. For researchers and research organisations, including those that curate and maintain research datasets, it is important to be very conscious about what license should be granted to research data sets. Open Data is circulated in parallel and sometimes, instead of FAIR Data. These two concepts must not the confused with each other. While broader access and easier scrutiny to research data are necessary, the existence of malicious intent should be recognised and further development of creative commons models should be undertaken.

Our research data collection protocol was designed to use citizen scientists (volunteers) alongside researchers' efforts. We also created short training materials to improve data collection, as the international community recommends. To attain the scale, diversity and geographical penetration of a full study, we think citizen science is a suitable approach for future work in this area as similar models exist (CSI-COP, 2021).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Materials**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

TI contributed to the conception and the design of the study, the introduction, designing and testing of the data collection protocol and qualitative interviews, the drawing up of essential questions and highlighting the risks of using the new technologies, and underlining the importance of further research. PA contributed to the introduction, the comparison of findings and the recommendations to the communities interested in scholarly communication. BG contributed to data collection and to helping formulate our group perspective on our findings and qualitative interviews. OS organised the dataset, performed the data analysis and contributed to qualitative interviews. DÖ–explained how to protect the research data. DB and YD contributed to data collection, qualitative interviews and the article bibliography. All authors contributed to manuscript revision, read, and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article (including Further Recommended Readings) can be found online at: https://www.frontiersin.org/articles/10.3389/frma.2021.748095/full#supplementary-material

Data Sheet 1 | Further Recommended Readings.

REFERENCES

Abdullah, H. (2020). "Proposition of a Framework for Consumer Information Privacy Protection," in International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD). doi:10.1109/icabcd49160.2020.9183822

ALLEA (2017a). *The European Code of Conduct for Research Integrity*. Berlin: ALLEA. Revised edition https://www.allea.org/wp-content/uploads/2017/05/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017.pdf.

ALLEA (2017b). *The European Code of Conduct for Research Integrity*. Berlin: ALLEA. Available at: https://allea.org/code-of-conduct/ (Accessed July 20, 2021).

Aries Systems (2020). Privacy Policy. Available at: https://www.ariessys.com/about/privacy-policy/ (Accessed July 12, 2021).

Ayris, P., San Román, A. L., Maes, K., and Labastida, I. (2018a). *Open Science and its Role in Universities: A Roadmap for Cultural Change*. Leuven: LERU. Available at: https://www.leru.org/publications/open-science-and-its-role-in-universities-a-roadmap-for-cultural-change.

Ayris, P., San Román, A. L., Maes, K., and Labastida, I. (2018b). *Open Science and its Role in Universities: A Roadmap for Cultural Change*. Leuven: LERU. Full Paper. Leru:Org: https://www.leru.org/files/LERU-AP24-Open-Science-full-paper.pdf.

Balaouras, S. (2019). *Protect Your Intellectual Property and Customer Data from Theft and Abuse*. Cambridge: USA.

Barbu, O. (2014). Advertising, Microtargeting and Social Media. *Proced. - Soc. Behav. Sci.* 163, 44–49. doi:10.1016/j.sbspro.2014.12.284

BBC News (2021). Google Boss Sundar Pichai Warns of Threats to Free and Open Internet - BBC News. Youtube: https://www.youtube.com/watch?v=eHMViKOOfg8 (Accessed July 20, 2021).

Beckers, M. (2021). Modern Recommender Systems. Available at: https://towardsdatascience.com/modern-recommender-systems-a0c727609aa8 (Accessed July 12, 2021).

Bielova, N. (2017). "Web Tracking Technologies and Protection Mechanisms" in Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. doi:10.1145/3133956.3136067

Bujlow, T., Carela-Espanol, V., Lee, B.-R., and Barlet-Ros, P. (2017). A Survey on Web Tracking: Mechanisms, Implications, and Defenses. *Proc. IEEE* 105, 1476–1510. doi:10.1109/jproc.2016.2637878

Carey, B. (2020). Need a Hypothesis? This A.I. Has One. Available at: https://www.nytimes.com/2020/11/24/science/artificial-intelligence-ai-psychology.html (Accessed July 15, 2021).

CSI-COP (2021). Citizen Scientists Investigating Cookies and App GDPR Compliance - about Us. Available at: https://csi-cop.eu/ (Accessed July 23, 2021).

Gibney, E. (2018). The Scant Science behind Cambridge Analytica's Controversial Marketing Techniques. *Nature*. doi:10.1038/d41586-018-03880-4

Hamed, A., and Ayed, H. K.-B. (2015). "Privacy Scoring and Users' Awareness for Web Tracking," in 6th International Conference on Information and Communication Systems (ICICS). doi:10.1109/iacs.2015.7103210

Lawrence, R., and Mendez, E. (2020). *Progress on Open Science : Towards A Shared Research Knowledge System : Final Report of the Open Science Policy Platform*. Brussels: Op.Europa.Eu. Available at: https://op.europa.eu/en/publication-detail/-/publication/d36f8071-99bd-11ea-aac4-01aa75ed71a1.

Michael, A. (2019). Ask the Chefs: AI and Scholarly Communications. Available at: https://scholarlykitchen.sspnet.org/2019/04/25/ask-chefs-ai-scholarly-communications/ (Accessed July 12, 2021).

Mittal, S. (2010). User Privacy and the Evolution of Third-Party Tracking Mechanisms on the World Wide Web. *SSRN J.* doi:10.2139/ssrn.2005252

Open Science EU (2020). Open Science Policy Platform: Final Report. Available at: https://openscience.eu/open-science-policy-platform-final-report/ (Accessed October 10, 2021).

Thaler, R. H., and Sunstein, C. R. (2003). Libertarian Paternalism. *Am. Econ. Rev.* 93, 175–179. doi:10.1257/000282803321947001

The Royal Society and The Alan Turing Institute (2019). The AI Revolution in Scientific Research. London: The Royal Society. Available at: https://royalsociety.org/-/media/policy/projects/ai-and-society/AI-revolution-in-science.pdf.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Research Metrics and Analytics

OPINION
published: 03 January 2022
doi: 10.3389/frma.2021.767869

Openness, Integrity, Inclusion, and Innovation in Scholarly Communication: Competing or Complementary Forces?

Virginia Barbour[*†]

Office for Scholarly Communication, Queensland University of Technology (QUT), Brisbane, QLD, Australia

Keywords: open access, integrity, infrastructure, COVID-19, preprint, Open Science, innovation

INTRODUCTION

In 2020, the importance of open and rapid communication of academic research came to the fore, as possibly never before, in the global effort to address the COVID-19 pandemic. The pandemic arrived at a time when much of the infrastructure for sharing research openly and rapidly was already in place, and to a large extent, the global publishing enterprise was able to fulfill its function of dissemination of information.

However, we are already seeing signs that publishing may revert to a more closed model post pandemic. It is also clear that the pandemic has exacerbated some of the problems in scholarly communication, such as a worsening participation by women and unequal distribution of funding globally. Furthermore, it is not clear that some of the innovations developed in the pandemic for sharing of information—such as the CORD-19 dataset of publications—will endure in their current state. Finally, the sheer volume of publishing, especially through relatively novel mechanisms, such as preprints, has led to uncertainty about how to support trust in research publications, both in the academic community and in the wider public.

INFRASTRUCTURE AND IDEAS READY FOR A PANDEMIC

The COVID-19 pandemic that emerged in 2020 and which at the time of writing is still ongoing led to probably the biggest disruption in scholarly communication seen since academic publishing began to move online at the end of the 20th century. What has been critical to the success thus far of much of this disruption is that it builds on emerging infrastructure and ideas that have primed the publishing system for change.

There are previous examples of publishing having to respond to a global medical emergency; the most recent relevant of these is SARS in 2003 (SARS | Basics Factsheet | CDC, n.d). That emergency was fortunately relatively short lived, and although the global medical research community rose to the challenge of investigating SARS, the global publishing community barely coped. A 2010 analysis showed that of the research done during the SARS global emergency, the majority of it was published well after the emergency was over: only 22% of the studies were submitted, 8% accepted, and 7% published during the epidemic of Xing et al. (2010). The contrast with the COVID-19 pandemic could not be clearer. There has been an outpouring of research, and most of this research is rapidly and freely available online, in the first instance predominantly on preprint servers (Fraser et al., 2021). Although preprint servers have a long history in some disciplines, until the COVID-19 pandemic their use and indeed their very acceptability in medical publishing were untested.

OPEN ACCESS

Edited by:
Linda Suzanne O'Brien, Griffith University, Australia

Reviewed by:
Juan Pablo Alperin, Simon Fraser University, Canada

***Correspondence:**
Virginia Barbour
ginny.barbour@qut.edu.au

***ORCID:**
Virginia Barbour
orcid.org/0000-0002-2358-2440

Specialty section:
This article was submitted to Scholarly Communication, a section of the journal Frontiers in Research Metrics and Analytics

Received: *31 August 2021*
Accepted: *08 November 2021*
Published: *03 January 2022*

Citation:
Barbour V (2022) Openness, Integrity, Inclusion, and Innovation in Scholarly Communication: Competing or Complementary Forces? Front. Res. Metr. Anal. 6:767869. doi: 10.3389/frma.2021.767869

medRxiv, founded by the BMJ and Yale University (Rawlinson and Bloom, 2019) in association with Cold Spring Harbor Press, the founders of *bioRxiv*, has perhaps been the standout success for publishing in the pandemic. Launched just before the pandemic, it was perfectly placed to support the publishing effort required but saw its submissions rise from a few hundred in 2019 (Bloom, 2020) to more than 13,000 preprints related to the pandemic to date (medRxiv, n.d) At the height of the pandemic, it was seeing millions of views per month of its content.

By the standards of preprints, more traditional publishing lagged far behind. Further, until a concerted call from a global coalition of government scientists and policy advisors, led by the Office of Science and Technology at the White House, it was not even clear under what terms research would be made available in the pandemic (Call to Action to the Tech Community on New Machine Readable COVID-19 Dataset – The White House, n.d.). The results of that call, the CORD-19 database, is another example of infrastructure in waiting. The key to the success of CORD-19 was its alignment with a set of key principles—the FAIR principles, originally described for data, which have come to have huge importance in the research data world. FAIR principles, which require high-quality metadata such as permanent identifiers and licenses, facilitate discoverability, machine readability, and interoperability, allowing sophisticated text mining and reuse of the research literature (Wang et al., 2020).

WILL OPEN ACCESS TO RESEARCH ENDURE POST PANDEMIC?

At the time of writing in July 2021, the pandemic continues globally with no obvious end date in sight. The rate of publication has decreased from the peaks of 2020, but the need for research to remain open remains in the face of coronavirus mutations and continuing societal challenges. Despite this, we are beginning to see publishers moving papers behind subscription barriers in a move that is reminiscent of publisher activities in earlier medical emergencies, such as following Ebola outbreaks. These moves illustrate clearly the stranglehold that traditional publishers retain over the dissemination of research publications. It reinforces the need for research publications to be fully open at the time of publication and that is only done by ensuring that articles are openly licensed with Creative Commons licenses. More generally, the pandemic has reinforced the need for a diversity of approaches to publishing models (Shearer et al., 2020) as well as a robust open infrastructure as championed by organization such as Invest in Open (Invest in Open Infrastructure, n.d) to support these models.

WHO LOST OUT IN PANDEMIC PUBLISHING?

The pandemic also laid bare many of the entrenched inequalities in scholarly communication, and indeed in research more generally. Money was poured into research globally on every possible aspect of the pandemic, from basic science such as genomic sequencing, through to analyses of public health. However, as in research in more normal times, the money did not flow equitably, nor were publications from the pandemic truly reflective either of research needs or of the wider researcher landscape. For women, the COVID-19 pandemic "exacerbated pre-existing gender inequity in the STEM workforce across the Asia-Pacific region" according to a 2021 report (Impact of COVID-19 on Women in the STEM Workforce | Asia-Pacific, n.d.) which further noted that "Additional domestic responsibilities, such as supervising school learning at home, caused competing priorities as domestic roles and professional roles overlapped. This resulted in negative impacts on productivity for many women, especially in terms of academic output such as journal publications." Nor were research projects on COVID-19 equally distributed globally. A summary of COVID-19 Funding Trends 2021 noted that "90 per cent of research projects are located in high income countries, with the greatest number in the US." (Special Report, 2021)

TRUST, INTEGRITY, AND REWARDS IN RESEARCH

The pandemic also exacerbated many of the trends in relation to trust—or lack of trust—in research. The rapid availability by necessity of non-peer-reviewed research in the form of preprints and the intense public interest and wide sharing of research through the news media triggered an intense discussion on trust in research. Confidence in research was highlighted by analysis that showed that the pandemic led to "a proliferation of research projects underpowered and unable to achieve their aims" (Norton et al., 2021) but which nonetheless were eagerly pored over and discussed widely. As traditional publishers tried to keep up with the flood of papers, it was notable that some of the most egregious examples of poor-quality research were actually published in high-profile peer-reviewed journals, which had apparently failed in proper scrutiny of research, especially in relation to access to underlying data (Two Elite Medical Journals Retract Coronavirus Papers over Data Integrity Questions | Science | AAAS, n.d.).

In some ways then, the pandemic also accelerated conversations about how to assess research for trustworthiness and how to balance speed of sharing versus scrutiny through peer review—which as is well known is, at best, an imperfect and partial way to assess the quality of research publications. *medRxiv*, which had to develop processes on the fly for the rapid screening of preprints, has, as a result, of the pandemic now a quality control process that, although no substitute for peer review, does seem able to reliably filter out research which has ethical or similar issues. The increase in the amount of research available as preprints—and the scrutiny of this publishing approach—has also led to a wider understanding in the press and wider public arena of what peer review means, and it is common to see now that news reports

will indicate the peer review status of reported research. Furthermore, publishing of research through preprints challenges one of key norms of research evaluation, which is currently overwhelmingly biased to rewarding researchers for publishing in specific journals. The new models of publishing can only accelerate discussions on the urgent need for reform of the incentive system as championed by DORA (Declaration on Research Assessment, n.d) and others.

BUILDING A BETTER FUTURE FOR RESEARCH SHARING

So what comes after the pandemic? In many ways, the pandemic has acted as an accelerator for discussions about open access and open science that previously had been caught up in bureaucratic niceties. In May, this year UNESCO provisionally agreed the text of its Open Science Recommendation (UNESCO Recommendation on Open Science, 2020). Its origin in 2019 was pre-pandemic, but by the time of its release the topic could not be more timely and its preamble referenced the pandemic as follows: "Noting that the global COVID-19 health crisis has proven worldwide the urgency of fostering an equitable access to scientific information, facilitating the sharing of scientific knowledge, data and information, enhancing scientific collaboration and science- and knowledge-based decision making to respond to global emergencies and increase the resilience of societies."

Further international work on open science inspired by the pandemic included an online UN Open Science meeting in July 2021 (United Nations. Open Science Conference, 2021) with more than 2,500 participants with global perspectives on the role of open science in the pandemic, and what lessons need to be learned from the pandemic in addressing the overarching emergency of our time, climate change. The clear consensus was that we cannot reverse the open research and sharing practices that have come to be normalized during the pandemic if we are going to collaborate effectively to combat climate change. In his keynote speech, Prof. Geoffrey Boulton highlighted the report of the International Science Council "Opening the record of science: making scholarly publishing work for science in the digital era" which calls for an urgent and robust reform of the scholarly publishing process according to the following seven principles:

1) There should be universal open access to the record of science, both for authors and readers.
2) Scientific publications should carry open licenses that allow reuse and text and data mining.
3) Rigorous and ongoing peer review is essential to the integrity of the record of science.
4) The data/observations underlying a published truth claim should be concurrently published.
5) The record of science should be maintained to ensure open access by future generations.
6) Publication traditions of different disciplines should be respected.
7) Systems should adapt to new opportunities rather than embedding inflexible infrastructures.

This last principle is perhaps the most important—the need for constant adaptation as needed. If there is one critical lesson that we have learned over the 18 months of the pandemic and which will surely be further reinforced before the pandemic is over, it is that previous models of publishing and research dissemination—in particular our reliance on proprietary publishing and infrastructure and associated incentive structures based solely on publication in specific journals—can no longer be considered fit for purpose.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Bloom, T. (2020). Shepherding Preprints Through a Pandemic. *BMJ.* 371, m4703. doi:10.1136/bmj.m4703

Call to Action to the Tech Community on New Machine Readable COVID-19 Dataset – The White House (n.d). Retrieved August 2, 2021, Available at: https://trumpwhitehouse.archives.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/.

Declaration on Research Assessment (n.d). DORA. Retrieved August 31, 2021, Available at: https://sfdora.org/.

Fraser, N., Brierley, L., Dey, G., Polka, J. K., Pálfy, M., Nanni, F., et al. (2021). The Evolving Role of Preprints in the Dissemination of COVID-19 Research and Their Impact on the Science Communication Landscape. *Plos Biol.* 19 (4), e3000959. doi:10.1371/journal.pbio.3000959

Impact of COVID-19 on women in the STEM workforce | Asia-Pacific. (n.d). 89.

Invest in Open Infrastructure (n.d). Invest in Open Infrastructure. Retrieved August 31, 2021, Available at: https://investinopen.org/.

medRxiv (n.d). medRxiv.org—The Preprint Server for Health Sciences. Available at: https://www.medrxiv.org/ (Accessed July 31, 2021).

Norton, A., Bucher, A., Antonio, E., and Mounier, P. (2021). A living mapping review for COVID-19 funded research projects: nine-month update [version 4; peer review: 2 approved]. *Wellcome Open Res.* 5, 209.

Rawlinson, C., and Bloom, T. (2019). New Preprint Server for Medical Research. *BMJ.* 365, l2301. doi:10.1136/bmj.l2301

SARS | Basics Factsheet | CDC (n.d). Retrieved August 2, 2021, Available at: https://www.cdc.gov/sars/about/fs-sars.html.

Shearer, K., Chan, L., Kuchma, I., and Mounier, P. (2020). Fostering Bibliodiversity in Scholarly Communications: A Call for Action. Zenodo. doi:10.5281/zenodo.3752923

Special Report: Covid-19 Funding Trends (2021). Research Professional News. Available at: https://www.researchprofessionalnews.com/rr-news-world-special-report-covid-19-funding-trends/.

Two elite medical journals retract coronavirus papers over data integrity questions | Science | AAAS (n.d). Retrieved August 2, 2021, Available at: https://www.sciencemag.org/news/2020/06/two-elite-medical-journals-retract-coronavirus-papers-over-data-integrity-questions.

UNESCO Recommendation on Open Science (2020). The UNESCO Open Science Recommendation was adopted on 23rd November 2021. Available at: https://en.unesco.org/science-sustainable-future/open-science/recommendation.

United Nations. Open Science Conference (2021). United Nations; United Nations. Available at: https://www.un.org/en/library/OS21 (Accessed August 31, 2021).
Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., et al. (2020). CORD-19: The COVID-19 Open Research Dataset. ArXiv:2004.10706 [Cs]. Available at: http://arxiv.org/abs/2004.10706.
Xing, W., Hejblum, G., Leung, G. M., and Valleron, A.-J. (2010). Anatomy of the Epidemiological Literature on the 2003 SARS Outbreaks in Hong Kong and Toronto: A Time-Stratified Review. *Plos Med.* 7 (5), e1000272. doi:10.1371/journal.pmed.1000272

Conflict of Interest: VB is employed part-time as Director at Open Access Australasia.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Research Metrics and Analytics

CONCEPTUAL ANALYSIS
published: 12 January 2022
doi: 10.3389/frma.2021.752336

Trust in Scholarly Communications and Infrastructure: Indigenous Data Sovereignty

*Katharina Ruckstuhl**

Otago Business School, University of Otago, Dunedin, New Zealand

Many Indigenous people have a deep mistrust of research, with some describing research as one of the "dirtiest" words in Indigenous language. The histories and experiences behind such mistrust are long and painful. Given what has been perceived as Indigenous objectification at the hands of largely Anglo-European others for research from which they fail to benefit, many communities now refuse research unless it is undertaken under certain, Indigenous-defined circumstances. Such refusal is a move away from others92 purposes and a move towards autonomy and self-determination. For some, this is a statement of sovereignty and it applies to all areas of endeavour, including the new frontiers of research and the structures that support them, such as datification of knowledge. This article examines data sovereignty from the perspective of Indigenous peoples. While data sovereignty has become a ubiquitous concern, Indigenous data sovereignty arises from contexts specific to Indigenous peoples. The focus of this article is to provide a brief overview of recent data sovereignty developments, along with the context that lies behind these activities. Through this examination, implications for trust in scholarly communications will be discussed.

Keywords: Indigenous data sovereignty, research infrastructure, decolonization, data governance, traditional knowledge, Nagoya Protocol, TK labels, metadata

OPEN ACCESS

Edited by:
Daniel W. Hook,
Digital Science, United Kingdom

Reviewed by:
Diane Rasmussen Pennington,
University of Strathclyde,
United Kingdom
Xing Zhao,
East China Normal University, China

***Correspondence:**
Katharina Ruckstuhl
katharina.ruckstuhl@otago.ac.nz

Specialty section:
This article was submitted to Scholarly Communication, a section of the journal Frontiers in Research Metrics and Analytics

Received: *02 August 2021*
Accepted: *13 December 2021*
Published: *12 January 2022*

Citation:
Ruckstuhl K (2022) Trust in Scholarly Communications and Infrastructure: Indigenous Data Sovereignty. Front. Res. Metr. Anal. 6:752336. doi: 10.3389/frma.2021.752336

INTRODUCTION

Many Indigenous people have a deep mistrust of research, with some describing research as one of the "dirtiest" words in Indigenous language. The histories and experiences behind such mistrust are long and painful. Given what has been perceived as Indigenous objectification at the hands of largely Anglo-European others for research from which they fail to benefit, many communities now refuse research unless it is undertaken under certain, Indigenous-defined circumstances. Such refusal is a move away from others' purposes and a move toward autonomy and self-determination. For some, this is a statement of sovereignty and it applies to all areas of endeavor, including the new frontiers of research and the structures that support them, such as datification of knowledge.

This article examines data sovereignty from the perspective of Indigenous peoples, focusing on data held in government or state-funded research organizations. While data sovereignty has become a ubiquitous concern, Indigenous data sovereignty arises from contexts specific to Indigenous peoples. The focus of this article is to provide a brief overview of recent data sovereignty developments, along with the context that lies behind these activities. Through this examination, implications for trust in scholarly communication and infrastructure will be discussed.

The article proceeds as follows. The first section examines the impact of colonialism in relation to research derived from Indigenous people, their lands and genetic and cultural resources. Particular attention is paid to Indigenous notions of sovereignty, in contrast to nation-state or individual notions, from which is derived more recent call for Indigenous data sovereignty (IDS). I then look at the various contexts and infrastructures of data—administrative data held in government databases, biologically based data in biobanks held in research organizations, and data in collecting organizations such as galleries, libraries, archives and museums. This section identifies how Indigenous people are developing policies and processes for data sovereignty. Drawing on the previous sections, the final section considers implications for trust in scholarly communication and infrastructure, and the actions needed to engender trust.

DATA SOVEREIGNTY AND ITS INDIGENOUS CONTEXT

Colonialism and Sovereignty

With the increased digitization of all forms of information, how data is stored, attributed, categorized, organized, owned, managed and used has become a ubiquitous concern from the micro level of the individual to the macro levels of nations and global organizations. This avalanche of data and the ease in which it crosses borders has seen some call for data sovereignty. This ranges from calls for personal data sovereignty (Micheli et al., 2020) to proposed policies for the sovereignty of European data and digital infrastructure (EIT Digital, 2020).

There is no one definition of data sovereignty, although there are overlapping features, many of which relate to the rights of individuals, collectives or nations to have control and power over data whether within territorial locations or cross-jurisdictionally. Data sovereignty is also associated with privacy and the constraint of information flows, ownership, inclusiveness and the representation of different groups into decisions about how data is used or re-used (Hummel et al., 2021).

That sovereignty is the word to describe a desired solution to the problems associated with data resonates with Indigenous people, however not necessarily for the reasons found in others' use of the word. In their analysis of the discourse on digital sovereignty, Couture and Toupin (2019) note territorial authority and thus sovereignty had been an aim of the nation-state. However, the sovereignty ambition of nation-states has always been contested by other formations such as those of kinship, religion, tribe or feudal ties. Moreover, an absolutist position on sovereignty is increasingly bounded or limited by mechanisms such as international treaties, pacts and agreements, the activities of trans-global organizations, and global infrastructures such as telecommunications networks. Thus, while nation states may "imagine" they are sovereign (Anderson, 1983), this has often been more limited in practice.

It is in relation to nation states, and their imagined sovereignty over bodies, resources and territories, that Indigenous demands for data sovereignty arise. Before exploring this, it is important to note that it is always difficult and sometime perilous to define a group, including Indigenous peoples who are not homogenous. Such definitions can be fraught, particularly when constructed outside of Indigenous peoples' views of themselves (Corntassel, 2003). The United Nations (UN) has shied away from defining the word "Indigenous" preferring to "identify rather than define" Indigenous peoples (United Nations, 2006), with the following as common characteristics (Daes, 2008):

(a) Priority in time, with respect to the occupation and use of a specific territory;
(b) The voluntary perpetuation of cultural distinctiveness, which may include the aspects of language, social organization, religion and spiritual values, modes of production, laws and institutions;
(c) Self-identification, as well as recognition by other groups, or by State authorities, as a distinct collectivity; and
(d) An experience of subjugation, marginalization, dispossession, exclusion or discrimination, whether or not these conditions persist.

All of the above characteristics are worth considering in relation to IDS and the consequent flow-on effects into trust in scholarly communication infrastructure.

First, as (a) states, Indigenous peoples occupied and continue to inhabit specific territories. Many of these Indigenous people were subsequently displaced from or dispossessed of these territories, for the most part forcibly, despite in some cases such as in Aotearoa New Zealand, Canada and the United States, treaties being signed to maintain or share territories. Whether displaced from territory or not, the overwhelming experience for many Indigenous people is of marginalization and discrimination. This is not a matter of colonial pasts from which Indigenous people have moved on as they become subsumed or assimilated into nation states. It is a structured and ongoing reality for Indigenous people that manifests itself in socio-economic disparities and, for some, ongoing violence for territories over which Indigenous peoples consider they have rights and obligations.

Despite the past and ongoing efforts of some nation states to eradicate the cultures, languages and practices of Indigenous peoples, whether in the cause of the one sovereign nation, or whether to civilize and promote "development" of Indigenous people, distinct Indigenous cultures remain. Again, while it is fraught to essentialize cultures, as (c) in the UN identification suggests, there remain patterns of worldview and practice to which many Indigenous peoples ascribe. These include:

- Distinct knowledge systems, variously described for example as Indigenous knowledge (IK), traditional ecological knowledge (TEK) or folk knowledge. Such knowledges rather than primitive or pre-modern are characterized by dynamism and adaptation (Pool, 2015). Such knowledges also do not discount that which is spiritual or "revealed" knowledge, but rather use such knowledge alongside traditional and empirical knowledge (Dei, 2000);
- A distinct relationship to place to which Indigenous people have a sense of guardianship and protection for future generations, whether such generations are human or not

(Colburn, 2021). From this relationship arises a sense not only of belonging but also connectedness, rights and obligations (Katerere et al., 2019);
- A collectivist rather than individualist approach to all facets of material life which can include how resources are used or distributed, who has the rights and obligations toward such resources, and how these resources are viewed such as being seen as "gifts" or "treasures" from creators (Colburn, 2021).
- Distinct languages through which knowledge, culture and relationship to place are transmitted intergenerationally despite 50% of the world's 6,500 languages under threat (Mackenzie and Davis, 2018)

In summary, Indigenous people have maintained their specific identities in a manner that that can be described as "survivance," which is "more than survival, more than endurance or mere response . . . [but is] an active repudiation of dominance, tragedy, and victimry" (Vizenor, 1998, p. 13). This active repudiation extends to how Indigenous people have been positioned within nation-states through the plethora of laws, institutions, structures and infrastructures that maintain colonialism or settler colonialism (Gover, 2015). This includes the innumerable scholarly mechanisms associated with disseminating knowledge and research about Indigenous people, their lands, resources and cultures.

Colonialism and Research

For many Indigenous people 'research' has been, and for some, continues to be one of the "dirtiest" words in the Indigenous world's language (Smith, 2009). Hence there has been little trust in the research mission and the pursuit of generic knowledge and universal "truths" that are divorced from Indigenous lives, with research viewed as complicit in past and ongoing colonialism.

One emblematic example is James Cook's Transit of Venus voyage to the South Pacific in 1769, for which the Royal Society successfully raised £4000. Such funding was forthcoming not only out of scientific curiosity but through the Admiralty's secret instructions to Cook to discover unknown countries and gain knowledge of these to advance British trade. Meanwhile, the Royal Society proposed that gentleman botanist Joseph Banks convince the discovered savage and brutal nations of European superiority (Igler, 2019). While such attitudes were typical of the era, Banks' reputation rests on the collection of 30,000 botanical and over 1,000 animal specimens gathered during the three-year voyage. These were the first specimens from the South Pacific seen in Britain and catapulted Banks' career and prestige, leading to his eventual Presidency of the Royal Society and Directorship of the Royal Gardens at Kew (Agnarsdóttir, 2019).

Banks' vision for Kew was that it might become a botanical exchange house, whereby collectors would taxonomically name and then bring back new plants that were economically useful to expand the British Empire (Hopper, 2013). And indeed this is what occurred, with for example, Brazilian rubber and Andean cinchona bark from which quinine is made eventually transferred via Kew to start industries in Malaya and India respectively. Expropriation of such specimens and the Indigenous knowledge that went along with them, converted science knowledge into imperial economic power. The ongoing consequence of this is that nations from which Indigenous knowledge was acquired to identify the utility of a specimen, now pay 'rents' in the forms of patents licenses or fees to access the biomedical or other technologies derived from these appropriated specimens (Brockway, 2011).

While Kew Gardens has recently acknowledged its role in Empire (Parveen, 2021), biocultural appropriation, as Brockway suggests, continues. This despite the Nagoya Protocol that, under the Convention of Biological Diversity, aims to provide a transparent legal framework for the fair and equitable sharing of benefits arising out of the use of a nation's genetic resources, including the traditional knowledge associated with genetic resources (United Nations, 2015). That such a protocol has been necessary speaks to the practices of what some call "biocolonialism," which can be seen as a process whereby genetic resources from traditional medicines and seeds are altered sufficiently to render them patentable, thereby allowing corporations or research organizations to commodify and profit from the sale of such knowledge (Harry, 2011). As Tauli-Corpus has argued, Indigenous people do not understand the logic whereby plants and seed varieties developed and preserved over thousands of years by Indigenous people become "improved" in laboratories that then confers an intellectual property right to the "inventor" (Tauli-Corpus, cited in Whitt, 1998, p. 39).

Intellectual property regimes fail to protect collective Indigenous knowledge, hence retrospective global attempts, such as the Nagoya Protocol, to address this through access and benefit-sharing. That this continues to be an issue can be seen in disputes brought by Indigenous people around patenting attempts of Ojibwe wild rice, Mexican maize and Hawaiian taro (McGonigle, 2016). Even when pharmaceutical companies attempt to recognize TEK, such as Shaman Pharmaceutical's trade agreements with Amazonian peoples in the 1990s (McGonigle, 2016) or the South Africa's Council for Scientific and Industrial Research benefit sharing agreement with San (Vermeylen, 2007), the ultimate benefits, economic or otherwise, to Indigenous communities remain uncertain or negligible.

This brings us to the issue of patenting of other life forms and, particularly in our current context of the COVID-19 pandemic, ongoing research into the human genome. From an Indigenous perspective, the "promise" of genomic research to alleviate health problems is undercut by the experience of unethical practice and misuse of data (Jacobs et al., 2010). For example, the Havasupai Tribe of northern Arizona filed and won a lawsuit in 2010 against the Arizona Board of Regents over the misuse of their genetic samples, collected for research on type 2 diabetes in 1989 but subsequently used for studies on schizophrenia, ethnic migration, and population inbreeding—areas disapproved of by the original donors (Garrison, 2013). While informed consent is a central tenet of ethical practice in the human sciences, Reardon and TallBear (2012) argue that at least in the US context, when it comes to Indigenous populations, there is an overwhelming belief of the right to pursue science to advance universal knowledge. In such cases, Indigenous peoples acting to protect their own interests might be seen as hampering the knowledge commons.

Such experiences are unfortunately common globally (Kowal et al., 2012).

A more recent example of this right to pursue knowledge involves the Institute for Development Research (IRD) in France, accused of biopiracy for patenting an anti-malaria drug without acknowledging the French Guianan indigenous community's traditional medicinal knowledge. As in the Havasupai case, the researchers initially saw themselves practicing a science based on the greater good, having collected the samples in 2009 "in good faith." In this case, rather than a direct payment, the IRD agreed to a benefit-sharing arrangement with Guianan authorities as recommended under the Nagoya Protocol (Pain, 2016). While the European Union, of which France is a member, only legally adopted the Nagoya Protocol in 2014, the IRD's retrospective agreement indicates the increasing pressure from Indigenous groups for fair and equitable benefit. Without due diligence of the sources of genetic materials, European researchers can face fines of up to €810,000 and imprisonment. Currently EU interpretation of the Protocol excludes information stored in databases, however, this is under contestation and may change (V.O. Patents Trademarks, 2019).

Raw genomic data has emerged as a global commodity in the last few years, with research organizations increasingly interested in small populations, such as Indigenous people (Fox, 2020). Such commodification, and the historic harms to Indigenous people of which the Havasupai is but one example, have hastened Indigenous efforts to control how such data is accessed, stored and used. There have been calls not only for Indigenous-framed ethical approaches to consent, but also for greater oversight and governance of both the original genetic material and the data that is derived from the material. This then brings us to IDS which has gained popularity as a terminology more recently, but has been in the policies of some tribal groups such as Cherokee since at least the 1990s (Bardill, 2017).

Given the historic experience of many Indigenous people, data sovereignty is a form of "corrective justice" after several centuries of policies that marginalized and diminished the rights of Indigenous peoples (Tsosie, 2021). What distinguishes IDS is an emphasis on tribal or tribal nation self-determination and autonomous decision-making (Hudson et al., 2017), and a rebalancing of power relationships. Thus, while data sovereignty shares some of the concerns of the nation state to control flows of data, IDS is in fact a challenge *against* the nation state and its ontological foundations and presumptions (Moreton-Robinson, 2020). And while individuals may call for personal data sovereignty, particularly in relation to privacy, IDS pushes against a solely individualist approach to espouse *collective* principles based on long-held worldviews and practices (GIDA, 2019).

Simply put, IDS is the right of Indigenous peoples to control the collection, governance, ownership, and application of data about their people, lifeways, land and resources (Kukutai and Taylor, 2016). Where those data reside, as suggested above, is overwhelmingly in various non-Indigenous repositories, both public and private. How then, can data sovereignty be exercised, and what implications does this have for trust in scholarly communications and infrastructure?

RECENT DEVELOPMENTS IN INDIGENOUS DATA SOVEREIGNTY POLICY AND PRACTICE

Administrative Data

In the public sphere, statistical administrative data collected for government policy purposes often categorizes Indigenous people from the "5D" perspective, i.e., difference, disparity, disadvantage, dysfunction and deprivation. It is not the data itself that is the problem but the purposes for which such data are analyzed and then used. These data are often gathered from a research perspective that aggregates different tribal collectives, decontextualizes them from their social and cultural context and analyses Indigenous people as problematic in contrast to other groups (Walter and Suina, 2019). This "deficit" data analysis fails to take account of Indigenous priorities, values, culture, lifeworlds and diversity (Walter and Suina, 2019) or address Indigenous ability to develop their own nation-building aspirations (Rainie et al., 2017). Hence, an increasing Indigenous focus is on the collection and analysis of data that prioritizes Indigenous-defined objectives thereby reframing narratives of Indigenous people as deficient and lacking in some decontextualized comparative metric such as health, education, housing (Rainie et al., 2019). This more strengths-based or capability approach (Sen, 2001) posits Indigenous people as more than proficient at solving their own issues, provided State infrastructure and resources are equitably provided.

From a practice perspective, there are examples of administrative data being either co-constructed with or controlled by Indigenous people. For example, the Canadian OCAP® principles of ownership, control, access, possession were a Canadian response to providing a framework for governance and statistical practices of health data. OCAP® asserts Indigenous rights to control and benefit from their data with impacts on other national bodies and educational institutions that have likewise altered their data practices to empower Indigenous data control (Walker et al., 2017). Flow-on effects have included broader Indigenous-led research protocols, jurisdictional control and development of best practices for research using First Nations, Inuit and Métis data (Rowe et al., 2021).

Biodata

As discussed previously in relation to genetic material there is a new Indigenous focus on not only ethical collection and consent but also secondary use of data (Garrison et al., 2020). Biobanks hold human biological materials and/or genetic information along with associated demographic and health information (Beaton et al., 2016). Given the global exchange of data, and the need to represent accurately population genetics to provide tailored health solutions, there is the need to include minority populations. One argument is that, contra to a belief that individuals are "gifting" their genetic biomarkers to help develop health breakthroughs such as precision medicine tools—a type of "public good"—there needs to be more of a focus on genetic stewardship. Such thinking arises from the observation that many Indigenous people fail to be the recipients of the proposed

benefits of health innovations, even when the data is in the public domain. Hence there are increasing calls for either the development of Indigenous-controlled biobanks or for increased governance over existing biobanks (Tsosie et al., 2021).

There are a number of examples of good practice, where Indigenous groups and genetic researchers have developed positive working relationships, grounded in Indigenous worldviews of health (McWhirter et al., 2012) and targeted at developing Indigenous capacity and governance (McWhirter et al., 2015). Likewise, there are emerging examples of biobank data governance, for example, Aotearoa New Zealand's He Tangata Kei Tua, a culturally informed policy and practice for biobanks in relation to governance, operational, and community engagement activities (Beaton et al., 2016). Similarly, the four-year funded Canadian "Silent Genomes" project that, along with aiming to reduce health-care disparities and improve diagnostic success for children with genetic diseases from Indigenous populations, also aims to develop a First Nations governed background variant library as a reference to allow effective precision diagnosis (Garrison et al., 2019). Another example of Indigenous biobank control is the Native BioData Consortium created to keep Indigenous research samples and data within the provenance and governance of Indigenous communities (Tsosie et al., 2021).

Turning to plant materials, at the aforementioned Kew Gardens, there is now a recognition that imperialist views still prevail in relation to its collections, with scientists continuing to report how new species are discovered every year, despite the knowledge of and use of such plants for thousands of years by local people (Antonelli, 2020). It does not take much to find related views in academic publications. A 2020 article in the journal Antibiotics describes how "many students wrote their masters and PhD theses on ethnomedicinal uses by the Karen people" [an Indigenous hill tribe on the border of Myanmar and Thailand] but "strangely" did not focus on how Karen people's botanical knowledge was used to treat ailments like fever. Therefore, the author complied "the most comprehensive list to date of botanical species that are treated as therapies against fever by the Karen people… cover[ing] 25 Karen villages in Thailand and compiled a list that includes 125 species," helpfully listing a taxonomy of the "high value plant species" on the open access mdpi site (Phumthum and Sadgrove, 2020). While the author does not claim to have discovered these plants, there is a "terra nullius" implication that Karen plant knowledge is "free" because the Karen do not have territorial sovereignty to the land on which the plants are found (Rojas-Páez and O'Brien, 2021). This carries on a mode of colonial thinking into science that was once used to dispossess many Indigenous people of their lands because it was "terra nullius" or "belonged to no one" (Harry, 2001).

However, there are also examples of scientists acknowledging that they are not the "discoverers" of new plants, with one Polish PhD student, Mateusz Wrazidlo, working with the Indigenous community of the Guiana Highlands to give a Pemón Arekuna name to an orchid species new to science. Wrazidlo states that this was aimed at "de-colonizing science nomenclature and giving more representation to indigenous [and] local languages" (Kimbrough, 2021). Such a practice embodies recent calls from ethnobiologists to decolonize institutions, projects and scholarship. The authors acknowledge that centralization of biocultural resources in Euro-American repositories and archives has been extractive and alienated Indigenous people from their cultural and biological heritage. Hence the authors recommend a set of practices that include repatriating biocultural collections to Indigenous stewards, ensuring that data around biocultural classifications accurately represents Indigenous understanding, showing reciprocal relationships in research rather than doing "parachute science" where researchers visit, collect and return to their home institutions, and respecting data sovereignty (Mcalvay et al., 2021).

Data in Galleries, Archives, Libraries and Museums

Much tangible and intangible knowledge, in the forms of stories, songs and oral traditions resides in art galleries, libraries, archives and museums. While Indigenous people have been demanding repatriation of human ancestors and their cultural artifacts over many years, the reality is that institutions continue to hold vast Indigenous collections. There is an accelerating movement to incorporate Indigenous framed archival practices (Callison et al., 2016) and an acknowledgment of the role of such institutions in perpetuating colonialism (Giblin et al., 2019). At a structural level, there are well-documented cases of histories of racist and offensive subject terms and classification schemes that homogenize and essentialize and that have remained static, retaining their colonialist roots. Far from being neutral classifications, library taxonomies are inherently biased, reflecting the dominant perspective of the "other" (Vaughan, 2018) For example, Indigenous people do not classify themselves as "indigenous," "native," "aboriginal," "Amer-Indian" or other such blanket description. As a Māori woman from Aotearoa New Zealand, I identify my tribal affiliations as Ngāi Tahu and Rangitāne. However, similar to government administrative data, library cataloging collectivizes groups of people to enable search, misnaming or using non-Indigenous terms to explain phenomena and maintaining a "rules-based" orientation to cataloging such as the Library of Congress, Dewey or Anglo-American Cataloguing Rules (Duarte and Belarde-Lewis, 2015). Such rules can be difficult to change, even when societal attitudes have.

One response to this has been to examine the metadata in archival classification systems. Metadata is the "data about data," or the cataloging information about a collection. It describes information, it enables administrative functions to ensure data is stored, preserved and able to be accessed technically, it identifies rights e.g., copyright, and it structures disparate individual components into larger more meaningful understandings. As such it is ideologically based, and neither neutral nor objective but rather subjective in what it includes, omits or describes (Gartner, 2016; Haberstock, 2020). While user or "social"-generated, as opposed to archival specialist generated metadata is becoming more a feature (Alemu, 2018), Indigenous-generated metadata functions additionally to address colonial power structures.

In order to decolonize archival metadata, some institutions are participating with Indigenous groups to develop more nuanced metadata labels or "tags." For example, in Aotearoa New Zealand librarians are adding Māori terms into subject headings, including authority files with Māori terms; instructions for faceting Western concepts such as "myths and legends" with Māori concepts of "history and genealogy"; and rules for faceting records to include the perspectives of the relevant tribes in a document (Duarte and Belarde-Lewis, 2015). In another project, Zuni elders worked with the A:shiwi A:wan Museum and Heritage Center to catalog Zuni items excavated in the 1920s. In this project, additional metadata schema were required to the "normal" to incorporate uses and practices of, and stories and narratives around objects (Haberstock, 2020). For some institutions, specificity about Indigenous material in collections can reveal a lack of knowledge, with metadata schema failing to associate content and the authorities of tribal nations, clans or families, their communities, or territories.

In a move similar to the repatriation of human remains or artifacts, Anderson and Christen (2019) advocate for "digital repatriation," which cedes decision making about access, narration, curation, and circulation of research materials to the original stewards that in turn affects future documentation, recording, metadata, as well as publication. For them, attribution is key given that photographs, sound recordings, films, artworks and manuscripts documenting Indigenous lives are the property of the "author" under copyright law. This is similar to the way that an inventor who develops a treatment based on Indigenous medicinal knowledge can be granted a property right in the form of a patent. Given that authorship circulates in perpetuity through the infrastructures of research—catalogs, records, publications and citations—digital repatriation acts as a rupture to colonialism through re-attribution to and control by originating communities. The example that Anderson and Christen highlight is that of sound recordings of Passamaquoddy singers, recorded in the 1890s by ethnographer Fewkes without attributions but through interactions with descendants of the original singers, re-attributed to the individuals who supplied the voices. More than that, however, the Library of Congress record contextualizes the recording, includes cultural and traditional narratives supplied by the elder descendants and applies "Traditional Knowledge Labels" to the record, including one that indicates that the material is non-commercial.

Traditional knowledge labels (TK labels) are an emergent digital rights tool aimed at enabling Indigenous control over their materials in a context of increasing digitization of cultural heritage, its global circulation via the internet with varying degrees of open access, and third-party use of such material (Reijerkerk, 2020). Anderson and Christen have adapted the Creative Commons licensing approach that ameliorates against copyright to develop the Local Contexts platform (https://localcontexts.org/) that hosts TK licenses, labels and notices. The labels are designed to highlight that local Indigenous values and appropriate use remain embedded within archival materials, even if they have been outside community ownership for generations (Anderson and Christen, 2013). The labels themselves have been extensively trialed with Indigenous communities and can be applied to tribal archives to explicate access permissions internal to the tribe or externally to others who may find tribal cultural material online. To the TK labels have been added Biocultural (BC) Labels and Notices that operate in a similar way but for data derived from genetic resources to enhance the capacity for Indigenous control of Indigenous data (Anderson and Hudson, 2020). Additionally, they provide a visible machine-readable, persistent and durable connection between Indigenous communities and researchers, genetic resources, generated digital sequence information, and knowledge that exists as metadata in sample/data repositories and can appear on published articles (Liggins et al., 2021).

INDIGENOUS DATA SOVEREIGNTY AND IMPLICATIONS FOR TRUST IN SCHOLARLY COMMUNICATION AND INFRASTRUCTURE

TK and BC Labels are at the forefront of data stewardship and data governance models (van Geuns and Brandusescu, 2020) that globally have become urgent areas of enquiry, as explained at the start of this article. Indigenous enquiry additionally extends into areas such as:

- artificial intelligence and its potential to re-inscribe coloniality based on its original faulty data sets (Lewis, 2020);
- the critical examination of open access data standards such as the FAIR principles for scientific data management and stewardship, developed to enhance the ability of machines to automatically find and use research data and to supporting its reuse by individuals (Wilkinson et al., 2016). While the FAIR principles (Findable, Accessible, Interoperable, Reusable) allow for open access, such principles can be at odds with Indigenous positions in relation to certain types of tribal data. Hence, alongside FAIR, the CARE principles (Collective Benefit, Authority to Control, Responsibility, and Ethics) have been proposed. The principles describe high-level actions applicable within various data settings with a goal to implement CARE and FAIR across the data lifecycle in tandem (Rainie et al., 2020);
- Indigenous data provenance and the rules by which Indigenous peoples' data should be described and recorded. This current working group of the IEEE will make recommendations for metadata fields that can be used across industry sectors, including machine learning, artificial intelligence, contexts, biodiversity and genomic science innovation and other associated databases. This will include connecting data to people and place, and when appropriate, supporting future benefit sharing options (IEEE Standards Association, 2020).

IDS has ongoing implications for trust in scholarly communication and infrastructure. Indigenous people expect that at every level of the research lifecycle—from the accessing of raw data, whether qualitative or quantitative, to its storage in databases, biobanks or herbaria, and then onto its analysis, eventual publication and potentially secondary re-use of originating data—there will be policies and institutional practices that reflect the realities of Indigenous peoples, be useful

for Indigenous purposes, and remain under Indigenous control, while promoting knowledge discovery and innovation (Rainie et al., 2020, p. 8).

For scholarly publishers, this is more than adopting diversity and inclusivity policies, although these are undoubtedly necessary (Dawson et al., 2020). It is also more than increasing Indigenous and other under-represented groups' accessibility to prestige publications, although this too is needed (Collyer, 2018). Rather, it is an examination of the "core" machinery of scholarly infrastructure—universities, ethics committees, funders, and others—of which data is increasingly its key component. Part of this examination and a consequent response may include applying digital rights management protocols, such as the TK and BC Labels and Notices, into publishing and related data management systems. For example, in 2020 ORCID, a not-for-profit software platform that provides a unique, persistent digital identifier to individual researchers ran a series of global webinars alongside the Global Indigenous Data Alliance and the US Indigenous Data Sovereignty Network to raise awareness of IDS. The webinars were an introduction to IDS aimed at research funders, institutions, publishers, and individual researchers (Akee et al., 2020). Following on from these webinars, ORCID is working with Local Contexts to create a "workflow" between the two organizations that enables researchers to request research or use existing Indigenous data. When the researcher is approved by the tribal group, the researcher's ORCID record will be updated with the metadata describing the work and tribal approval. This will enable Local Contexts to update a researcher's record to indicate they have permission or consent from the tribal group to conduct research or use the data.

Partnering with Indigenous groups, supporting conscientization of Indigenous issues, diversifying the workface are important, but they are insufficient. Research infrastructures need to move beyond the metaphoric rhetoric of "decolonization" (Tuck and Yang, 2012), to the actuality: making room for Indigenous decision-making and authority over their materials, wherever they may be located. Seen in this light, ORCID's approach to Indigenous data management is a core infrastructure response to IDS. It is a small but significant way by which Indigenous groups may have some control over access to and use of their own cultural, bio-cultural or genomic data. Potentially this then acts as a mechanism whereby provenance of such data is "on-the-record" and hence helps to identify those tribal groups that may need to be in discussions should benefits be eventually derived.

Indigenous trust in scholarly communication and infrastructure will be derived from the sum of the sets of activities described in this article. These include: reciprocal relationships; using Indigenous nomenclature and language; access and benefit-sharing arrangements; avoidance of "terra nullius," "common good" or "universalist" thinking and methodologies that are then re-embedded in publications and open access data; global, national and institutional governance protocols and standards around Indigenous data; re-inscribing attribution and provenance into metadata; and using digital tools that reinforce Indigenous rights and stewardship.

As has been explained, IDS is but the latest field in a long history of Indigenous action to assert sovereignty. What is different now is that there are theories, tools, approaches and protocols that can be applied across a range of research infrastructure and settings to acknowledge and respond to Indigenous demands for data sovereignty. Non-response or inadequate response may lead to financial and reputational penalties, as the Havasupai and IRD examples suggest. Conversely, genuine efforts to apply IDS tools and methods can enhance reputation and trust. Given the newness of many of these tools and approaches, this will not be an easy or "quick-fix" process.

As the global breadth and depth of activity explained in this article suggest, demands for IDS in state and government run research organizations are increasing. The demand on private sector organizations such as academic publishers and the dissemination infrastructures they rely on are less well-canvassed, although no less pressing. Tools that have been developed to address IDS in the state sector, such as the TK and BC labels, may have relevance, as may policy and ethics approaches. However, it is too early to say to what extent these may be applicable, and further research in this area is warranted. What is clear is that organizations both public and private are increasingly asked to respond to the questions that IDS raises.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

National Science Challenge SfTI Spearhead Project. 2019-S1-CRS Building New Zealands' Innovation Capacity.

REFERENCES

Agnarsdóttir, A. (2019). The young Joseph Banks: naturalist explorer and scientist, 1766–1772. *J. Maritime Res.* 21, 23–44. doi: 10.1080/21533369.2020.1746090

Akee, R., Anderson, J., Carroll, S. R., Cousijn, H., Haak, L., Hudson, M., et al. (2020). *Indigenous Data Sovereignty: Activating Policy and Practice*. Available online at: https://orcid.figshare.com/articles/presentation/Indigenous_Data_Sovereignty_Activating_Policy_and_Practice/12844439/1 (accessed July 10, 2021).

Alemu, G. (2018). Metadata enrichment for digital heritage: users as co-creators. *Int. Inf. Libr. Rev.* 50, 142–156. doi: 10.1080/10572317.2018.1449426

Anderson, B. (1983). *Imagined Communities: Reflections on the Origin and Spread of Nationalism, 2nd Edn*. London: Verso.

Anderson, J., and Christen, K. (2013). "Chuck a copyright on it": dilemmas of digital return and the possibilities for traditional knowledge licenses and labels. *Museum Anthropol. Rev.* 7, 105–126. Available online at: https://www.proquest.com/scholarly-journals/chuck-copyright-on-dilemmas-digital-return/docview/2387825868/se-2

Anderson, J., and Christen, K. (2019). Decolonizing attribution: traditions of exclusions. *J. Radic. Libr.* 5, 113–152. Available online at: https://static1.squarespace.com/static/5d3799de845604000199cd24/t/5d73f62134d32b4f17e85532/1567880738264/Decolonizing+Attribution.pdf

Anderson, J., and Hudson, M. (2020). The biocultural labels initiative: supporting Indigenous rights in data derived from genetic resources. *Biodivers. Inf. Sci. Stand*. 4: e59230. doi: 10.3897/biss.4.59230

Antonelli, A. (2020). Director of science at Kew: it's time to decolonise botanical collections. *The Conversation*. Available online at: https://theconversation.com/director-of-science-at-kew-its-time-to-decolonise-botanical-collections-141070 (accessed July 10, 2021).

Bardill, J. (2017). Comparing tribal research and specimens policies: models, practices, and principles. *Int. Indigenous Policy J.* 8:4. doi: 10.18584/iipj.2017.8.4.4

Beaton, A., Hudson, M., Milne, M., Port, R. V., Russell, K., Smith, B., et al. (2016). Engaging Māori in biobanking and genomic research: a model for biobanks to guide culturally informed governance, operational, and community engagement activities. *Genet. Med.* 19, 345–351. doi: 10.1038/gim.2016.111

Brockway, L. (2011). "Science and colonial expansion. The role of the British Royal Botanic Gardens," in *The Postcolonial Science and Technology Studies Reader*, ed S. Harding (Durham; London: Duke University Press), 126–139. doi: 10.1215/9780822393849-008

Callison, C., Roy, L., and LeCheminant, G. A. (2016). *Indigenous Notions of Ownership and Libraries, Archives and Museums*. Berlin; Boston, MA: Walter de Gruyter. doi: 10.1515/9783110363234

Colburn, R. (2021). "Indigenous entrepreneurship," in *The Palgrave Handbook of Minority Entrepreneurship*, ed T. Cooney (Cham: Palgrave Macmillan), 319–348. doi: 10.1007/978-3-030-66603-3_15

Collyer, F. M. (2018). Global patterns in the publishing of academic knowledge: global north, global south. *Curr. Sociol.* 66, 56–73. doi: 10.1177/0011392116680020

Corntassel, J. (2003). Who is indigenous? 'peoplehood' and ethnonationalist approaches to rearticulating indigenous identity. *Nationalism Ethnic Polit.* 9, 75–100. doi: 10.1080/13537110412331301365

Couture, S., and Toupin, S. (2019). What does the notion of "sovereignty" mean when referring to the digital? *New Media Soc.* 21, 2305–2322. doi: 10.1177/1461444819865984

Daes, E.-I. A. (2008). "Standard-setting activities: evolution of standards concerning the rights of indigenous peoples: on the concept of "indigenous people"," in *The Concept of Indigenous Peoples in Asia. A Resource Book*, ed E. Christian (Copenhagen; Chiang Mai: International Work Group for Indigenous Affairs/Asia Indigenous Peoples Pact Foundation), 29–50.

Dawson, J., Coggins, N. L., Dolechek, M., and Fosad, G. (2020). Toolkits for equity: an antiracist framework for scholarly publishing. *Serials Rev.* 46, 170–174. doi: 10.1080/00987913.2020.1806653

Dei, G. (2000). REthinking the roled o Indigenous knowledges in the academy. *Int. J. Inclusive Educ.* 4, 111–132. doi: 10.1080/136031100284849

Duarte, M. E., and Belarde-Lewis, M. (2015). Imagining: creating spaces for indigenous ontologies. *Catalog. Classif. Q.* 53, 677–702. doi: 10.1080/01639374.2015.1018396

EIT Digital (2020). *European Digital Infrastructure- and Data Sovereignty. A Policy Perspective*. Available online at: https://eit.europa.eu/news-events/news/new-report-european-digital-infrastructure-and-data-sovereignty (accessed July 14, 2021).

Fox, K. (2020). "The illusion of inclusion: large scale genomic data sovereignty and indigenous populations," in *KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, eds R. Gupta, and Y. Liu (New York, NY: Association for Computing Machinery), 3591. Available online at: https://dl.acm.org/doi/pdf/10.1145/3394486.3411071 (accessed February 10, 2021).

Garrison, N. A. (2013). Genomic justice for Native Americans: impact of the Havasupai case on genetic research. *Sci. Technol. Hum. Values* 38, 201–223. doi: 10.1177/0162243912470009

Garrison, N. A., Hudson, M., Ballantyne, L. L., Garba, I., Martinez, A., Taualii, M., et al. (2019). Genomic research through an indigenous lens: understanding the expectations. *Annu. Rev. Genom. Hum. Genet.* 20, 495–517. doi: 10.1146/annurev-genom-083118-015434

Garrison, N. A., Rainie, S. C., and Hudson, M. (2020). Entwined processes: rescripting consent and strengthening governance in genomics research with indigenous communities. *J. Law Med. Ethics* 48, 218–220. doi: 10.1177/1073110520917020

Gartner, R. (2016). *Metadata. Shaping Knowledge From Antiquity to the Semantic Web*. London: Springer Nature. doi: 10.1007/978-3-319-40893-4

Giblin, J., Ramos, I., and Grout, N. (2019). Dismantling the master's house. Third *Text* 33, 471–486. doi: 10.1080/09528822.2019.1653065

GIDA (2019). *Onati Indigenous Data Sovereignty (ID-SOV) Communique*. Availabl online at: GIDA-global.org

Gover, K. (2015). Settler-state political theory, 'CANZUS' and the UN Declaration on the Rights of Indigenous Peoples. *Eur. J. Int. Law* 6, 345–373. doi: 10.1093/ejil/chv019

Haberstock, L. (2020). Participatory description: decolonizing descriptive methodologies in archives. *Arch. Sci.* 20, 125–138. doi: 10.1007/s10502-019-09328-6

Harry, D. (2001). *Biopiracy and Globalization: Indigenous Peoples Face a New Wave of Colonialism*. Available online at: http://ipcb.org/publications/other_art/globalization.html (accessed July 10, 2021).

Harry, D. (2011). Biocolonialism and indigenous knowledge in United Nations discourse. *Griffith L. Rev.* 20, 702–728. doi: 10.1080/10383441.2011.10854717

Hopper, S. (2013). From botany bay to breathing planet: an Australian perspective on plant diversity and global sustainability. *Pac. Conserv. Biol.* 19, 356–365. doi: 10.1071/PC130356

Hudson, M., Anderson, K., Dewes, T. K., Temara, P., Whaanga, H., and Roa, T. (2017). "He matapihi ki te mana raraunga – Conceptualising big data through a Māori lens," in *Apperley, He Whare Hangarau Māori – Language, Culture and Technology*, eds H. Whaanga, T. T. Keegan (Hamilton: Te Pua Wānanga ki te Ao/Faculty of Māori and Indigenous Studies), 64–73.

Hummel, P., Braun, M., Tretter, M., and Dabrock, P. (2021). Data sovereignty: a review. *Big Data Soc.* 8, 1–17. doi: 10.1177/2053951720982012

IEEE Standards Association (2020). *P2890 - Recommended Practice for Provenance of Indigenous Peoples' Data*. Available online at: https://standards.ieee.org/project/2890.html (accessed July 20, 2021).

Igler, D. (2019). The questions they asked: Joseph Banks and naturalists in the Pacific Ocean. *J. Maritime Res.* 21, 63–75. doi: 10.1080/21533369.2019.1705574

Jacobs, B., Roffenbender, J., Collmann, J., Cherry, K., Bitso, L. L., and Bassett, K. (2010). Bridging the divide between genomic science and indigenous peoples. *J. Law Med. Ethics* 38, 684–696. doi: 10.1111/j.1748-720X.2010.00521.x

Katerere, D. R., Applequist, W., Aboyade, O. M., and Togo, C. (2019). *Traditional and Indigenous Knowledge for the Modern Era : A Natural and Applied Science Perspective*. Boca Raton: CRC Press. doi: 10.1201/b21965

Kimbrough, L. (2021). *New Orchid Species From Guiana Highlands Named by Indigenous Group*. Available online at: https://news.mongabay.com/2021/01/new-orchid-species-from-guiana-highlands-named-by-indigenous-group/ (accessed July 30, 2021).

Kowal, E., Pearson, G., Peacock, C. S., Jamieson, S. E., and Blackwell, J. M. (2012). Genetic research and aboriginal and torres strait Islander Australians. *Bioethical Inq.* 9, 419–432. doi: 10.1007/s11673-012-9391-x

Kukutai, T., and Taylor, J. (2016). *Indigenous Data Sovereignty. Toward an Agenda. Canberra. Canberra: Australian National University Press*. doi: 10.22459/CAEPR38.11.2016

Lewis, J. E. (2020). *Indigenous Protocol and Artificial Intelligence Position Paper*. Available online at: https://spectrum.library.concordia.ca/986506/7/Indigenous_Protocol_and_AI_2020.pdf (accessed July 01, 2021).

Liggins, L., Hudson, M., and Anderson, J. (2021). Creating space for indigenous perspectives on access and benefit-sharing: encouraging researcher use of the local contexts notices. *Mol. Ecol.* 30, 2477–2482. doi: 10.1111/mec.15918

Mackenzie, I., and Davis, W. (2018). "Why lexical loss and culture death endanger science," in *The Oxford Handbook of Endangered Languages*, ed K. L. Rehg, and L. Campbell (Oxford: Oxford University Press). doi: 10.1093/oxfordhb/9780190610029.013.36

Mcalvay, A., Armstrong, C., Baker, J., Elk, L., and Bosco, S. E. (2021). Ethnobiology phase VI: decolonizing institutions, projects, and scholarship. *J. Ethnobiol.* 5, 170–191. doi: 10.2993/0278-0771-41.2.170

McGonigle, I. (2016). Patenting nature or protecting culture? Ethnopharmacology and indigenous intellectual property rights. *J. Law Biosci.* 3, 217–226. doi: 10.1093/jlb/lsw003

McWhirter, R., Nicol, D., and Savulescu, J. (2015). Genomics in research and health care with Aboriginal and Torres Strait Islander peoples. *Monash Bioethics Rev.* 33, 203–209. doi: 10.1007/s40592-015-0037-8

McWhirter, R. E., Mununggirritj, D., Marika, D., Dickinson, J., and Condon, J. R. (2012). Ethical genetic research in Indigenous communities: challenges and successful approaches. *Trends Mol. Med.* 18, 702–708. doi: 10.1016/j.molmed.2012.08.003

Micheli, M., Ponti, M., Craglia, M., and Berti, S. A. (2020). Emerging models of data governance in the age of datafication. *Big Data Soc.* 7:1–15. doi: 10.1177/2053951720948087

Moreton-Robinson, A. (2020). "Incommensurable sovereignties. Indigenous ontology matters," in *Routledge Handbook of Critical Indigenous Studies*, eds B. Hokowhitu, A. Moreton-Robinson, L. Tuhiwai-Smith, C. Andersen, and S. Larkin (Abingdon; New York, NY: Routledge, Taylor and Francis Group), 257–362. doi: 10.4324/9780429440229-23

Pain, E. (2016). *French Institute Agrees to Share Patent Benefits After Biopiracy Accusations*. Available online at: https://www.sciencemag.org/news/2016/02/french-institute-agrees-share-patent-benefits-after-biopiracy-accusations doi: 10.1126/science.aaf4036 (accessed May 01, 2021).

Parveen, N. (2021). Kew Gardens director hits back at claims it is 'growing woke'. *The Gaurdian*. Available online at: https://www.theguardian.com/science/2021/mar/18/kew-gardens-director-hits-back-at-claims-it-is-growing-woke (accessed April 10, 2021).

Phumthum, M., and Sadgrove, N. (2020). High-value plant species used for the treatment of "fever" by the Karen Hill Tribe People. *J. Antibiot.* 9:220. doi: 10.3390/antibiotics9050220

Pool, I. (2015). *Colonization and Development in New Zealand Between 1769 and 1900. The Seeds of Rangiatea*. Cham; Heidelberg; New York, NY; Dordrecht; London: Springer International Publishing. doi: 10.1007/978-3-319-16904-0

Rainie, S. C., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., et al. (2020). The CARE principles for indigenous data governance. *Data Sci. J.* 19, 1–12. doi: 10.5334/dsj-2020-043

Rainie, S. C., Rodriguez-Lonebear, D., and Martinez, A. (2017). Policy *Brief: Data Governance for Native Nation Rebuilding*. Tucson: Native Nations Institute.

Rainie, S. C., Rodriguez-Lonebear, D., and Martinez, A. (2019). Indigenous data governance: strategies from United States native nations. *Data Sci. J.* 18:31. doi: 10.5334/dsj-2019-031

Reardon, J., and TallBear, K. (2012). Your DNA is our history. Genomics, anthropology, and the construction of whiteness as property. *Curr. Anthropol.* 53, 233–245. doi: 10.1086/662629

Reijerkerk, D. (2020). UX design in online catalogs: practical issues with implementing traditional knowledge (TK) labels. First *Monday* 25. doi: 10.5210/fm.v25i8.10406

Rojas-Páez, G., and O'Brien, C. A. (2021). "Challenges of indigenous data sovereignty in Colombia's transitional setting," in *Indigenous Data Sovereignty*, eds M. Walter, T. K. Kukutai, S. R. Carroll, and D. Rodriguez-Lonebear (Abingdon; New York, NY: Routledge), 170–186.

Rowe, R. K., Bull, J. R., and Walker, J. D. (2021). "Indigenous self-determination and data governance in the Canadian policy context," in *Indigenous Data Sovereignty*, eds M. Walter, T. K. Kukutai, S. R. Carroll, and D. Rodriguez-Lonebear (Abingdon; New York, NY: Routledge), 81–98.

Sen, A. K. (2001). *Development as Freedom*. New York, NY: Alfred A. Knopf.

Smith, L. T. (2009). *Decolonizing Methodologies: Research and Indigenous Peoples*. Dunedin: University of Otago Press.

Tsosie, K. S., Yracheta, J. M., Kolopenuk, J. A., and Geary, J. (2021). We have "gifted" enough: indigenous genomic data sovereignty in precision medicine. *Am. J. Bioethics* 21, 72–75. doi: 10.1080/15265161.2021.1891347

Tsosie, R. (2021). "The legal and policy dimensions of Indigenous Data Sovereignty (IDS)," in *Indigenous Data Sovereignty and Policy*, eds M. Walter, T. K. Kukutai, S. R. Carroll, and D. Rodriguez-Lonebear (Abingdon; New York, NY: Routledge), 204–225. doi: 10.4324/9780429273957-14

Tuck, E. K., and Yang, W. (2012). Decolonization is not a metaphor. *Decolonization* 1, 1–40. Available online at: https://jps.library.utoronto.ca/index.php/des/article/view/18630/15554

United Nations (2006). *Who Are Indigenous Peoples?* Available online at: http://www.un.org/esa/socdev/unpfii/documents/5session_factsheet1.pdf (accessed May 15, 2021).

United Nations (2015). *About the Nagoya Protocol*. Available online at: https://www.cbd.int/abs/about/ (accessed April 07, 2021).

V.O. Patents and Trademarks (2019). *The Nagoya Protocol and Its Impact on Your Research*. Available online at: https://www.vo.eu/dossier/nagoya-protocol/ (accessed July 01, 2021).

van Geuns, J., and Brandusescu, A. (2020). *Shifting Power Through Data Governance*. Available online at: https://foundation.mozilla.org/en/data-futures-lab/data-for-empowerment/shifting-power-through-data-governance/

Vaughan, C. (2018). The language of cataloguing: deconstructing and decolonizing systems of organizationin libraries. *DJIM* 14. doi: 10.5931/djim.v14i0.7853

Vermeylen, S. (2007). Contextualizing 'fair' and 'equitable': the san's reflections on the hoodia benefit-sharing agreement. *Local Environ.* 12, 423–436. doi: 10.1080/13549830701495252

Vizenor, G. (1998). *Fugitive Poses: Native American Indian Scenes of Absence and Presence*. Lincoln, Nebraska: University of Nebraska Press.

Walker, J., Lovett, R., Kukutai, T., Jones, C., and Henry, D. (2017). Routinely collected indigenous health data: governance, ownership and the path to healing. *Lancet* 390, 2022–2023. doi: 10.1016/S0140-6736(17)32755-1

Walter, M., and Suina, M. (2019). Indigenous data, indigenous methodologies and indigenous data sovereignty. *Int. J. Soc. Res. Methodol.* 22, 233–243. doi: 10.1080/13645579.2018.1531228

Whitt, L. A. (1998). Biocolonialism and the commodification of knowledge. *Sci. Cult.* 7, 33–67. doi: 10.1080/09505439809526490

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Research Metrics and Analytics

ORIGINAL RESEARCH
published: 21 January 2022
doi: 10.3389/frma.2021.751734

RipetaScore: Measuring the Quality, Transparency, and Trustworthiness of a Scientific Work

Josh Q. Sumner[1,2]*, *Cynthia Hudson Vitale*[1,3] *and Leslie D. McIntosh*[1]

[1] Ripeta, LLC, St. Louis, MI, United States, [2] Washington University School of Medicine, Donald Danforth Plant Science Center, Washington University in St. Louis, St. Louis, MI, United States, [3] Association of Research Libraries, Washington, DC, United States

A wide array of existing metrics quantifies a scientific paper's prominence or the author's prestige. Many who use these metrics make assumptions that higher citation counts or more public attention must indicate more reliable, better quality science. While current metrics offer valuable insight into scientific publications, they are an inadequate proxy for measuring the quality, transparency, and trustworthiness of published research. Three essential elements to establishing trust in a work include: trust in the paper, trust in the author, and trust in the data. To address these elements in a systematic and automated way, we propose the ripetaScore as a direct measurement of a paper's research practices, professionalism, and reproducibility. Using a sample of our current corpus of academic papers, we demonstrate the ripetaScore's efficacy in determining the quality, transparency, and trustworthiness of an academic work. In this paper, we aim to provide a metric to evaluate scientific reporting quality in terms of transparency and trustworthiness of the research, professionalism, and reproducibility.

Keywords: research metrics, research quality, scientific indicators, reproducibility, research integrity

OPEN ACCESS

Edited by:
Stephen Pinfield,
The University of Sheffield,
United Kingdom

Reviewed by:
Houqiang Yu,
Sun Yat-sen University, China
Rajesh Piryani,
University of Tsukuba, Japan

***Correspondence:**
Josh Q. Sumner
josh.s@ripeta.com

Specialty section:
This article was submitted to Scholarly Communication, a section of the journal Frontiers in Research Metrics and Analytics

Received: *01 August 2021*
Accepted: *21 December 2021*
Published: *21 January 2022*

Citation:
Sumner JQ, Vitale CH and McIntosh LD (2022) RipetaScore: Measuring the Quality, Transparency, and Trustworthiness of a Scientific Work.
Front. Res. Metr. Anal. 6:751734.
doi: 10.3389/frma.2021.751734

INTRODUCTION

Misinformation, disinformation, and a general distrust in research and science by members of the general public has been the topic of many news stories in the last few years. This has cascaded into a series of funding policies, executive memos, and national and international task forces being established to increase research integrity and restore trust in scientific outcomes and policies that have resulted from those outcomes (United States White House, 2021). One critical factor in enhancing trust in research is through the increased transparency of reporting research (Moher et al., 2020). Yet, few tools and even fewer assessment metrics exist to evaluate the responsible reporting of research.

Existing research assessment metrics purport to measure a paper's quality or an author's clout. Often the fields of quality and prominence are lumped together, with authors that consistently have high citation counts being assumed to conduct the best research. Despite this conflation of quantity and prestige, there is plenty to be gained in various parts of the research world from metrics such as the H-index (Hirsch, 2005), RG score (ResearchGate, 2021), and Altmetric (Digital Science, 2018a) which all provide valuable insight for certain applications. Still, none of these measures serve as an appraisal of how trustworthy or reproducible a publication is based on the paper's content. Instead, these measures tend to track popularity or impact using publicly available information about the spread and influence of a paper or an author. While these are useful quantities, they should not be treated as direct measures of credibility, rigor, or quality of a publication. In light of the publishing

frenzy and heightened media attention on research through the COVID-19 pandemic there is a growing need for a user-oriented guide to understand the quality of a specific scientific paper. We propose our novel ripetaScore to address this need and serve as a direct measurement of the quality of a paper. In this paper, we aim to introduce this metric to evaluate scientific reporting quality in terms of transparency and trustworthiness through use of the three-part ripetaScore, measuring research, professionalism, and reproducibility.

The trust in reproducibility score is centered around the elements of a paper, which may facilitate a future researcher to most accurately replicate the study. Ripeta is a technology company that has developed tools and services to automatically assess the responsible reporting of research. Ripeta tools extract and show the responsible reporting of key scientific quality indicators within a scientific paper. Ripeta scans the paper's text for a selection of indicators spanning methodology, availability of data or code, analysis process, and analysis software citation. Though reproducibility is not guaranteed, if these variables are present in a paper then there is higher potential for reproducibility and are clear scientific quality indicators. Reproducibility is important for using research funds appropriately, providing the most reliable information possible, and for strengthening ethical scientific practices.

The trust in professionalism score aims to measure the legitimacy of the paper's authors and their thoroughness in reporting outside influences on their work. Two pieces currently play in determining the trust in professionalism: (1) Identifying if an author is who they say they are; and (2) Determining whether or not the authors are adhering to reporting standards put in place, such as being open about ethical declarations, conflicts of interest, and funding sources.

Finally, the trust in research component determines whether a paper meets general specifications for what "research" is. While this is normally obvious when reading a manuscript, it is important to factor into any analyses using automated methods. Some publishers do not make obvious distinctions in paper metadata between editorials or communications and research articles. For that reason, our score contains a "Trust in Research" component.

MATERIALS AND METHODS

The goal for the ripetaScore is to provide a meaningful high-level score summarizing the process of verifying the quality reporting of research and manuscript structure so experts can then more easily check the science. While other tools evaluate some scientific reporting practices, none are implemented as a complete summary of a publication like the ripetaScore. For the ripetaScore, we leveraged a locally developed corpus of publications to create the training dataset as well as personal experiences evaluating papers through the research lifecycle. The corpus of papers was selected based on several criteria and in response to internal needs for more data and external requests for analysis. Broadly, the training dataset was comprised of publications:

- Searchable by DOI
- Published either through peer-reviewed journal or hosted on a preprint server
- Recorded in the Dimensions database (Digital Science, 2018b)
- Licensed CC-BY or CC-0 or with access allowed via contractual permission.

Papers meeting the criteria above were collected and their text stored for use by leveraging various natural language processing (NLP) models. Ripeta has developed several NLP models, each tuned to a specific scientific quality indicator and based upon previous research conducted in developing the Ripeta reproducibility framework (McIntosh et al., 2017). Trained to read like humans, these NLP models scan articles for seed phrases and terms that indicate the presence of their respective quality indicator. These models were developed iteratively through a workflow utilizing human annotations and machine learning algorithms. The first stage of development involved manually annotating scientific papers for such aspects as data availability statements, explanations of statistical procedures and software, or study purposes to provide seed terms and known true positive results for a set of publications. These annotations were carried out in prodigy (Prodigy, 2021) for convenient integration with Ripeta's corpus of papers and with SpaCy (Honnibal and Montani, 2017). Next, a SpaCy model was built that used the extracted terms and examples to look for linguistic patterns to find similar phrases in the manually annotated papers as well as new publications. These steps were repeated many times until our NLP models could reliably return accurate results for new publications without any human guidance. Besides yielding very precise models this process has ensured that as the scientific landscape develops, Ripeta can retrain these NLP models to react to emerging challenges or to uphold more stringent standards. Now able to process the manuscripts, the NLP models extract text they recognize as matches for their respective criteria, based on those criteria's definitions.

From our total corpus of over a half a million articles, a subset of 12,000 CC-BY and CC-0 publications was selected to develop and test the ripetaScore. That sample included a variety of subject areas, funders, publishers, and journals. Publications that were not research articles were excluded as part of the scoring process.

As shown in **Figure 1**, once papers were identified, a persistent identifier such as a DOI or PMCID was submitted for each publication that should be collected by Ripeta via a POST request. We then validated the ID format and searched the Dimensions database for the identifier. If the publication exists in Dimensions, the DOI and other paper metadata were collected and a unique identifier to be used internally was assigned to the paper. CrossRef (2021) and Unpaywall (2021) were checked for additional paper metadata and license information is checked against Ripeta policy. If the license information did not meet policy, the paper was not stored in Ripeta's corpus and the harvesting process was terminated. Otherwise, the source document URLs were collected then the source document was parsed and stored for later use. Harvested papers were cleaned and sectioned using the papers' XML to allow for algorithm development.

FIGURE 1 | Manuscript collection workflow.

Next, papers were run through appropriate NLP models. These NLP models were created using Python version 3.7 (Van Rossum, 2007), SpaCy version 2.3.5 (spaCy, 2021), and Prodigy version 1.10.3 (Prodigy, 2021) at time of writing. The models have been trained to read similarly to how humans do. Model training started with a selection of seed terms, keywords, or phrases expected to be in a given statement, and proceeded from there to find more complex patterns among human-annotated training and testing papers. Seed terms and other model parameters were iterated upon until our desired accuracy was achieved, thereafter performance was measured continuously to ensure high accuracy. Results from the various NLP models were analyzed in R version 4.0.2 and/or Python version 3.7 to create reports, interactive dashboards, or other data summaries.

To best capture the trustworthiness of a publication the score needed both a professionalism and reproducibility component. Please see **Figure 2** for a breakdown of how scientific quality indicators were categorized into areas of research, professionalism, and reproducibility by Ripeta. Many of the scientific quality indicators were further subsectioned into more granular components based on the returned text response from the manuscript.

THE RIPETASCORE SCORING WORKFLOW

Components of the RipetaScore: Research, Professionalism, and Reproducibility

The ripetaScore combines three aspects of trust for a total of 30 points. See **Figure 3** for the ripetaScore scoring workflow. First, a paper is analyzed across our "Trust in Research" criteria to determine whether the paper is a research paper, which determines whether the paper will continue to be scored or not. Research articles are then evaluated for the presence of our reproducibility quality indicators and receive up to 20 points from those criteria. The last 10 points come from our trust in professionalism quality indicators.

The first step in scoring a manuscript is determining whether the document is a research paper or not. There are a variety of things that are not considered "real research" for one reason or another. An example criteria is that scientific research should have certain enumerated divisions separating the manuscript into recognizable sections. If key sectioning is not present, such as methods or conclusions, that would contribute to an indication that an article may not be scientific research. Additionally, the content of some work may flag it as something other than research. Titles and language outside of the normal scientific lexicon may simply be authors expressing themselves in their work, but in some cases it can be a useful tool in evaluating a paper. To make science better the community needs to be able to quickly and effectively determine what is research and what is not. For the purposes of scientific betterment, it does not matter if the statistics are reported well, if the publication was churned out of a paper mill. For such a publication, arguing minutiae of the methods misses the real issue and frames the discussion in an unproductive manner. With the goal of scientific betterment in mind, papers in our corpus were evaluated by our NLP algorithms and selections of papers were manually reviewed to examine correlations between quality of publications and different quality indicators Ripeta has developed. Removing these non-research articles from the corpus of scored papers increases Ripeta's efficiency through the rest of the scoring process and clearly differentiates research that is lacking in key quality indicators from submissions that simply are not research articles.

Papers determined to be real research are evaluated to gauge the trust in their reproducibility. Trust in reproducibility encompasses the majority of the ripetaScore for research articles. Since the widespread acknowledgment of a reproducibility crisis in science there has been much discussion about how to remedy these problems (Strech et al., 2020). The trust in reproducibility component of the ripetaScore is based on quality indicators designed to address this crisis. Primarily papers are evaluated with regards to their data/code sharing practices, thoroughness in explaining methods, and citing software. These indicators were picked due to their role in improving the likelihood of a study being well-enough documented as to be fully reproducible (Vickers, 2006; Baggerly, 2010). While these indicators are important, they cannot guarantee that a published finding is correct and fully reproducible. For example, currently we identifies whether data was shared and if so where the data was shared, but we are not yet making efforts to retrieve the data to assess its quality. Similarly, we look for evidence that the methods are sufficient to describe the work in detail, but we do not assess the methods for their appropriateness to a given field or study design. As we continue to develop new algorithms this component of the score may grow in scope.

Finally, papers are evaluated for trust in professionalism. Trust in professionalism is about components of research such as authorship and scientific etiquette. Some of the main contributors to this trust in professionalism aspect of the ripetaScore include whether ethical approval is properly cited, how corresponding authors can be reached, and whether funding sources are disclosed. Authorship concerns are another

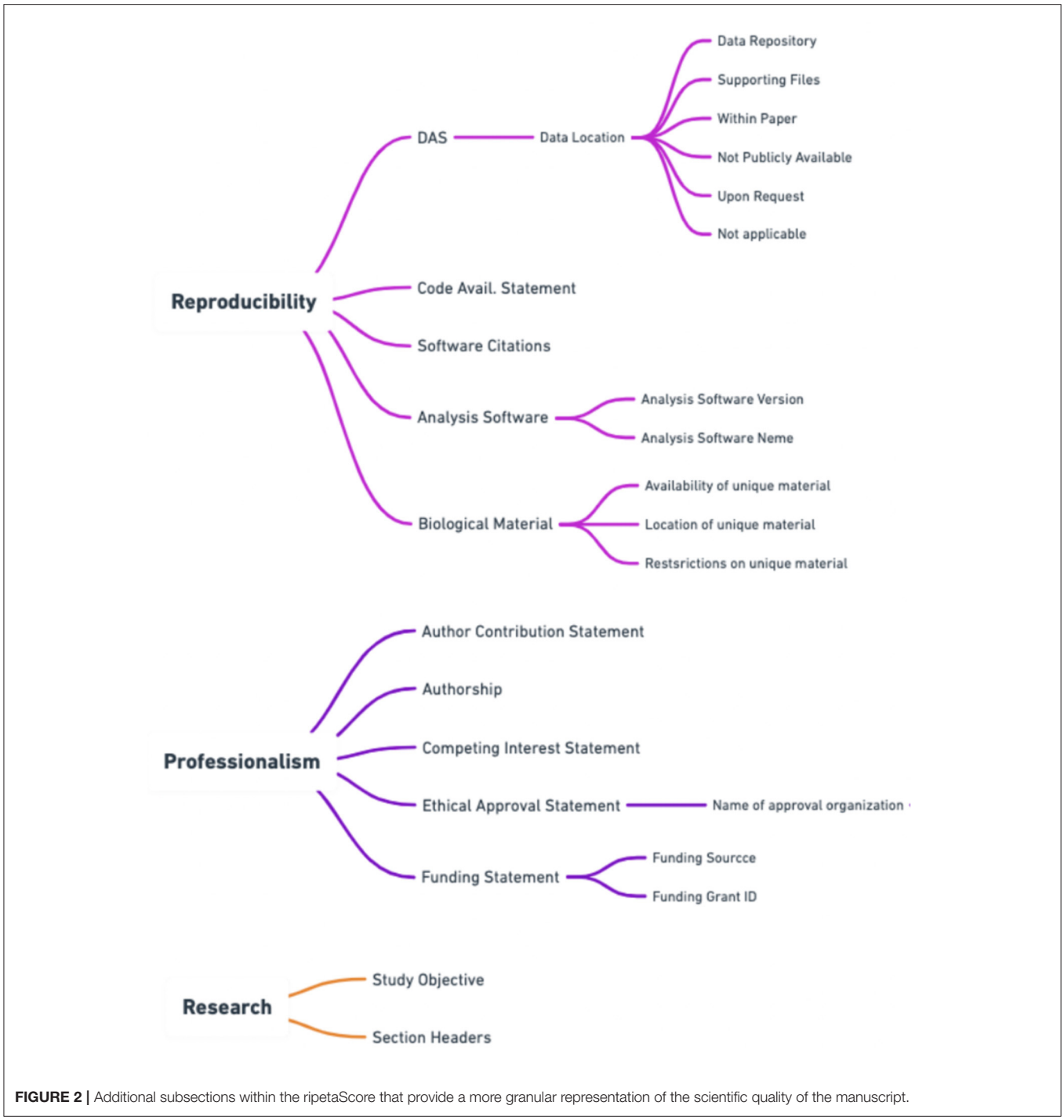


FIGURE 2 | Additional subsections within the ripetaScore that provide a more granular representation of the scientific quality of the manuscript.

factor being incorporated into trust in professionalism. Over the past decade there have been nearly 2,800 retractions due to authorship issues (Retraction Watch Database, 2021). The reasons for these retractions range from forged authorship and faked peer review to uncovered paper mills or author misconduct. Retractions are not only harmful to a journal's reputation, they also waste a tremendous amount of resources and can lead to negative repercussions in public policy as seen with several COVID-19 publications and preprints (Stern et al., 2014; Davido et al., 2020; Retracted COVID-19 papers, 2021). We used retrospective analysis of these retractions as well as exploratory analysis of preprints as they were submitted during the COVID-19 pandemic to develop a list of use cases for authorship trust, or trust in professionalism. Mainly we are interested in reducing the burden to journals and publishers by separating manuscripts where the authorship requires manual

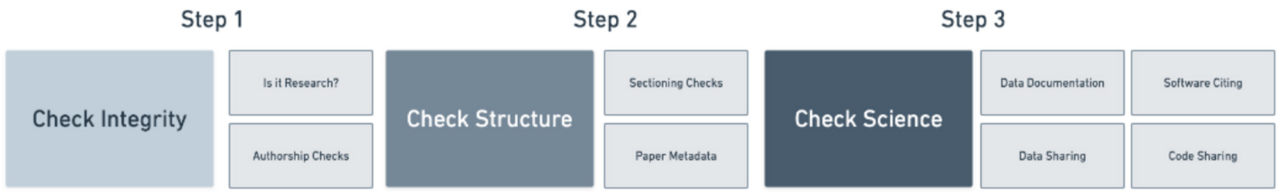


FIGURE 3 | RipetaScore scoring workflow.

TABLE 1 | RipetaScore breakdown across scientific quality indicators.

	Ripeta's scientific quality indicators			
	Research check (pass/fail)	**Professionalism (0–10)**	**Reproducibility (0–20)**	**Total score**
Perfect paper All indicators available	Pass	10	20	30
Research paper 1 Missing a few indicators but overall robust documentation and potential for reproducibility	Pass	6	17	23
Research paper 2 Includes only a minimum number of indicators such as an ethics statement and software used for analyses	Pass	3	7	10
Commentary Does not meet the requirements to be considered research	Fail	–	–	–

review from those manuscripts where professionalism can be established using existing data sources and the content of the paper. Through developing these use cases, collecting paper metadata, and analyzing the content of an academic paper our scoring criteria provide a useful metric for detecting potentially worrisome authorship issues. The journal putting forth an article also plays an important role in professionalism. Namely, suspected predatory journals and publishers should be monitored and their influence needs to appear in any evaluation of scientific trust. Finally, there are important parts of scientific etiquette that are evaluated as part of trust in professionalism. Widely accepted best practices such as stating funding sources or listing ethical approvals are important to the integrity of research and to professional scientific conduct. These best practices along with authorship checks inform our trust in professionalism score. In aggregate, trust in professionalism reflects on journal practices but for a single paper this reflects on individual trustworthiness of the work.

Together these three components of the ripetaScore make for an automated, holistic evaluation of the quality of a scientific text, which is useful to everyone from casual readers to journal editors looking to save time and money during costly procedures.

RESULTS

Evaluating the RipetaScore

There are many components of scientific quality and all of them should be taken into account when evaluating a publication. To help make these concepts more concrete we will go over a few examples of papers that score well or that score poorly using our ripetaScore. In order to avoid drawing unwanted attention to individual authors these papers are presented anonymously but with some context surrounding the field of research, journal, or time of publication.

The calculation for the ripetaScore is weighted across Ripeta's scientific quality indicators. The first component, the check for "Research" is based on whether or not the article is a true research paper or not. This is calculated as a simple pass (1) or fail (0) in the score creation. The next components of the ripetaScore—"Professionalism" and "Reproducibility" —assess how the paper performs across quality indicators. Each "Professionalism" quality indicator is assigned a numerical representation based on its importance for responsible reporting practices with a maximum score of 10. Finally, a "Reproducibility" score is calculated based on the numerical representation assigned for each quality indicator supporting the potential to reproduce the work with a maximum score of 20. The calculation is: ripetaScore = Research Check (pass/fail) * [Professionalism (0–10) + Reproducibility (0–20)].

The first example (**Table 1**) paper scores quite well with a ripetaScore of 23. This paper was published in 2019 in PLoS Computational Biology, well after open access practices have become commonplace and in a journal known for high standards of transparency. This paper's score reflects that it includes a clear study purpose, states the funding sources and their roles, and has an ethical statement (although the ethical statement does not list specific IRB approval). Looking at the reproducibility focused criteria this paper scores nearly a perfect 20. Data and

code, which overlap for this particular paper, are on Github with links provided, both of which factor heavily into the ripetaScore's reproducibility component. One place where this paper could score higher is in properly referencing MATLAB software that was used for analyses. Additionally, every author on the paper is on ORCID, although they do not all have their ORCID profiles listed in the paper's authorship information suggesting that they may not have been on ORCID when the paper was originally published. There is no evidence of any authorship issues regarding this paper. A high score such as this one is a code of confidence but would carry slightly different connotations depending on how the ripetaScore is implemented. On a preprint server it may serve to expedite the publication process and to provide readers of preprints with some baseline information about the writing. For a journal editor reviewing publications for acceptance the ripetaScore provides a quick way of assessing general quality which would both aide in most efficient use of expensive reviewer time and help serve as a check on journal policy compliance (such as whether or not submissions are adhering to a data sharing policy that is in place).

Our next example (**Table 1** "Research Paper 2") comes from a paper that scores in the lower middle of the distribution with a ripetaScore of 10 but which could score much higher with a few small improvements. This article was published in 2019 by Nature in the Scientific Reports journal. The authors of paper #2 fulfilled many of the components that go into the ripetaScore but several key indicators are still missing. For example, the paper does contain a data availability statement, but the data are listed as being available upon request from the authors, a method for data sharing that is dubiously helpful at best with under 20% of such statements enabling data to be found in many cases (Vines et al., 2013). While the authors make mention of using Graphpad Prism there is no citation to the specific version or way of examining any analysis code, both of which negatively impact the reproducibility of the work. Due to these and other factors this paper scores in the lower middle section of our possible distribution, along with the majority of other papers. A paper scoring in this range could indicate to preprint readers that they may want to pay extra attention to details about the analyses or check the author's publishing history depending on how the work did on each aspect of the ripetaScore. For a journal editor this type of score may suggest that the standard review process will be sufficient but that the paper should not be put on a fast-track to publication based on quality alone. As an author, a ripetaScore in this range is a sign that your manuscript may be lacking in some key factors aiding reproducibility, many of which are easy to add in and greatly improve the spirit of open science.

Finally, some writings such as our last example, measure up extremely poorly using the ripetaScore with a score of zero. This article was published in late 2020 by Openventio in the "Epidemiology—Open Journal." This publication contains none of the key quality indicators that currently factor into the ripetaScore and it raises questions of authorship trustworthiness. While this combination of missing information and abnormal authorship could be evidence of a new researcher unfamiliar with best practices it may also be evidence of predatory exploitation of open science, particularly considering the subject matter (Heimstadt, 2020). When dealing with a socially or politically charged subject matter it is also important to bear in mind that the ripetaScore does not take conclusions or press attention into account. Thus, good science can be conducted then co-opted by any number of agenda's without that misuse being tracked in the ripetaScore. As a reader either of preprints or published literature, a very low ripetaScore or outlandish media claims should or a score of zero should lead to careful consideration of the claims and whether they make sense given the rest of the scientific literature. From the perspective of a journal editor this sort of ripetaScore should raise red flags and suggest that extensive review and revisions may be necessary to get the paper to meet a journal's quality standards. Lastly, for an author this ripetaScore provides feedback that there are some very important components of scientific literature that are missing in your work

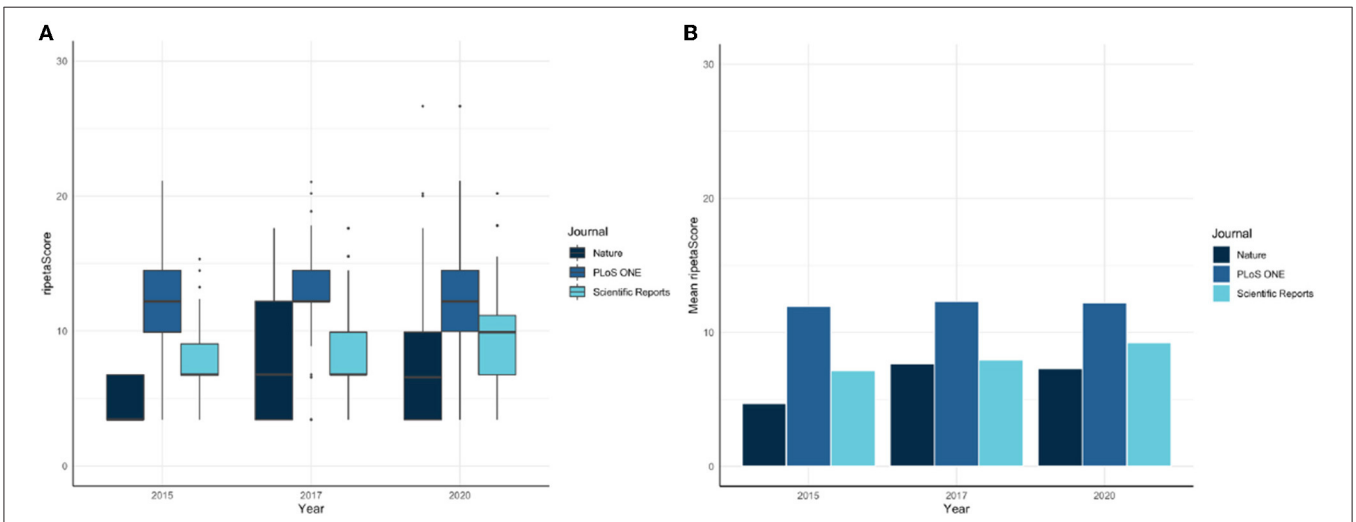


FIGURE 4 | (A,B) RipetaScore comparison across Nature, PloS One, and Scientific Reports.

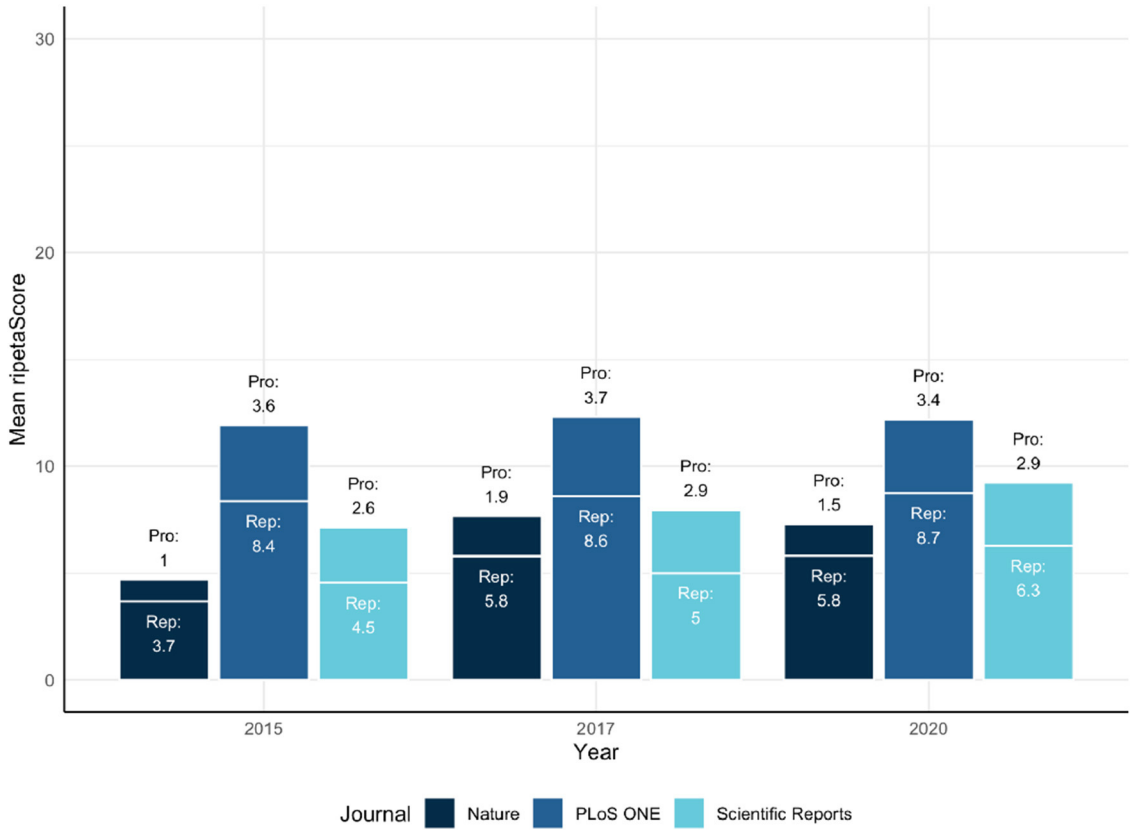


FIGURE 5 | Mean ripetaScore for Nature, Plos One, and Scientific Reports.

and that the discrepancy should be addressed. In such situations we encourage authors to reference journal policies and consider existing tools meant to aid in transparency and reproducibility. There are a variety of existing options to make sharing data, protocols, and code easier such as gigantum, github, protocols.io, codeocean, figshare, and jupyter. By implementing these existing options a low ripetaScore can often be greatly improved with relatively little added effort.

The ripetaScore is most useful when aggregated across time for a scholarly entity. As an example, **Figure 4** illustrates a comparison across Nature, PloS ONE, and Scientific Reports with the average ripetaScore of publications from 2015, 2017, and 2020 in those journals.

In these comparisons it is clear that PLoS ONE is leading the other journals on average, but has relatively constant ripetaScores over time (**Figure 5**). Nature and Scientific Reports on the other hand have lower averages but have shown improvement over time. Looking at the score as it's component pieces it becomes clear that most of the improvements being seen are coming from reproducibility practices improving while professionalism has stayed relatively similar in most cases.

Next Steps for the RipetaScore

While making the ripetaScore we realized that authorship was a critical component of scientific trust which could not be evaluated using our other textual variables aimed at reproducibility. Once we decided to explicitly include score aspects aimed at trustworthy authorship we investigated possible avenues of examining authorship such as network analysis and name disambiguation. As our authorship identification and evaluation processes become more refined the ripetaScore will only become more accurate and more helpful in establishing authorship trust. Similarly, as expectations or best practices for research reproducibility continue to develop, our ripetaScore will respond to those changes and to the size of our paper corpus growing.

CONCLUSION

Transparency, reproducibility, and responsible scientific practices are of utmost importance to the furtherance of research and scientific betterment. The ripetaScore provides an easily accessible metric to evaluate scientific reporting quality and trustworthiness toward evaluating these ends. The ripetaScore comprises three parts, trust in research, in reproducibility, and in professionalism. These categories and their contributions to the total ripetaScore have been developed through extensive testing of Ripeta's growing corpus of scientific papers. The ripetaScore is useful in evaluating single papers or conglomerated research and with continuing development of new NLP models,

new standards for reproducibility, and integration of more authorship checks the ripetaScore will only show more insights into research papers.

LIMITATIONS

All research metrics have limitations on their applicability and use. Additionally, for all metrics the inputs to conducting these calculations change over time and as more research is published. The ripetaScore has these same limitations. While we have built a score that is extensible as the number of scientific quality indicators increases, any metric should not be the sole basis for evaluation and assessment. Rather, any of this metric should be used in context with additional evaluative techniques.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://ripeta.figshare.com/.

AUTHOR CONTRIBUTIONS

LM and CV contributed to conception and design of the study. JS analyzed the data with input from LM. JS drafted the manuscript with editing from LM and CV. All authors contributed to manuscript revision, read, and approved the submitted version.

REFERENCES

Baggerly, K. (2010). Disclose all data in publications. *Nature* 467, 401. doi: 10.1038/467401b

Davido, B., Lansaman, L., Bessis, S., Lawrence, C., Alvarez, J. C., Mascitti, H., et al. (2020). Hydroxychloroquine plus azithromycin: a potential interest in reducing in-hospital morbidity due to COVID-19 pneumonia (HI-ZY-COVID)? *medRxiv Preprint*. doi: 10.1101/2020.05.05.20088757

Digital Science. (2018a). *Altmetric [Software]*. Available online at: https://www.altmetric.com/ (accessed July 31, 2021), under licence agreement.

Digital Science. (2018b). *Dimensions [Software]*. Available online at: https://app.dimensions.ai (accessed July 31, 2021), under licence agreement.

Heimstadt, M. (2020). *Between Fast Science and Fake News: Preprint Servers Are Political*. Available online at: https://blogs.lse.ac.uk/impactofsocialsciences/2020/04/03/between-fast-science-and-fake-news-preprint-servers-are-political/ (accessed May 20, 2021).

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. U.S.A.* 102, 16569–16572. doi: 10.1073/pnas.0507655102

Honnibal, M., and Montani, I. (2017). spaCy 2: Natural *Language Understanding* with Bloom *Embeddings, Convolutional Neural Networks and Incremental Parsing*.

McIntosh, L. D., Juehne, A., Vitale, C. R. H., Liu, X., Alcoser, R., Lukas, J. C., et al. (2017). Repeat: a framework to assess empirical reproducibility in biomedical research. *BMC Med. Res. Methodol.* 17, 143. doi: 10.1186/s12874-017-0377-6

Moher, D., Bouter, L., Kleinert, S., Glasziou, P., Sham, M. H., Barbour, V., et al. (2020). The Hong Kong Principles for assessing researchers: fostering research integrity. *PLoS Biol.* 18, e3000737. doi: 10.1371/journal.pbio.3000737

Prodigy. (2021). Available online at: https://prodi.gy/ (accessed July 31, 2021).

ResearchGate. (2021). *Research Gate Score*. Available online at: https://www.researchgate.net/ (accessed July 31, 2021).

Retracted COVID-19 papers. (2021). *Retracted Coronavirus COVID-19 Papers*. Available online at: https://retractionwatch.com/retracted-coronavirus-covid-19-papers/ (accessed July 31, 2021).

Retraction Watch Database. (2021). Available online at: http://retractiondatabase.org (accessed July 31, 2021).

spaCy. (2021). Available online at: https://spacy.io/ (accessed July 31, 2021).

Stern, A. M., Casadevall, A., Steen, R. G., and Fang, F. C. (2014). Financial costs and personal consequences of research misconduct resulting in retracted publications. *eLife* 3, e02956. doi: 10.7554/eLife.02956

Strech, D., Weissgerber, T., Dirnagl, U., and QUEST Group. (2020). Improving the trustworthiness, usefulness, and ethics of biomedical research through an innovative and comprehensive institutional initiative. *PLoS Biol.* 18, e3000576. doi: 10.1371/journal.pbio.3000576

United States White House. (2021). *Memorandum on Restoring Trust in Government Through Scientific Integrity and Evidence-Based Policymaking*. Available online at: https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/27/memorandum-on-restoring-trust-in-government-through-scientific-integrity-and-evidence-based-policymaking/ (accessed September 9, 2021).

Van Rossum, G. (2007). Python *Programming Language*.

Vickers, A. J. (2006). Whose data set is it anyway? Sharing raw data from randomized trials. *Trials* 7, 15. doi: 10.1186/1745-6215-7-15

Vines, T., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M.T., et al. (2013). The availability of research data declines rapidly with article age. *Curr. Biol.* 24, 94–97. doi: 10.1016/j.cub.2013.11.014

Conflict of Interest: JS, CV, and LM advise or work for Ripeta as employees.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Research Metrics and Analytics

REVIEW
published: 31 January 2022
doi: 10.3389/frma.2021.751553

The Scholarly Knowledge Ecosystem: Challenges and Opportunities for the Field of Information

Micah Altman[1]* *and Philip N. Cohen*[2]

[1] *Center for Research in Equitable and Open Scholarship, MIT Libraries, Massachusetts Institute of Technology, Cambridge, MA, United States,* [2] *Department of Sociology, University of Maryland, College Park, MD, United States*

The scholarly knowledge ecosystem presents an outstanding exemplar of the challenges of understanding, improving, and governing information ecosystems at scale. This article draws upon significant reports on aspects of the ecosystem to characterize the most important research challenges and promising potential approaches. The focus of this review article is the fundamental scientific research challenges related to developing a better understanding of the scholarly knowledge ecosystem. Across a range of disciplines, we identify reports that are conceived broadly, published recently, and written collectively. We extract the critical research questions, summarize these using quantitative text analysis, and use this quantitative analysis to inform a qualitative synthesis. Three broad themes emerge from this analysis: the need for multi-sectoral cooperation and coordination, for mixed methods analysis at multiple levels, and interdisciplinary collaboration. Further, we draw attention to an emerging consensus that scientific research in this area should by a set of core human values.

Keywords: scholarly communications, research ethics, scientometrics, open access, open science

OPEN ACCESS

Edited by:
Linda Suzanne O'Brien, Griffith University, Australia

Reviewed by:
Markus Stocker, Technische Informationsbibliothek (TIB), Germany
Ricardo Arencibia-Jorge, Universidad Nacional Autónoma de México, Mexico

***Correspondence:**
Micah Altman
micah_altman@alumni.brown.edu

Specialty section:
This article was submitted to Scholarly Communication, a section of the journal Frontiers in Research Metrics and Analytics

Received: *01 August 2021*
Accepted: *15 December 2021*
Published: *31 January 2022*

Citation:
Altman M and Cohen PN (2022) The Scholarly Knowledge Ecosystem: Challenges and Opportunities for the Field of Information. Front. Res. Metr. Anal. 6:751553. doi: 10.3389/frma.2021.751553

THE GROWING IMPORTANCE OF THE SCIENTIFIC INFORMATION ECOSYSTEM

> "The greatest obstacle to discovery is not ignorance—it is the illusion of knowledge."
> —Daniel J. Boorstin

Over the last two decades, the creation, discovery, and use of digital information objects have become increasingly important to all sectors of society. And concerns over global scientific information production, discovery, and use reached a fever-pitch in the COVID-19 pandemic, as the life-and-death need to generate and consume scientific information on an emergency basis raised issues ranging from cost and access to credibility.

Both policymakers and the public at large are making increasingly urgent demands to understand, improve, and govern the large-scale technical and human systems that drive digital information. The scholarly knowledge ecosystem[1] presents an outstanding exemplar of the challenges of understanding, improving, and governing information ecosystems at scale.

Scientific study of the scholarly knowledge ecosystem has been complicated by the fact that the topic is not the province of a specific field or discipline. Key research in this area is scattered across many fields and publication venues. This article integrates recent reports from multiple disciplines to characterize the most significant research problems—particularly grand challenges problems—that pose a barrier to the scientific understanding of the scholarly research ecosystem, and traces the contours of the approaches that are most broadly applicable across these grand challenges.[2]

The remainder of the article proceeds as follows: Characterizing the Scholarly Knowledge Ecosystem section describes our bibliographic review approach and identifies the most significant reports summarizing the scholarly knowledge ecosystem. Embedding Research Values section summarizes the growing importance of scientific information and the emerging recognition of an imperative to align the design and function of scholarly knowledge production and dissemination with societal values. Scholarly Knowledge Ecosystem Research Challenges section characterizes—impact scientific research problems selected from these reports. Commonalities Across the Recommended Solution Approaches to Core Scientific Questions section identifies the common shared elements of solution approaches to these scientific research problems. Finally, Summary section summarizes and comments on the opportunities and strategies for library and information science researchers to engage in new research configurations.

[1]Throughout this paper, we follow Altman et al. (2018) in using the terms "scholarship," "scholarly record," "evidence base," and "scholarly knowledge ecosystem" broadly. These denote (respectively), communities and methods of systematic inquiry aimed at contributing to new generalizable knowledge; all of the informational outputs of that system (including those outputs commonly referred to as "scholarly communications"); the domains of evidence that are used by these communities and methods to support knowledge claims (including quantitative measures, qualitative descriptions, and texts); and the set of stakeholders, laws, policies, economic markets, organizational designs, norms, technical infrastructure, and educational systems that strongly and directly affect the scholarly record and evidence base, and/or are strongly and directly affected by it (which encompasses the system of scholarly communication, and the processes generated by this system).

[2]In order to create a review that spanned multiple disciplines while maintaining concision and lasting relevance we deliberately concentrate the focus of the article in three respects: First, we focus on enduring research challenges rather than on shorter-lived research challenges (e.g., with a time horizon of under a decade). Second, we focus on fundamental challenges to scientific understanding (theorizing, inference, and measurement) rather than on cognate challenges to scholarly practice such as the developing of infrastructure, education, standardization of practice, and the mobilization and coordination of efforts within and across specific stakeholders. Third, we limit discussion of solutions to these problems to describing the contours of broadly applicable approaches—rather than recapitulate the plethora of domain and problem-specific approaches covered in the references cited.

CHARACTERIZING THE SCHOLARLY KNOWLEDGE ECOSYSTEM

The present and future of research—and scholarly communications—is "more." By some accounts, scientific publication output has doubled every 9 years, with one analysis stretching back to 1650 (Bornmann and Mutz, 2015). This growth has been accompanied by an increasing variety of scholarly outputs and dissemination channels, ranging from nanopublications to overlay journals to preprints to massive dynamic community databases.[3] As its volume has multiplied, we have also witnessed public controversies over the scholarly record and its application. These include intense scrutiny of climate change models (Björnberg et al., 2017), questions about the reliability of the entire field of forensic science (National Research Council, 2009), the recognition of social biases embedded in algorithms (Obermeyer et al., 2019; Sun et al., 2019), and the widespread replication failures across medical (Leek and Jager, 2017) and behavioral (Camerer et al., 2018) sciences.

The COVID-pandemic has recently provided a stress test for scholarly communication, exposing systemic issues of volume, speed, and reliability, as well as ethical concerns over access to research (Tavernier, 2020). In the face of the global crisis, the relatively slow pace of journal publication has spurred the publication of tens of thousands of preprints (Fraser et al., 2020), which in turn generated consternation over their veracity (Callaway, 2020) and the propriety of reporting on them in major news media (Tingley, 2020).

This controversy underscores calls from inside and outside the academy to reexamine, revamp, or entirely re-engineer the systems of scholarly knowledge creation, dissemination, and discovery. This challenge is critically important and fraught with unintended consequences. While calls for change reverberate with claims such as "taxpayer-funded research should be open," "peer review is broken," and "information wants to be free," the realities of scholarly knowledge creation and access are complex. Moreover, the ecosystem is under unprecedented stress due to technological acceleration, the disruption of information economies, and the divisive politics around "objective" knowledge. Understanding large information ecosystems in general and the scientific information ecosystem in particular, presents profound research challenges with huge potential societal and intellectual impacts. These challenges are a natural subject of study for the field of information science. As it turns out, however, much of the relevant research on scholarly knowledge ecosystems is spread across a spectrum of other scientific, engineering, design, and policy communities outside the field of information.

We aimed to present a review that is useful for researchers in the field of information in developing and refining research agendas and as a summary for regulators and funders of

[3]For prominent examples of nanopublication, overlay journals, preprint servers and massive dynamic community databases see (respectively) (Lintott et al., 2008; Groth et al., 2010; Bornmann and Leydesdorff, 2013; Fraser et al., 2020).

areas where research is most needed. To this end, we sought publications that met the following three criteria:

- *Broad*
  - Characterizing a broad set of theoretical, engineering, and design questions relevant to how people, systems, and environments create, access, use, curate, and sustain scholarly knowledge.
  - Covering multiple research topics within scholarly knowledge ecosystems.
  - Synthesizing multiple independent research findings.
- *Current*
  - Indicative of current trends in scholarship and scholarly communications.
  - Published within the last 5 years, with substantial coverage of recent research and events.
- *Collective*
  - Reflecting the viewpoint of a broad set of scholars.
  - Created, sponsored, or endorsed by major research funders or scholarly societies.
  - Or published in a highly visible peer-reviewed outlet.

To construct this review, we conducted systematic bibliographic searches across scholarly indices and preprint archives. This search was supplemented by forward- and backward- citation analysis of highly cited articles; and a systematic review of reports from disciplinary and academic societies. We then filtered publications to operationalize the selection goals described above. This selection process yielded the set of eight reports, listed in **Table 1**.

Collectively the reports in **Table 1** integrate perspectives from scores of experts, based on examination of over one thousand research publications and scholarship from over a dozen fields. In total, these reports span the primary research questions associated with understanding, governing, and reengineering the scholarly knowledge ecosystem.

To aid in identifying commonalities across these reports, we coded each report to identify important research questions, broad research areas (generally labeled as opportunities or challenges), and statements declaring core values or principles needed to guide research. We then constructed a database by extracting the statements, de-duplicating them (within work), standardizing formatting, and annotating them for context.[4] **Table 2** summarizes the number of unique coded statements in each category by type and work.

EMBEDDING RESEARCH VALUES

Science and scholarship have played a critical role in the dramatic changes in the human condition over the last three centuries. The scientific information ecosystem and its governance are now recognized as essential to how well science works and for whom. Without rehearsing a case for the value of science itself, we observe that the realization of such value is dependent on a system of scholarly knowledge communication.

In recent years we have seen that the system for disseminating scholarly communications (including evaluation, production, and distribution) is itself a massive undertaking, involving some of the most powerful economic and political actors in modern society. The values, implicit and explicit, embodied in that system of science practice and communication are vital to both the quality and quantity of its impact. If managing science information is essential to the potential positive effects of science, then the values that govern that ecosystem are essential building blocks toward that end. The reports illustrate how these values emerge through a counter-discourse, the contours of which are visible across fields.

All of the reports underscored[5] the importance of critical values and principles for successful governance of the scholarly ecosystem and for the goals and conduct of scientific research itself. [6] These values overlapped but were neither identical in labeling nor substance, as illustrated in **Table 3**.

Although the reports each tended to articulate core values using somewhat different terminology, many of these terms referred to the same general normative concepts. To characterize the similarities and differences across reports, we applied the 12-part taxonomy developed by AIETHICS in their analysis of ethics statements to each of the reports. As shown in **Figure 1**, these 12 categories were sufficient to match almost all of the core principles across reports, with two exceptions: several reports advocated for the value of organizational or institutional sustainability, as distinct from the environmental sustainability category; And the EAD referenced a number of principles, such as "competence" and (technical) "dependability" that generally referred to the value of sound engineering.

The value of *transparency* acts as a least-common-denominator across reports (as shown in **Figure 2**). However, transparency never appeared alone and was most often included with social equity and solidarity or inclusion. These values are distinct, and some, such as privacy and transparency, are in direct tension.

A dramatic expression of science's dependency on the values embedded in the knowledge ecosystem is the "reproducibility crisis" that has emerged at the interface of science practice and science communication (NASEM-BCBSS, 2019). Reproducibility is essentially a function of transparent scientific information management (Freese and King, 2018), contributing to meta-science, which furthers the values of equity and inclusion as

[4]For replication purposes, this database and the code for all figures and tables, are available through GitHub https://doi.org/10.7910/DVN/DJB8XI and will be archived in dataverse before publication, and this footnote will be updated to include a formal data citation.

[5]Almost all of the reports stated these values explicitly and argued for their necessity in the design and practice of science. The one exception is (Hardwicke et al., 2020)—which references core values and weaves them into the structure of its discussion—but does not argue explicitly for them.

[6]This set of ethical values constitute ethical principles *for* scientific information and its use. This should be distinguished from research programs such as (Fricker, 2007; Floridi, 2013) who propose ethics *of* information—rules that are inherently normative to information, e.g., Floridi's principle that "entropy ought to be prevented in the infosphere."

TABLE 1 | Key reports relevant to the scholarly knowledge ecosystem.

Year	Title	Description	Citation/References
2020	NDSA agenda for Digital Stewardship	Community/expert synthesis report conducted through *National Digital Stewardship Alliance*	(NDSA, 2020) (Digital stewardship)
2020	Calibrating the scientific ecosystem through meta-research	Scientific review published in *Annual Review of Statistics and Its Application*	(Hardwicke et al., 2020) (Meta research)
2019	The global landscape of AI ethics guidelines	PRISMA-review of AI ethics principles from 84 large organizations, societies, governments	(Jobin et al., 2019) (AI ethics)
2019	Reproducibility and replicability in science	Expert consensus report on reproducibility, convened by National Academies Committee on Reproducibility and Reliability	(NASEM-BCBSS, 2019) (reproducibility)
2018	A Grand Challenges-Based Research Agenda for Scholarly Communication and Information Science	Community-based synthesis report convened by MIT Center for Research on Equitable and Open Scholarship and Mellon Foundation	(Altman et al., 2018) (grand challenges)
2019	Open and Equitable Scholarly Communications: Creating a More Inclusive Future	Community-based synthesis report convened by Association of College and Research Libraries	(Maron et al., 2019) (SCHOLCOM)
2018	Open science by design: Realizing a vision for 21st-century research	Expert consensus report on open science convened by National Academies Board on Research Data and Information.	(NASEM–BRDI, 2018) (Open SCI)
2016	Ethically aligned design	Community/expert synthesis report convened by Institute of Electrical and Electronics Engineers	(Leek and Jager, 2017) (EAD)

much as those of interpretability and accountability. Open science enhances scientific reliability and human well-being by increasing access to both the process and the fruits of scientific research.

The values inherent in science practices also include the processes of assigning and rewarding value in research, which are themselves functions of science information management: this is the charge that those developing alternatives to bibliometric indicators should accept. Academic organizations determine the perceived value and impact of scholarly work by allocating attention and resources through promotion and tenure processes, collection decisions, and other recognition systems (Maron et al., 2019). As we have learned with economic growth or productivity measures, mechanistic indicators of success do not necessarily align with social and ethical values. Opaque expert and technical systems can undermine public trust unless the values inherent in their design are explicit and communicated clearly (IEEE Global Initiative et al., 2019).

When the academy delegates governance of the scholarly knowledge ecosystem to economic markets, scholarly communication tends toward economic concentration driven by the profit motives of monopolistic actors (e.g., large publishers) and centered within the global north (Larivière et al., 2015). The result has been an inversion of the potential for equity and democratization afforded by technology, leading instead to a system that is:

> "plagued by exclusion; inequity; inefficiency; elitism; increasing costs; lack of interoperability; absence of sustainability and/or durability; promotion of commercial rather than public interests; opacity rather than transparency; hoarding rather than sharing; and myriad barriers at individual and institutional levels to access and participation." (Altman et al., 2018, p. 5)

The imperative to bring the system under a different values regime requires an explicit and coordinated effort that is generated and expressed through research. The reports here reflect the increasing recognition that these values must also inform information research.

TABLE 2 | Extent of coded content.

Work	Research questions	Research areas	Values	Total
AI ETHICS	0	1	11	12
DIGITAL STEWARDSHIP	0	7	4	11
EAD	7	3	8	18
GRAND CHALLENGES	32	6	5	43
META RESEARCH	0	4	2	6
OPEN SCI	5	5	2	12
REPRODUCIBILITY	3	3	3	9
SCHOLCOM	0	18	3	21

Despite emerging as a "loose, feel-good concept instead of a rigorous framework" (Mehra and Rioux, 2016, p. 3), *social justice* in information science has grown into a core concern in the field. Social justice—"fairness, justness, and equity in behavior and treatment" (Maron et al., 2019, p. 34)—may be operationalized as an absence of pernicious discrimination or barriers to access and participation, or affirmatively as the extension of agency and opportunity to all groups in society. A dearth of diversity in the knowledge creation process (along the lines of nationality, race, disability, or gender) constrains the positive impact of advances in research and engineering (Lepore et al., 2020).

Many vital areas of the scientific evidence base, the legal record, and broader cultural heritage are at substantial risk of disappearing in the foreseeable future. Values of information *durability* must be incorporated into the design of the technical, economic, and legal systems governing information to avoid catastrophic loss (NDSA, 2020). The unequal exposure to the risk of such loss is itself a source of inequity. Durability is also

TABLE 3 | Core values and principles identified in each report.

Work	Values implicated
AI ETHICS	Transparency; justice, fairness, and equity; non-maleficence; responsibility; privacy; beneficence; freedom and autonomy; trust; sustainability; dignity; solidarity
DIGITAL STEWARDSHIP	Information ethics and privacy; trustworthiness; (organizational) sustainability; environmental sustainability
EAD	Universal human values (well-being); political self-determination and data agency; technical dependability; effectiveness; transparency; accountability; awareness of misuse; competence
GRAND CHALLENGES	Inclusion; openness; social equity; (organizational) sustainability; durability
META RESEARCH	Transparency; reproducibility
Open SCI	Openness; transparency
REPRODUCIBILITY	Science is a communal enterprise; science aims for refined degrees of confidence; scientific knowledge is durable and mutable
SCHOLCOM	Openness; inclusion; social equity

linked to the value of *sustainability*, applying both to impact the global environment (Jobin et al., 2019) and the durability of investments and infrastructure in the system, ensuring continued access and functioning across time and space (Maron et al., 2019).

As the information ecosystem expands to include everyone's personal data, the value of *data agency* has emerged to signify how individuals "ensure their dignity through some form of sovereignty, agency, symmetry, or control regarding their identity and personal data" (IEEE Global Initiative et al., 2019, p. 23). The scale and pervasiveness of information collection and use raises substantial and urgent theoretical, engineering, and design questions about how people, systems, and environments create, access, use, curate, and sustain information.

These questions further implicate the need for core values to govern information research and use: if individuals are to be more than objects in the system of knowledge communication, their interaction within that system requires not only access to information but also its *interpretability* beyond closed networks of researchers in narrow disciplines (Altman et al., 2018; NDSA, 2020). Interpretability of information is a prerequisite for the value of *accountability*, which is required to assess the impacts and values of scholarship. Accountability also depends on transparency, as the metrics for monitoring the workings of the scholarly knowledge ecosystem cannot perform their accountability functions unless the underlying information is produced and disseminated transparently.

SCHOLARLY KNOWLEDGE ECOSYSTEM RESEARCH CHALLENGES

Governing large information ecosystems presents a deep and broad set of challenges. Collectively, the reports we review touched on a broad spectrum of research areas—shown in **Table 4**. These research areas range from developing broad theories of epistemic justice (Altman et al., 2018) to specific questions about the success of university-campus strategies for rights-retention (Maron et al., 2019). This section focuses on those research areas representing grand challenges—areas with the potential for broad and lasting impact in the foreseeable future.

Altman et al. (2018) covered the broadest set of research areas. It identified six challenges for creating a scholarly knowledge ecosystem to globally extend the "true opportunities to discover, access, share, and create scholarly knowledge" in ways that are democratic in their processes—while creating knowledge that is durable as well as trustworthy. These imperatives shape the research problems we face. Such an ecosystem requires expanding *participation* beyond the global minority that dominates knowledge production and dissemination. It must broaden the *forms of knowledge* produced and controlled within the ecosystem, including, for example, oral traditions and other ways of knowing. The ecosystem must be built on a foundation of *integrity and trust*, which allows for the review and dissemination of growing quantities of information in an increasingly politicized climate. With the exponential expansion of scientific knowledge and digital media containing the traces of human life and behavior, problems of the *durability of knowledge*, and the inequities therein, are of growing importance. Opacity in the generation, interpretation, and use of scientific knowledge and data collection, and the complex algorithms that put them to use, deepens the challenge to maintain *individual agency* in the ecosystem. Problems of privacy, safety, and control, intersect with diverse norms regarding access and use of information. Finally, innovations and improvements to the ecosystem must incorporate *incentives for sustainability* so that they do not revert to less equitable or democratic processes.[7]

We draw from the frameworks of all the reports to identify several themes for information research. **Figure 3** highlights common themes using a term-cloud visualization summarizing research areas and research questions.[8] The figure shows the importance that the documents place on the values discussed above

[7] Any enumeration of grand challenge problems inevitably tends to the schematic. This ambitious map of challenges, intended to drive research priorities, has the benefit of reflecting the input of a diverse range of participants. Like the other reports in our review, Altman et al. (2018) lists many contributors (14) from among even more (37) workshop participants, and followed by a round of public commentary. Such collaboration will also be required to integrate responses, as these challenges intertwine at their boundaries. Thus, successful interventions to change the ecosystem at scale will require working in multiple, overlapping problem areas. Notwithstanding, these problems are capacious enough that any one of them could be studied separately and prioritized differently by different stakeholders.

[8] **Figure 3** is based on terms generated through skip n-gram analysis and ranked by their importance within each document relative to the entire corpus. Specifically, the figure uses TIF*DF (term frequency by inverse document frequency) to select and scale 2 by 1 skip-n-grams extracted from the entire corpus after minimal stop-word removal. This results in emphasizing pairs of words such as "transparency reproducibility" that do not appear in most documents overall, but appear together frequently within some documents.

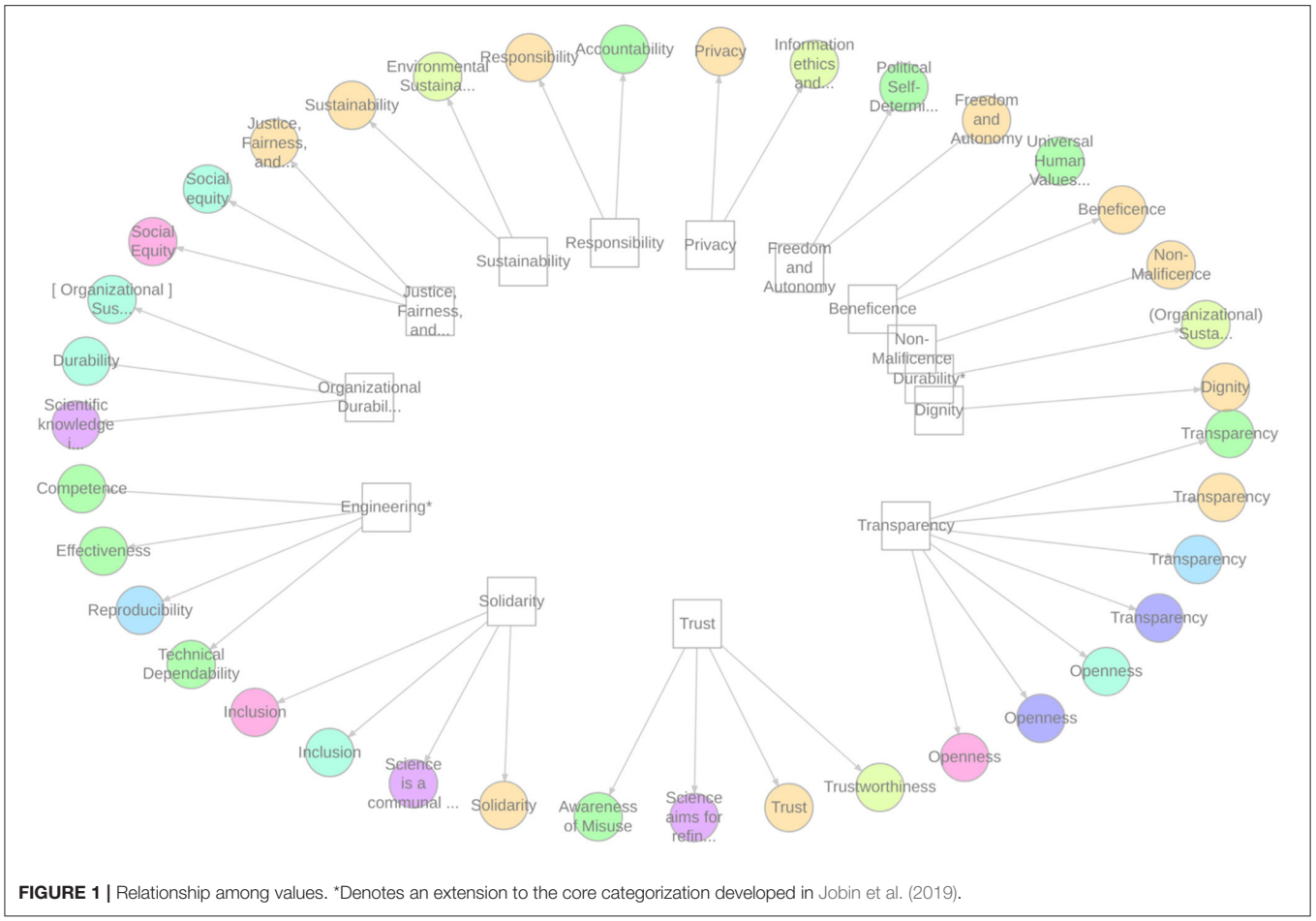


FIGURE 1 | Relationship among values. *Denotes an extension to the core categorization developed in Jobin et al. (2019).

and the importance of governance, technology, policy, norms, incentives, statistical reproducibility, transparency, and misuse.

For illustration, we focus on several exemplar proposals that reflect these themes. IEEE Global Initiative et al. (2019) asks how "the legal status of complex autonomous and intelligent systems" relates to questions of liability for the harms such systems might cause. This question represents a challenge for law and for ethical AI policy, as Jobin et al. (2019) outlined. Maron et al. (2019) raise questions about how cultural heritage communities limiting access to their knowledge while also making it accessible according to community standards poses additional problems for AI-using companies and the laws that might govern them. There is a complex interaction of stakeholders at the intersections of law, ethics, technology, and information science, and a research agenda to address these challenges will require interdisciplinary effort across institutional domains.

Consider the Grand Challenge's call for research into the determinants of engagement and participation in the scholarly knowledge ecosystem. Understanding those drivers requires consideration of a question raised by Maron et al. (2019) regarding the costs of labor required for open-source infrastructure projects, including the potentially inequitable distribution of unpaid labor in distributed collaborations. Similarly, NASEM–BRDI (2018) and NDSA (2020) delineate the basic and applied research necessary to develop both the institutional and technical infrastructure of stewardship, which would enable the goal of long-term durability of open access to knowledge. Finally, NASEM-BCBSS (2019) and Hardwicke et al. (2020) together characterize the range of research needed to systematically evaluate and improve the trustworthiness of scholarly and scientific communications.

The reports taken as a collection underscore the importance of these challenges and the potential impact that solving them can have far beyond the academy. For example, the NDSA 2020 report clarifies that resolving questions of predicting the long-term value of information and ensuring its durability and sustainability are critical for the scientific evidence-base and for preserving cultural heritage and maintaining the public record for historical government, and for legal purposes. Further, IEEE Global Initiative et al. (2019) and Jobin et al. (2019) demonstrate the ubiquitous need for research into effectively embedding ethical principles into information systems design and practice. Moreover, the IEEE report highlights

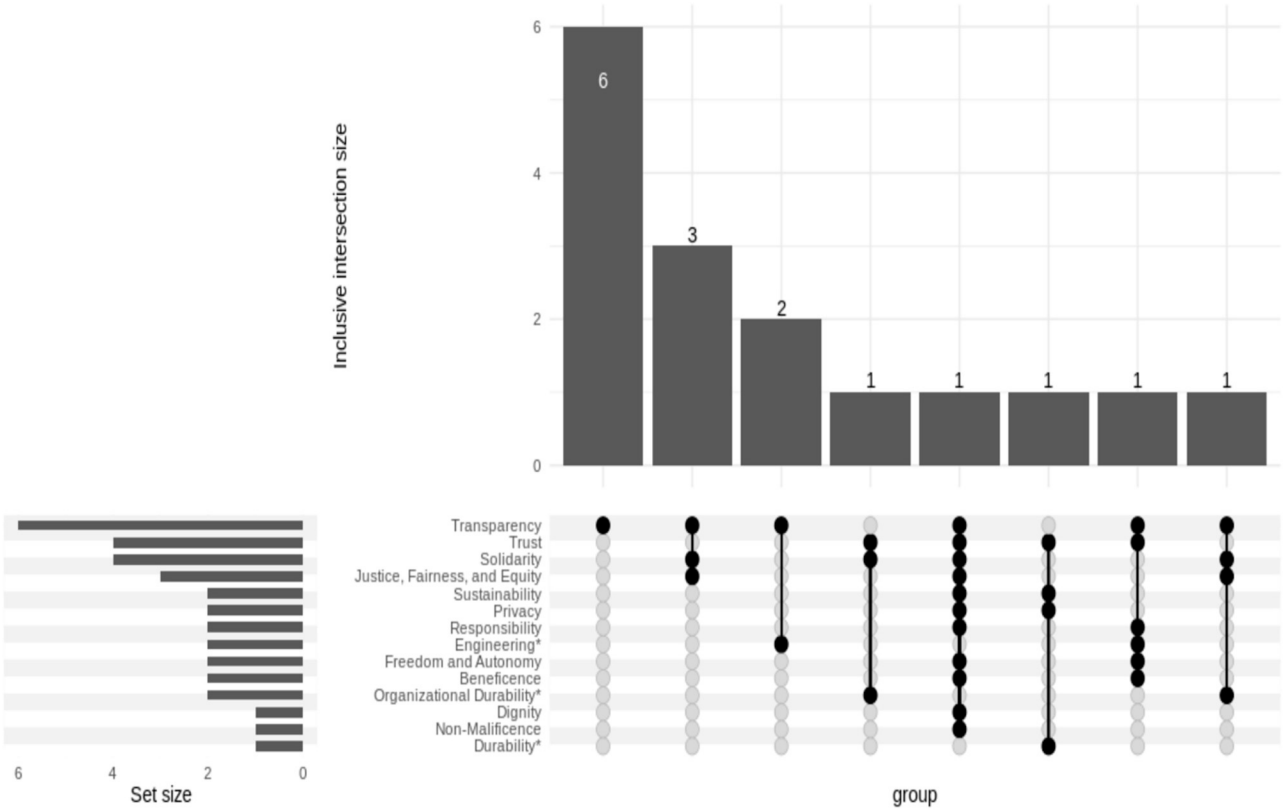


FIGURE 2 | Common core of values. *Denotes an extension to the core categorization developed in Jobin et al. (2019).

the need for trustworthy information systems in all sectors of society.

COMMONALITIES ACROSS THE RECOMMENDED SOLUTION APPROACHES TO CORE SCIENTIFIC QUESTIONS

The previous section demonstrates that strengthening scientific knowledge's epistemological reliability and social equity implicates a broad range of research questions. We argue that despite this breadth, three common themes emerge from the solution approaches in these reports: the need for multi-sectoral cooperation and coordination; the need for mixed methods analysis at multiple levels; and the need for interdisciplinary collaboration.

Cooperate Across Sectors to Intervene and Measure at Scale

As these reports reiterate, information increasingly "lives in the cloud."[9] Almost everyone who creates or uses information, scholars included, relies on information platforms at some point of the information lifecycle (e.g., search, access, publication). Further, researchers and scholars are generally neither the owners of, nor the most influential stakeholder in, the platforms that they use. Even niche platforms, such as online journal discovery systems designed specifically for dedicated scholarly use and used primarily by scholars, are often created and run by for-profit companies and (directly or indirectly) subsidized and constrained by government-sector funders (and non-profit research foundations).

A key implication of this change is that information researchers must develop the capacity to work within or through these platforms to understand information's effective properties, our interactions with these, the behaviors of information systems, and the implications of such properties, interactions, and behaviors for knowledge ecosystems. Moreover, scholars and scientists must be in dialogue with platform stakeholders to develop the basic research needed to embed human values into information platforms, to understand the needs of the practice, and to evaluate both.

Employ a Full Range of Methodologies Capable of Measuring Outcomes at Multiple Levels

Many of the most urgent and essential problems highlighted through this review require solutions at the ecosystem (macro-)

[9]Specifically, see NASEM-BIRDI (2018, chapters one and two), Lazer et al. (2009), and NDSA et al. (2020, sections 1.1, 4.1, and 5.2).

TABLE 4 | Research areas.

AI ETHICS	OPEN SCI
(Integrating, aligning, and implementing ethical principles through) public policy, technology governance, and research ethics	Costs and infrastructure
DIGITAL STEWARDSHIP	Disciplinary differences
Content preservation at scale	Lack of supportive culture, incentives, and training
Content selection at scale	Privacy, security, and proprietary barriers to sharing
Environmental sustainability of digital collections	Structure of scholarly communications
Information cost and value modeling	**REPRODUCIBILITY**
Stewardship at scale	Barriers in the culture of research
Strengthening the evidence base for digital preservation	Fraud and misconduct
Trust frameworks	Obsolescence of digital artifacts
EAD	**SCHOLCOM**
(Designing for) political self-determination and data agency	Assessing implicit and explicit bias
(Designing for) universal human values (well-being)	Building business models to support (mission-aligned) scholarly communications
(Designing for) technical dependability	Creating a broader scholarly communications workforce
GRAND CHALLENGES	Creating incentives for participation (in scholarly communications)
(Broadening) participation in the research community	Creating metrics built on value: expanding which values we measure
(Overcoming) restrictions on forms of knowledge	Designing systems that focus on users and audience
Incentives to sustain a (ethical) scholarly knowledge ecosystem	Determining the right scale and scope for (technological) infrastructure (that is organizationally sustainable)
Threats to durability of knowledge	Driving transformation within (academic) libraries
Threats to individual agency	Enacting effective strategies for revisiting copyright
Threats to integrity and trust	Encouraging technological innovation and ongoing development (in academic libraries)
META RESEARCH	Enhancing representations within academic libraries
Incentives and norms	Ensuring diversity of collections
Reproducibility	Facilitating access for those with disabilities
Statistical misuse	Intentionally limiting openness and knowledge sharing
Transparency	Investing in community-owned infrastructure
	Managing research data and enhancing discovery
	Retaining and protecting intellectual rights
	Understanding the costs of un(der)recognized and un(der)compensated labor (in scholarly communications)

level.[10] In other words, effective solutions must be implementable at scale and be self-sustaining once implemented. A key implication is that both alternative metrics and vastly greater access to quantitative data from and about the performance of the scholarly ecosystem are required.[11]

Engage Interdisciplinary Teams to Approach Ecosystem-Level Theory and Design Problems

Selecting, adapting, and employing methods capable of reliable ecosystem-level analysis will require drawing on the experience of multiple disciplines.[12] Successful approaches to ecosystem-level

[10]Ecosystem-level analysis and interventions are an explicit and central theme of Altman et al. (2018), NASEM–BRDI (2018), Maron et al. (2019) and Hardwicke et al. (2020) refer primarily to ecosystems implicitly in emphasizing throughout on the global impacts of and participation in interconnected networks of scholarship. NDSA (2020) explicitly addresses ecosystem issues through discussion of shared technical infrastructure and practices (see section 4.1) and implicitly through multi-organizational coordination to steward shared content and promote good practice.

[11]Metrics are a running theme of IEEE Global Initiative et al. (2019)—especially the ubiquitous need for open quantitative metrics of system effectiveness and impact, and the need for new (alternative) metrics to capture impacts of engineered systems on human well-being that are currently unmeasured. Altman et al. (2018, see, e.g., section 2) notes the severe limitations of the current evidence base and metrics for evaluating scholarship and the functioning of the scholarly ecosystem. Jobin et al. (2019, p. 389) also note the importance of establishing a public evidence base to evaluate and govern ethical AI use. Similarly, NDSA et al. (2020, section 5.2) emphasize the need to develop a shared evidence base to evaluate the state of information stewardship. Maron et al. (2019) call for new (alternative) metrics and systems of evaluation for scholarly output and contents as a central concern for the future of scholarship (p. 11–13, 16–20). NASEM–BRDI (2018), NASEM-BCBSS (2019), and Hardwicke et al. (2020) emphasize the urgent need for evidential transparency in order to evaluate individual outputs and systemic progress toward scientific openness and reliability—and emphasize broad sharing of data and software code.

[12]IEEE Global Initiative et al. (2019) emphasized interdisciplinary research and education as one of the three core approaches underpinning ethical engineering research and design (pp. 124–129), and identifying the need for interdisciplinary approaches in specific key areas (particularly engineering and well-being, affective computing, science education, and science policy). Altman et al. (2018) emphasize the need for interdisciplinarity to address grand challenge problems, arguing

FIGURE 3 | Research problems.

problems will, at minimum, require the exchange and translation of methods, tools, and findings between research communities. Moreover, many of the problems outlined above are inherently interdisciplinary and multisectoral—and successful solutions are likely to combine insights from theory, method, and practice from information- and computer- science, social- and behavioral- science, and from law and policy scholarship.

These three implications reflect broad areas of agreement across these reports regarding necessary conditions for approaching the fundamental scientific research questions about the scholarly knowledge ecosystem in general. Of course these three conditions are necessary, but far from sufficient—and only scratch the surface of what will be needed to restructure the ecosystem. Developing a comprehensive proposal for such a restructuring is a much larger project—even if the individual scientific questions we summarize above were to be substantially answered. For details on promising approaches to the individual areas summarized in **Table 4** see the respective reports, and especially (Altman et al., 2018; Hardwicke et al., 2020; NDSA, 2020).

Moreover, the development of a blueprint to effectively restructure the scholarly ecosystem will require addressing a range of issues. These include the development of effective science practices; effective advocacy in favor or an improved scholarly ecosystem; the development of model information policies and standards (e.g., with respect to licensing, or formats); the construction and operation of information infrastructure; effective education and training; and processes for allocating research funding in alignment with a better functioning ecosystem. Most of the reports discuss above recognize that these issues are critical to any future successful restructuring, and some—especially (Altman et al., 2018; NASEM–BRDI, 2018; Maron et al., 2019; NASEM-BCBSS, 2019)—suggest specific paths forward.

Although the function of this review is to characterize the core scientific challenges to understanding the scholarly ecosystem necessary for a restructuring. We note that there is a growing consensus, as reflected by these reports, around a number of operational principles, practices, and infrastructure that many believe necessary for a positive restructuring of the scholarly knowledge ecosystem. The most broadly recognized examples of these include the FAIR principles for scientific data management (Wilkinson et al., 2016), the TOP guidelines for journal transparency and openness (Nosek et al., 2015), arXiv and the increasingly robust infrastructure for preprints (McKiernan, 2000; Fraser et al., 2020), and the expansion of the infrastructure for data archiving, citation, and discovery (King, 2011; Cousijn et al., 2018; NASEM-BCBSS, 2019; NDSA, 2020) that has been critical to science for over 60 years.

that an improved scholarly knowledge ecosystem "will require exploring a set of interrelated anthropological, behavioral, computational, economic, legal, policy, organizational, sociological, and technological areas." Maron et al. (2019, sec. 1) call out the need for situating research in the practice and the engagement of those in the information professions. NDSA (2020) argue that solving problems or digital curation and preservation require transdisciplinary (sec. 2.5) approaches and drawing on research from a spectrum of disciplines, including computer science, engineering, and social sciences (sec 5). NASEM-BCBSS (2019) note that reproducibility in science is a problem that applies to all disciplines. While NASEM–BRDI (2018) and Hardwicke et al. (2020) both remark that the body of methods, training, and practices (e.g., meta-science, data science) required for achieving open and reproducible (respectively) science require approaches that are inherently inter-/cross-disciplinary.

SUMMARY

Since its inception, the field of information has been a leader in understanding how information is discovered, produced, and accessed. It is now critical to answer these questions as applied to the conduct of research and scholarship itself.

Over the last three decades, the information ecosystem has changed dramatically. The pace of information collection and dissemination has broadened; the forms of scientific information and systems for managing them have become more complex, and the stakeholders and participants in information production and use have vastly expanded. This expansion and acceleration have placed great stress on the system's reliability and heightened internal and external attention to inequities in participation and impact of scientific research and communication.

More recently, the practices and infrastructure for disseminating and curating scholarly knowledge have also begun to change. For example, infrastructure for sharing communications in progress (see, e.g., in preprints, or through alternative forms of publications) is now common in many fields, as is infrastructure to share data for replication and reuse.

These changes present challenges and opportunities for the field of information. While the field's traditional scope of study has broadened from a focus on individual people, specific technologies, and interactions with specific information objects (Marchionini, 2008) to a focus on more general information curation and interaction lifecycles, theories and methods for evaluating and designing information ecologies remain rare (Tang et al., 2021). Further, information research has yet to broadly incorporate approaches from other disciplines to conduct large-scale ecological evaluations or systematically engage with stakeholders in other sectors of society to design and implement broadly-used information platforms. Moreover, while there has been increased interest in the LIS field in social justice, the field lacks systematic frameworks for designing and evaluating systems to promote this value (Mehra and Rioux, 2016).

For scholarship to be epistemologically reliable, policy-relevant, and socially equitable, the systems for producing, disseminating, and sustaining scientific information must be re-theorized, reevaluated, and redesigned. Because of their broad and diverse disciplinary background, information researchers and schools could have an advantage in convening and catalyzing effective research. The field of information science can make outstanding contributions by thoughtful engagement in multidisciplinary, multisectoral, and multimethod research focused on values-aware approaches to information-ecology scale problems.

Thus reimagined and reengineered through interdisciplinary and multisectoral collaborations, the scientific information ecosystem can support enacting evidence-based change in service of human values. With such efforts, we could ameliorate many of the informational problems that are now pervasive in society: from search engine bias to fake news to improving the conditions of life in the global south.

AUTHOR CONTRIBUTIONS

The authors describe contributions to the paper using a standard taxonomy (Allen et al., 2014). Both authors collaborated in creating the first draft of the manuscript, primarily responsible for redrafting the manuscript in its current form, contributed to review and revision, contributed to the article's conception (including core ideas, analytical framework, and statement of research questions), and contributed to the project administration and to the writing process through direct writing, critical reviewing, and commentary. Both authors take equal responsibility for the article in its current form.

FUNDING

This research was supported by MIT Libraries Open Access Fund.

REFERENCES

Allen, L., Scott, J., Brand, A., Hlava, M., and Altman, M. (2014). Publishing: credit where credit is due. *Nature* 508, 312–313. doi: 10.1038/508312a

Altman, M., Bourg, C., Cohen, P. N., Choudhury, S., Henry, C., Kriegsman, S., et al. (2018). "A grand challenges-based research agenda for scholarly communication and information science," in *MIT Grand Challenge Participation Platform,* Cambridge, MA

Björnberg, K. E., Karlsson, M., Gilek, M., and Hansson, S. O. (2017). Climate and environmental science denial: a review of the scientific literature published in 1990–2015. *J. Clean. Prod.* 167, 229–241. doi: 10.1016/j.jclepro.2017.08.066

Bornmann, L., and Leydesdorff, L. (2013). The validation of (advanced) bibliometric indicators through peer assessments: a comparative study using data from InCites and F1000. *J. Inform.* 7, 286–291. doi: 10.1016/j.joi.2012.12.003

Bornmann, L., and Mutz, R. (2015). Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. *J. Assoc. Inform. Sci. Technol.* 66, 2215–2222. doi: 10.1002/asi.23329

Callaway, E. (2020). Will the pandemic permanently alter scientific publishing? *Nature* 582, 167–169. doi: 10.1038/d41586-020-01520-4

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Human Behav.* 2, 637–644. doi: 10.1038/s41562-018-0399-z

Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., et al. (2018). A data citation roadmap for scientific publishers. *Sci. Data* 5, 1–11. doi: 10.1038/sdata.2018.259

Floridi, L. (2013). *The Ethics of Information*. Oxford: Oxford University Press.

Fraser, N., Brierley, L., Dey, G., Polka, J. K., Pálfy, M., and Coates, J. A. (2020). Preprinting a pandemic: the role of preprints in the COVID-19 pandemic. *BioRxiv* 2020.05.22.111294. doi: 10.1101/2020.05.22.111294

Freese, J., and King, M. M. (2018). Institutionalizing transparency. *Socius* 4:2378023117739216. doi: 10.1177/2378023117739216

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Groth, P., Gibson, A., and Velterop, J. (2010). The anatomy of a nanopublication. *Inform. Serv. Use*. 30, 51–56. doi: 10.3233/ISU-2010-0613

Hardwicke, T. E., Serghiou, S., Janiaud, P., Danchev, V., Crüwell, S., Goodman, S. N., et al. (2020). Calibrating the scientific ecosystem through meta-research. *Annu. Rev. Stat. Appl.* 7, 11–37. doi: 10.1146/annurev-statistics-031219-041104

IEEE Global Initiative, Chatila, R., and Havens, J. C. (2019). "The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems," in *Robotics and Well-Being*, eds M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, and E. E. Kadar (Springer International Publishing), 11–16.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* 1, 389–399. doi: 10.1038/s42256-019-0088-2

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science* 331, 719–721. doi: 10.1126/science.1197872

Larivière, V., Haustein, S., and Mongeon, P. (2015). The oligopoly of academic publishers in the digital era. *PLoS ONE* 10:e0127502. doi: 10.1371/journal.pone.0127502

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., et al. (2009). Social science: computational social science. *Science* 323, 721–723. doi: 10.1126/science.1167742

Leek, J. T., and Jager, L. R. (2017). Is most published research really false? *Annu. Rev. Stat. Appl.* 4, 109–122. doi: 10.1146/annurev-statistics-060116-054104

Lepore, W., Hall, B. L., and Tandon, R. (2020). The Knowledge for Change Consortium: a decolonising approach to international collaboration in capacity-building in community-based participatory research. *Can. J. Dev. Stud.* 2020, 1–24. doi: 10.1080/02255189.2020.1838887

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., et al. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Month. Notices R. Astron. Soc.* 389, 1179–1189. doi: 10.1111/j.1365-2966.2008.13689.x

Marchionini, G. (2008). Human–information interaction research and development. *Library Inf. Sci. Res.* 30, 165–174. doi: 10.1016/j.lisr.2008.07.001

Maron, N., Kennison, R., Bracke, P., Hall, N., Gilman, I., Malenfant, K., et al. (2019). *Open and Equitable Scholarly Communications: Creating a More Inclusive Future*. Chicago, IL: Association of College and Research Libraries.

McKiernan, G. (2000). ArXiv. Org: The Los Alamos National Laboratory e-print server. *Int. J. Grey Literat.* 1, 127–138. doi: 10.1108/14666180010345564

Mehra, B., and Rioux, K. (eds.). (2016). *Progressive Community Action: Critical Theory and Social Justice in Library and Information Science*. Sacramento, CA: Library Juice Press.

NASEM-BCBSS, Board on Behavioral, Cognitive, and Sensory Sciences, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, Nuclear and Radiation Studies Board, Division on Earth and Life Studies, Board on Mathematical Sciences and Analytics, et al. (2019). *Reproducibility and Replicability in Science*. Washington, DC: National Academies Press.

NASEM–BRDI, Board on Research Data and Information, Policy and Global Affairs and National Academies of Sciences Engineering and Medicine (2018). *Open Science by Design: Realizing a Vision for 21st Century Research*. Washington, DC: National Academies Press, p. 25116

National Research Council (2009). *Strengthening Forensic Science in the United States: A Path Forward*. Washinton, DC: National Research Council.

NDSA, National Digitial Stewardship Alliance, and National Agenda Working Group (2020). *2020 NDSA Agenda*. Washinton, DC: NDSA.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., et al. (2015). Promoting an open research culture. *Science* 348, 1422–1425. doi: 10.1126/science.aab2374

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 447–453. doi: 10.1126/science.aax2342

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., et al. (2019). Mitigating gender bias in natural language processing: literature review. ArXiv:1906.08976 [Cs]. http://arxiv.org/abs/1906.08976

Tang, R., Mehra, B., Du, J. T., and Zhao, Y. (Chris). (2021). Framing a discussion on paradigm shift(s) in the field of information. *J. Assoc. Inf. Sci. Technol.* 72, 253–258. doi: 10.1002/asi.24404

Tavernier, W. (2020). COVID-19 demonstrates the value of open access: what happens next? *College Res. Libr. News* 81:226. doi: 10.5860/crln.81.5.226

Tingley, K. (2020, April 21). *Coronavirus Is Forcing Medical Research to Speed Up*. The New York Times. https://www.nytimes.com/2020/04/21/magazine/coronavirus-scientific-journals-research.html

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Research Metrics and Analytics

ORIGINAL RESEARCH
published: 28 March 2022
doi: 10.3389/frma.2022.779097

Measuring Research Information Citizenship Across ORCID Practice

*Simon J. Porter**

Digital Science, London, United Kingdom

Over the past 10 years, stakeholders across the scholarly communications community have invested significantly not only to increase the adoption of ORCID adoption by researchers, but also to build the broader infrastructures that are needed both to support ORCID and to benefit from it. These parallel efforts have fostered the emergence of a "research information citizenry" between researchers, publishers, funders, and institutions. This paper takes a scientometric approach to investigating how effectively ORCID roles and responsibilities within this citizenry have been adopted. Focusing specifically on researchers, publishers, and funders, ORCID behaviors are measured against the approximated research world represented by the Dimensions dataset.

Keywords: ORCID, scientometrics, Dimensions, research infrastructure, scholarly communications

OPEN ACCESS

Edited by:
Linda Suzanne O'Brien, Griffith University, Australia

Reviewed by:
Christopher Sean Burns, University of Kentucky, United States

***Correspondence:**
Simon J. Porter
s.porter@digital-science.com
orcid.org/0000-0002-6151-8423

Specialty section:
This article was submitted to Scholarly Communication, a section of the journal Frontiers in Research Metrics and Analytics

Received: *17 September 2021*
Accepted: *01 March 2022*
Published: *28 March 2022*

Citation:
Porter SJ (2022) Measuring Research Information Citizenship Across ORCID Practice. Front. Res. Metr. Anal. 7:779097. doi: 10.3389/frma.2022.779097

1. INTRODUCTION

In 2012, the founding members of ORCID Consortium asked the scholarly community to join them in imagining a new version of the scholarly record: One in which researchers were globally and uniquely identified (Haak et al., 2012). Although this sounds like a simple, incremental step, it was much more fundamental, at once solving information ambiguities and addressing issues of identity in an increasingly international community where trust in the validity of authorship is a critical currency. On a practical level, by attaching their ORCID iD to research objects such as publications, researchers would be able to reduce the administrative burden of communicating who they are and what they do across multiple domains including publishing, institutional assessment, research funding, and scholarly information discovery. Institutions within these domains would, in turn, gain greater strategic insight from the scholarly record not readily realizable within their own information silos.

Even at the beginning of the ORCID project, it was understood that to realize the benefits of ORCID, social and cultural change would be required in addition to technical change. Sustained community investment and collaboration around the development of ORCID and related infrastructures would need to be established amongst a disparate group of stakeholders with different drivers and motivations. All would need to be committed to developing and adopting new workflows and methods of information exchange. By connecting themselves to, and relying on each other, this newly networked community of researchers, institutions, funding bodies, publishers, and research service providers would establish the foundations of a new research information citizenship (Porter, 2016), defined by researcher agency, and distributed metadata stewardship.

When we speak about researcher agency we are specifically referring to the combination of a researcher-owned digital representation in the form of an ORCID record together with the set of interactions with the digital world through that representation. By implicitly establishing this as the de facto definition of researcher agency, ORCID upended passive assumptions about how a research identifier could be deployed. An ORCID iD was not just an identifier for a researcher that could be

added by anybody to a record, it simultaneously served as an identity through which a researcher could exert digital agency—this constituted a major step in establishing an infrastructural norm in the emergent digital research landscape. In addition to creating trusted assertions within publisher, funder, and other administrative workflows, a researcher could also gain access to research services. These services could include research facilities and collaboration tools, both at an administrative level of securing that access, as well as at the practical level of logging into a piece of equipment to perform their work. This merging of the worlds of describing research and conducting it created the possibility that trusted metadata about who was doing what research could be a byproduct of research itself.

Distributed metadata stewardship arises as a natural consequence of researcher agency in a complex ecosystem of stakeholders: It is simply not efficient, desirable, or practical to try to centralize permissions and the transaction logs associated with intrinsically distributed activities (typically those where researchers usually transact with any number of distributed stakeholders)[1]. As researchers engage across the activities in the research life cycle, different parts of the data contained in the ORCID registry of scholarly activities are made available to, and shared across, many different systems. In the case of publishing, a set of authenticated ORCID relationships between a set of researchers and a publication is collected at the time of submission or during the publication process. This distributed authentication is important as capturing these relationships at the point of submission is one of the few times when there is an incentive that can be applied in favor of data quality. A further consequence of distributed metadata stewardship is that the scholarly record itself becomes distributed, with different stakeholders holding differing levels of detail about each ORCID in their own systems. For instance, a publication identified by a DOI supplied by Crossref records the link between an ORCID iD and a specific author on the paper (Clark, 2020), whilst an ORCID record at orcid.org records the direct connection between a publication and researcher (ORCID, 2021b). While the distributed nature of this approach to data holding adds a level of privacy for an individual (since no one actor or system has access to all the information about that individual) there are also pitfalls - specifically, the opportunity for data loss or data inconsistency. Without a single source of truth or a set of mechanisms to homogenize data (such as a distributed data ledger), there is always the possibility of data ambiguity.

In addition to changes in workflows and responsibilities, global adoption of ORCID has also required a global network of change agents. Rather than being "top-down" initiatives led by governments, the mainstay of these activities has been done slowly with a mixture of bottom-up approaches and mid-level interventions. Country-led ORCID Consortia have organized to help researchers understand the benefits of maintaining their ORCID record. For their part, funders and publishers initially made ORCID optional in their grant and publication submission processes. In the last few years this has increasingly moved to requiring researchers to supply their ORCID as part of these processes (ORCID, 2016). Some countries have also chosen to act at a higher level and now mandate the use of ORCID iDs as part of their researcher reporting processes (Puuska, 2020).

While nudges and mandates can be powerful in gaining adoption, it is easier to achieve compliance if there is a tangible benefit to researchers and other stakeholders. In parallel with the development of the technology and compliance landscape, infrastructure has been developed to facilitate these benefits. Change has not been uniform, with funders and publishers moving toward ORCID support at different rates depending on their capacity to change their systems to conform with ORCID best practice (Mejias, 2020).

Almost a decade on and the success of ORCID can readily be measured by the number of participants actively engaged with ORCID. In 2018, UNESCO reports that the global researcher population had reached 8.9 Million FTE (UNESCO, 2021). At the end of 2018, there were 5.8 Million live ORCID registrations, 1.4M of whom had recorded at least one work (ORCID, 2021a). By July 2021, the number of ORCIDs that had an authenticated relationship with at least one scholarly work had increased to 3.9M. That these numbers are even within the same order of magnitude as the UNESCO figure is a significant achievement. While compelling, what these headline numbers do not indicate is the degree to which behavior and citizenship around ORCID research information has changed. Gaining an insight into the following questions would provide a better understanding of how far research citizenship now extends: Are researchers actively using their ORCID throughout the research process, or does the observed behavior simply reflect a compliance response to mandates? Beyond the ORCID registry itself, how are the responsibilities of distributed metadata stewardship being met? Does behavior differ between countries and disciplines? How far have publishers changed their practices to accommodate ORCID workflows? What is the quality of ORCID metadata outside of the ORCID registry (particularly in the Crossref registry)?

To address these questions, this paper takes a scientometric (Leydesdorff, 1995) approach and analyses ORCID behaviors with reference to the approximated world of researchers as embodied in the Dimensions database. Although not 100% accurate for all the reasons that ORCID was created in the first place, Dimensions provides a global set of algorithmically created researcher identities against which ORCID uptake can be measured. Additionally, Dimensions global coverage of publications and grants and the links between them provides a sufficient background dataset against which to conduct the analysis. Section 2 of this paper provides a description of the methodology used to link ORCID assertions from both Crossref and ORCID with the Dimensions dataset. Section 3 provides an analysis of the ORCID behaviors that we are able to observe. Finally, Section 4 reflects on the consequences of these findings.

[1]The idea of a centralized identity and authentication mechanism for academia is an alluring one. However, the idea that, at the current time, publishers, funders and academic institutions would all make themselves reliant on a centralized third-party is difficult to imagine. This is fundamentally counter-cultural in an academic context. Furthermore, we live in an era where the direction of movement in technology is toward the decentralization of trust or, more specifically, the distribution of trust across networks. Hence, it seems unlikely that centralization in this context would be a wise structural choice at this time.

2. METHODS

A previous analyses of ORCID uptake and usage used ORCID's public data file and publication level integration with metadata from Web of Science (Dasler et al., 2017). Comparative observations about researcher population by discipline and country were made by using reference researcher populations that were created programmatically from the Web of Science dataset by the Centre for Science & Technology Studies (CWTS) at Leiden University. In this investigation we have used the combined ORCID statements from both the ORCID (Blackburn et al., 2020) and Crossref public files (Clark, 2021) to examine ORCID-related behavior in publishing as a whole. This distinction is significant as it allows the flow of ORCID records between Crossref and the ORCID registry to be observed. Our approach also differs from the previous analysis in that we have integrated researcher identities from Dimensions, as well as matching records at the publication level. Integrating ORCID and Dimensions researcher identities allows for measures of individual record completeness to be approximated. Since the original study several large-scale initiatives have had an impact on ORCID adoption including funder and publisher mandates. Dimensions is well suited to provide insights into these developments as both funders and publishers are uniquely identified, allowing for publications to be easily aggregated and analyzed along these axis. The methodology for integrating the three datasets is described below.

2.1. Data Integration

To begin our analysis we needed to create a baseline dataset to facilitate comparisons. We generated this baseline by integrating Crossref and ORCID data with Dimensions (Hook et al., 2018) so that researchers without ORCID iDs could be identified. Inclusion of the Dimensions data allows us to access enhanced metadata concerning author affiliations, as well as publisher-level and funder-level information. Dimensions serves as a convenient intersection between the Crossref and ORCID datasets since the construction of Dimensions is predicated on persistent unique identifiers (PIDs) with information from orcid.org already matched back to Dimensions, and the Crossref data forming a key part of Dimensions' publications data spine (Visser et al., 2021). Data from the Crossref public file can be easily integrated at the author level, as the author level names largely match those in Dimensions. ORCID and Crossref data were loaded into Google BigQuery, allowing easy integration with Dimensions data, which is also available as a Google BigQuery dataset (Hook and Porter, 2021).

Table 1 provides a breakdown of the fields used in the analysis. Data was analyzed along the following axis: Publication, Researcher Affiliation (Country), Publisher, Funder, and Researcher Discipline. Of these, Publisher, Funder, and Researcher Discipline are described in further detail below.

TABLE 1 | Data sources and fields used in the analysis.

Source	Entity	Metadata analyzed
ORCID	Researcher	First name, last name, ORCID, date ORCID created
ORCID	Publication	DOI
Crossref	Publication	DOI
Crossref	Author.	First name, last name, ORICD iD
Dimensions	Researcher	Researcher_id, ORCID iD, and most recent institution & country affiliation
Dimensions	Author	First name, last name
Dimensions	Publisher	Publisher and journal references
Dimensions	Funder	Links between funders and researcher

2.2. Publications

Publication data from Crossref was integrated with publication data in Dimensions by matching on DOI, first name and surname. Reflecting the differences in metadata schemas, publications in the ORCID registry were not matched at the author level, but instead on ORCID iD and DOI. Publications without Crossref DOIs were also ignored as they did not have bearing on the practices measured in this investigation.

2.3. Researchers

Having matched Publications from Dimensions and Crossref at the author level, the corresponding researcher_id (Dimensions), and ORCID iD (Crossref) could be associated. This match could only be done after having addressed a data quality issue in the Crossref file (described below).

2.4. Affilitions

For this analysis the richer set of information around affiliation data in the ORCID record was not used in favor of Dimensions data that provided a consistent method of assigning institutional affiliation across researchers with and without ORCID iDs. The most recent affiliation for a researcher was calculated based on the affiliations associated with their most recent publications and grants.

2.5. Researcher Discipline

To facilitate the analysis of ORCID adoption by discipline, a researcher's discipline was defined as the two-digit Field of Research classification (Australian Bureau of Statistics, 2020) in which they most commonly publish (Porter, 2021). These classifications were assigned to publications using an NLP approach, ensuring consistency across a global dataset.

2.6. Data Quality

Before integrating Crossref and ORCID author assertions with Dimensions, Crossref records were first adjusted to address the phenomenon of "author shuffling." (Author shuffling is an effect where by an ORCID iD is assigned to the wrong author on a paper Baglioni et al., 2021). By joining raw Crossref records to Dimensions records, it was possible to estimate the size of the author shuffling problem by identifying papers where authors appeared to be collaborating with themselves. In the case of author shuffling, for an author with a reasonably sized publication history, an ORCID iD will be matched to more than one Dimensions researcher_id. For shuffled records, the research_id to which they are matched will be one of their

collaborators. Shuffled records can be identified when more than one of the researcher_ids that the ORCID has been associated with appears on the same paper. As **Figure 1** shows, the percentage of shuffled records in Crossref rose to just over .7% in 2018 before dropping slightly to approximately 0.5% in 2020. This is almost certainly an underestimate as this method only identifies cases where Dimensions has a researcher_id for the shuffled author as well as the actual author.

To increase the chances of finding all shuffled records so that they could be cleaned before matching, suspect author assertions were identified based on the following criteria:

1. The author appears to be collaborating with themselves (as above), or the match with Crossref results in more than one ORCID iD being assigned to a researcher_id;
2. The ORCID iD author matched identified by Dimensions disagrees with the author ORCID assertion in Crossref;
3. Dimensions does not have a researcher_id for the author ORCID assertion in Crossref.

For these records, a simple string matching algorithm using a Levenshtein Distance calculation was used to establish the most likely match between the name recorded in the ORCID record, and the names of the author on the paper (Cohen, 2015). If this approach returned the same match as Crossref with a ratio score of greater than or equal to 70%, the Crossref match was kept. If the name could be matched to another author on the paper with a confidence score of greater than 90%, then the ORCID author assertion was reassigned to that author. The difference in confidence cutoffs places a value on the Crossref assertion, as well as addresses a problem with the matching approach that gave very high scores to incorrectly matched authors with very short first names and surnames.

One drawback of the above approach to fixing shuffled records is that it creates a bias against some of the very use cases that ORCID was established to help solve, including changes in married names, names with few characters, and names with non-Latin characters. In addition, some authors used the native version of their name in their ORCID record, but published with the anglicized version. To help reduce the number of times these instances were rejected due to low name matching scores, author name ORCID matches that could be found across publications from multiple publishers were also accepted as true.

Using the combination of these methods, 1.7% connections asserted in the Crossref data were removed, and 0.5% reassigned to other authors. That 1.2% of connections were not easily recoverable is illustrative of the difficulty of name matching based on strings.

3. RESULTS

3.1. ORCID Adoption and Engagement

With the integrated ORCID, Crossref, and Dimensions datasources, we are able to measure ORCID adoption as the percentage of researchers in a given year who have at least one publication with a DOI linked to their ORCID iD either in ORCID directly or identified within the Crossref file. ORCID record completeness was also approximated by comparing the number of publications linked to an ORCID iD vs. the number of publications linked to the Dimensions researcher_id against which the ORCID identifier was matched. As defined, ORCID *adoption* is intended as a measure of active usage, whereas ORCID record *completeness* is a proxy for engagement.

Researchers with only a few publications are difficult to identify algorithmically as there are few data points to base a decision on. To increase the chances of Dimensions accurately identifying researchers, researchers with less than 5 years publishing history have been excluded from the analysis. Completeness calculations have also been restricted to publications between 2015 and 2019.

We argue that completeness can be thought of as a proxy for engagement, since a researcher needs to take responsibility for their own record in order for it to be maintained accurately. Firstly, they must set up their ORCID to receive automatic updates from Crossref, and secondly, they must update their own record with ORCID publication assertions not captured during publisher submission. By including publications in the Crossref record, this measure of completeness is able to include ORCID assertions are not present in a researcher's public record. ORCID assertions that have been made private by the researcher and are not included in the Crossref record have not been included in the analysis.

3.1.1. ORCID Adoption and Engagement by Country

Breaking measures of ORCID adoption and completeness down, by Country (**Figure 2**), it is clear that just as factors other than economic wealth strongly influence the scientific wealth of nations (Allik et al., 2020), local research environments significantly influence ORCID researcher engagement. Looking at the years between 2015 and 2019, Portugal ranks most highly in both Adoption (67%), and Engagement (70%). Poland, Australia, Denmark, Columbia and South Africa and New Zealand then follow with adoption levels between 50 and 60%. Of the countries with an identified researcher pool of $>$ 100,000, the more established and larger scale research economies, Italy, Spain, and the United Kingdom have adoption rates in the region just below or just above 40%. However, not all the established research economies show the same level of engagement for a cadre of different reasons: The United States, China, and Japan are notable for their relatively low adoption and engagement rates compared to countries in the same World Bank income bands. In the case of the United States, this is likely to be due to the lack of centralized, government-led research evaluation and levers associated with block funding the other countries such as those mentioned have available. Japan has adopted its own system of researcher identification with the researchmap.jp system, which stands apart from all other global systems. China, while moving quickly, is simply at an earlier stage of engagement with globalized research infrastructure and has unique challenges in terms of name disambiguation.

Countries with high engagement have also demonstrated concerted enrolment efforts. These efforts can be detected in the publication record by looking for ORCID iDs that are used in publications between the time they were created and the end of the next full publication year (**Figure 3**). Using this

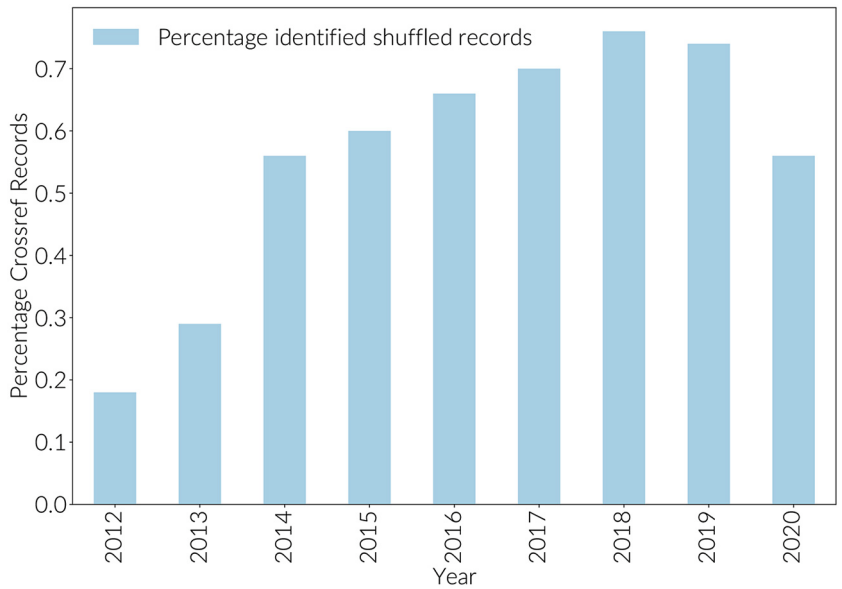


FIGURE 1 | Percentage identified shuffled ORCID assertions.

methodology, it is possible to observe that Portugal started early with a concentrated effort in 2012, and 2013 at the launch of the ORCID initiative, with Spain following over 2013 and 2014, Italy and Denmark in 2014-2015. Both Australia and the United Kingdom showed a sustained engagement at or slightly below 10% between 2013 and 2016. Poland is distinct in initiating renewed engagement activities in 2016.

Countries with low engagement show a different pattern (**Figure 4**). Since 2016, there has been a steep increase in ORCID iD assertions that are present in Crossref, but are not displayed in a researcher's own ORCID record. This is particularly prominent with Chinese authors where 50% of researchers in 2019 do not have any 2019 statements from Crossref that have made it back to their public ORICD record. For United States authors, this value 40%, compared to 10% for Portugal, and just above 20% for Italy and Australia. This result is despite the fact that there is an established workflow to push ORCID assertions back from Crossref to ORCID, and that all researchers are required to do is to provide consent in response to an email (Brown et al., 2016). At least two scenarios might explain this behavior with the strength of this effect varying by country:

1. An increasing number of researchers are registering for an ORCID iD because they are encouraged to during early career studies or because they need one to engage in certain formal processes within their country. Motivated from a position of compliance, these researchers are not sufficiently engaged to go further and keep their ORCID record up to date either by entering in details directly, or by authorizing the systems that they engage with to update their record on their behalf (such as the Crossref auto update functionality.) (Mejias, 2020).
2. An increasing number of researchers are choosing to keep their record private due to growing privacy concerns associated with digital existence as a whole.

The first scenario is concerning. It suggests that a growing number of researchers will not be able to use their ORCID iD as a tool to reduce academic burden. These researchers will likely be frustrated when the act of supplying their ORCID iD in a funder workflow does not result in their record being populated. This scenario is reasonably likely. In 2017, after the initial release of the Crossref auto-update functionality, only 50% of researchers were reported as choosing to respond to the email from Crossref offering to auto update their ORCID record when new publications were detected (Meadows and Haak, 2017). For some countries, it does not appear as if this number has significantly improved since this time.

The second scenario, although not necessarily preventing any ORCID use cases, would indicate an increasing desire by researchers not to be 'known' by their ORCID iD, and perhaps a lack of buy-in to open identifier infrastructure. Both scenarios would be regional examples of less than enthusiastic research information citizens.

Part of difference between country cultures can be explained by the interventions local funding agencies have made in integrating ORCID iDs into their processes. Funding agencies can impact ORCID behavior by requiring researchers to have an ORCID (adoption,) as well as by driving engagement by making it easy for researchers to use information from their ORCID records in their publications, or implying a strong preference for complete ORCID records. Beyond publication workflows, funders will also play an increasing role in linking ORICD iDs to open public records of grants (ORCID Funder Working Group, 2019) creating similar data reuse patterns to publications.

Figure 5 shows the top 60 funders by the number of researchers with ORCID iDs that they have funded between 2015 and 2019. Across these 60 funders, a much higher ORCID adoption rate can be observed for funded researchers than

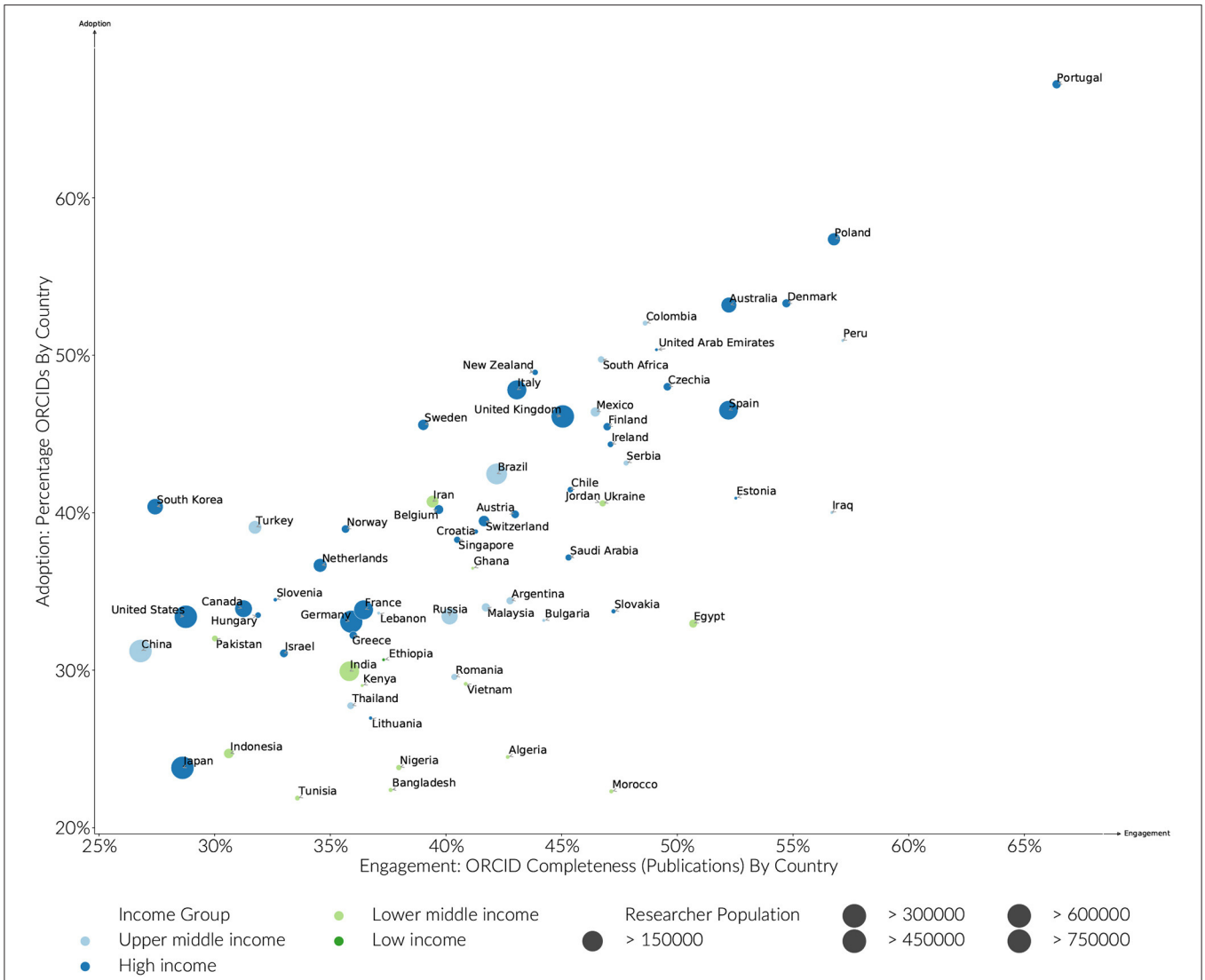


FIGURE 2 | Estimated ORCID adoption and engagement by country. Active researchers in the analysis must have (A) published between 2015 and 2019, (B) have a publication history of greater than 5 years, and (C) published more than 5 papers.

compared with country averages. This is to be expected to some degree, as there will be a greater overlap between researchers that receive funding, and researchers required to have an ORCID iD as part of publisher ORCID policies. A similar shift is not observed in the engagement rates by funder when compared to overall country rates.

Even with the overall increase in ORCID adoption rates, distinct funder patterns can be observed. The United Kingdom, Finland, Portugal, Australia, Austria and Czechia have very high adoption rates (between 80% and 90%). Many of these funders are associated with funder ORCID policies that either mandate, or strongly recommend the use of ORCID iDs in funder submissions. That engagement rates for these funders do not differ significantly from country norms, suggests an impact beyond just those who were funded to applicants and the broader community. An underlying information systems capacity for a country to accept a funder mandate may also be in play, with the United Kingdom, Australia, Finland and Portugal, and Czechia all having strong research reporting practices at the country and institutional level. High levels of research engagement implies a high level of ORCID record maintenance. Countries with a mature network of Institutional Current Research Information Systems will be better supported with these maintenance activities.

A separate band of funders including funders from the United States, Canada, Germany, Russia and Israel sees adoption rates between 60 and 80%. Within this second band, where identifiable in funder policies listed by ORCID (ORCID, 2020), ORICD integration funder appears to be more technical and optional rather than policy driven. Other funders within this band have more recently launched ORCID initiatives, the effects of which would not be seen in the analyzed period.

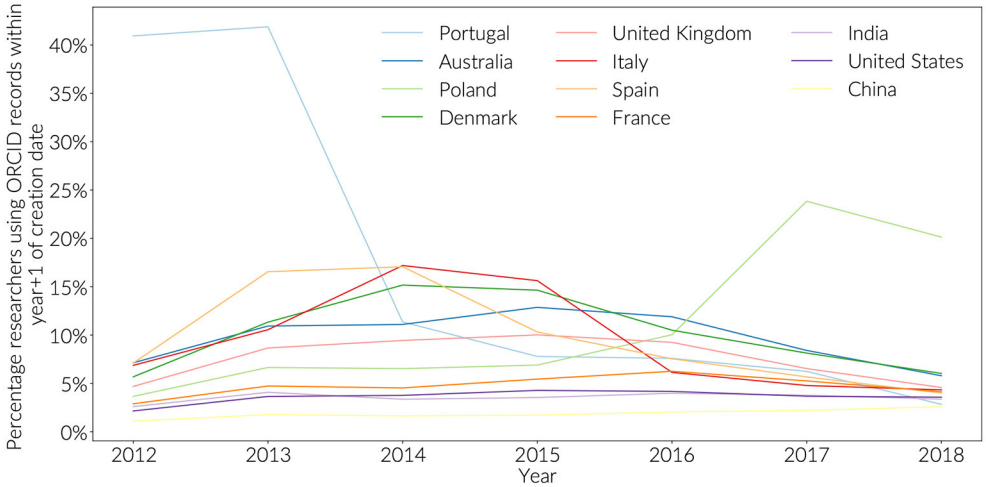


FIGURE 3 | New ORCID registrations by year (+1). Totals are not cumulative, showing early peaks in adoption.

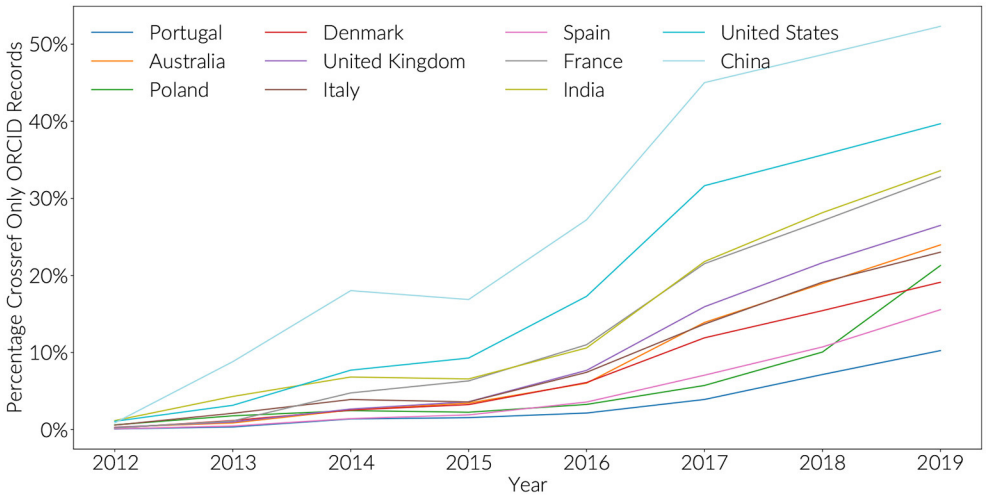


FIGURE 4 | Percentage of ORCID records with only Crossref assertions by year.

3.2. ORCID Adoption by Research Category

Overall, funder adoption and engagement rates are clustered more by country than they are by discipline, however some discipline effects can still be observed. Medically focused funders in particular have lower engagement rates on average when compared to other funders in the same country. These differences in discipline are also borne out more generally. As shown in **Figure 6**, ORCID adoption by discipline ranges 25–45%, and engagement from 30 to 50%. Earth Sciences and Chemical Sciences have both high adoption and engagement rates. Humanities research areas are distinguished by having lower adoption levels, but higher engagement levels. The large difference between adoption and engagement levels for these fields is partly explained by the articles in these fields having fewer authors per paper, and therefore fewer middle authors that are unlikely to receive ORCIDs given current publishing workflows. The average number of authors per paper does not explain the disparity in engagement across all disciplines, however. For instance, researchers in Medical and Health Sciences have a much lower engagement rate when compared to the relatively high adoption and engagement rates of disciplines with a similar average number of authors per paper such as Chemical or Biological Sciences (**Figure 7**).

As disciplines cross different country and funder environments, a high engagement and adoption level by discipline suggests that there are pockets of research practice that are closer to normalizing the use of ORCID for all authors.

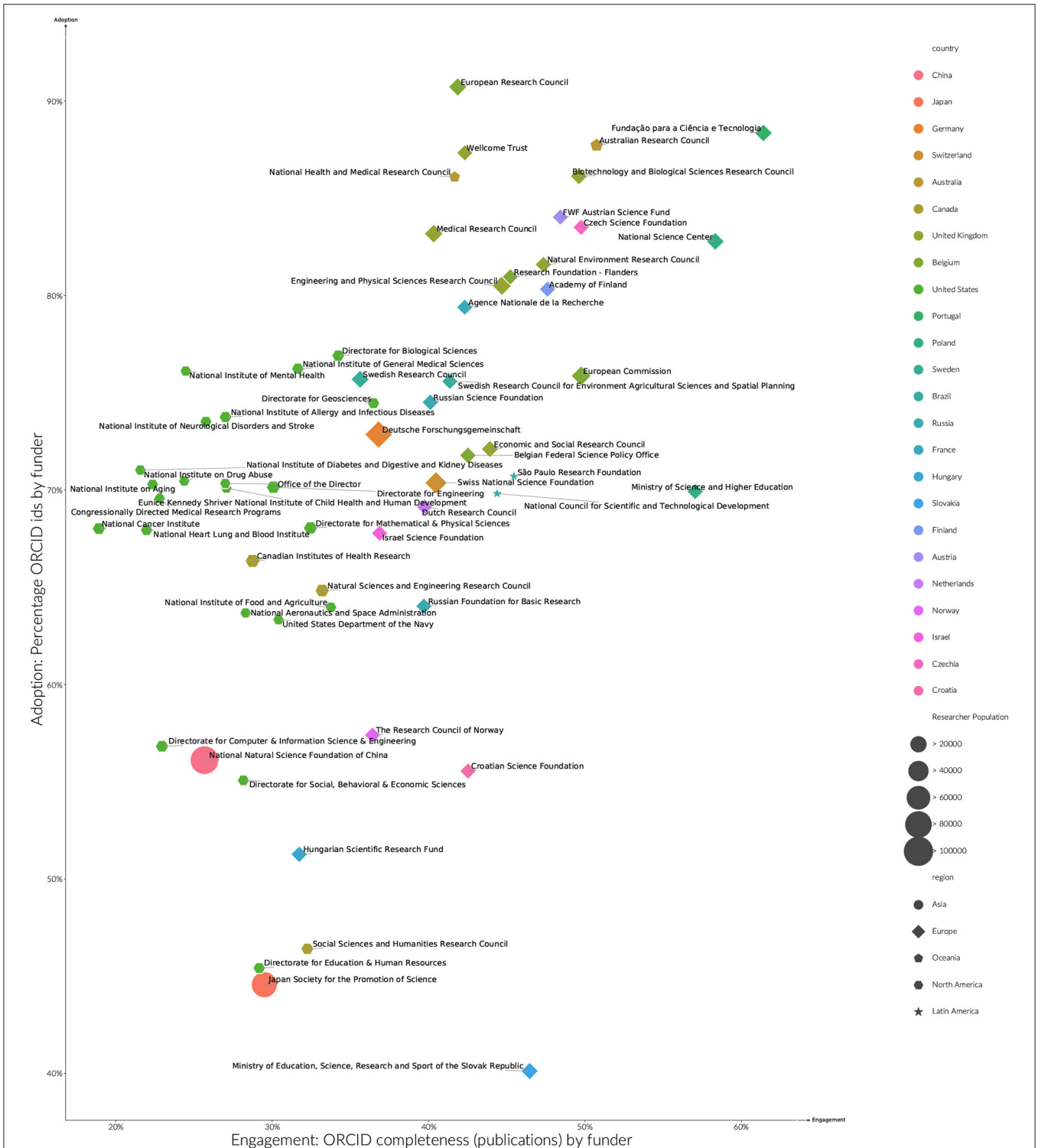


FIGURE 5 | Estimated ORCID adoption and engagement by funder researchers > 5 years publication history, publications (2015–2019). Sizes indicate the number of researchers by funder, with shapes denoting world regions. To aid readibility of funder names, axis do not start at zero.

3.3. ORCID Adoption - Publisher Level

Like funders, publishers support ORCID adoption and engagement via different mechanisms. ORCID adoption can be driven by publisher mandates. Engagement is supported most fully by providing all authors on a paper the opportunity to assert their ORCID iD. Publishers complete their responsibilities

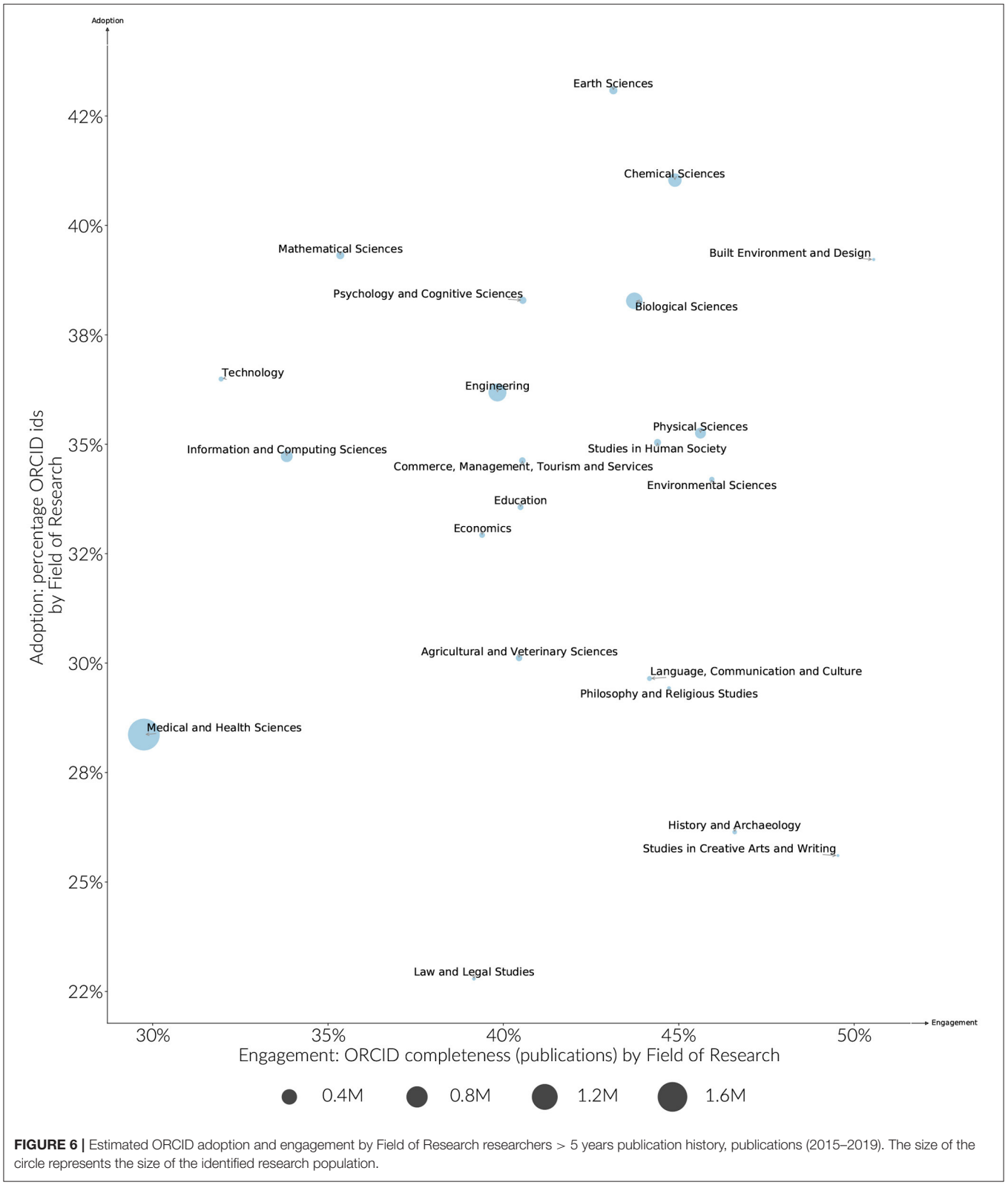


FIGURE 6 | Estimated ORCID adoption and engagement by Field of Research researchers > 5 years publication history, publications (2015–2019). The size of the circle represents the size of the identified research population.

as research information citizens by passing the ORCID metadata through to Crossref.

With a few notable exceptions, support for ORCID in publication metadata by journals and publishers has increased significantly, particularly since 2016. For the top 16 publishers by volume, **Figure 8** outlines the number of journals per publisher that have evidence of ORCID metatdata support in their Crossref records. With the exception of Wolters Kluwer, De Gruyter,

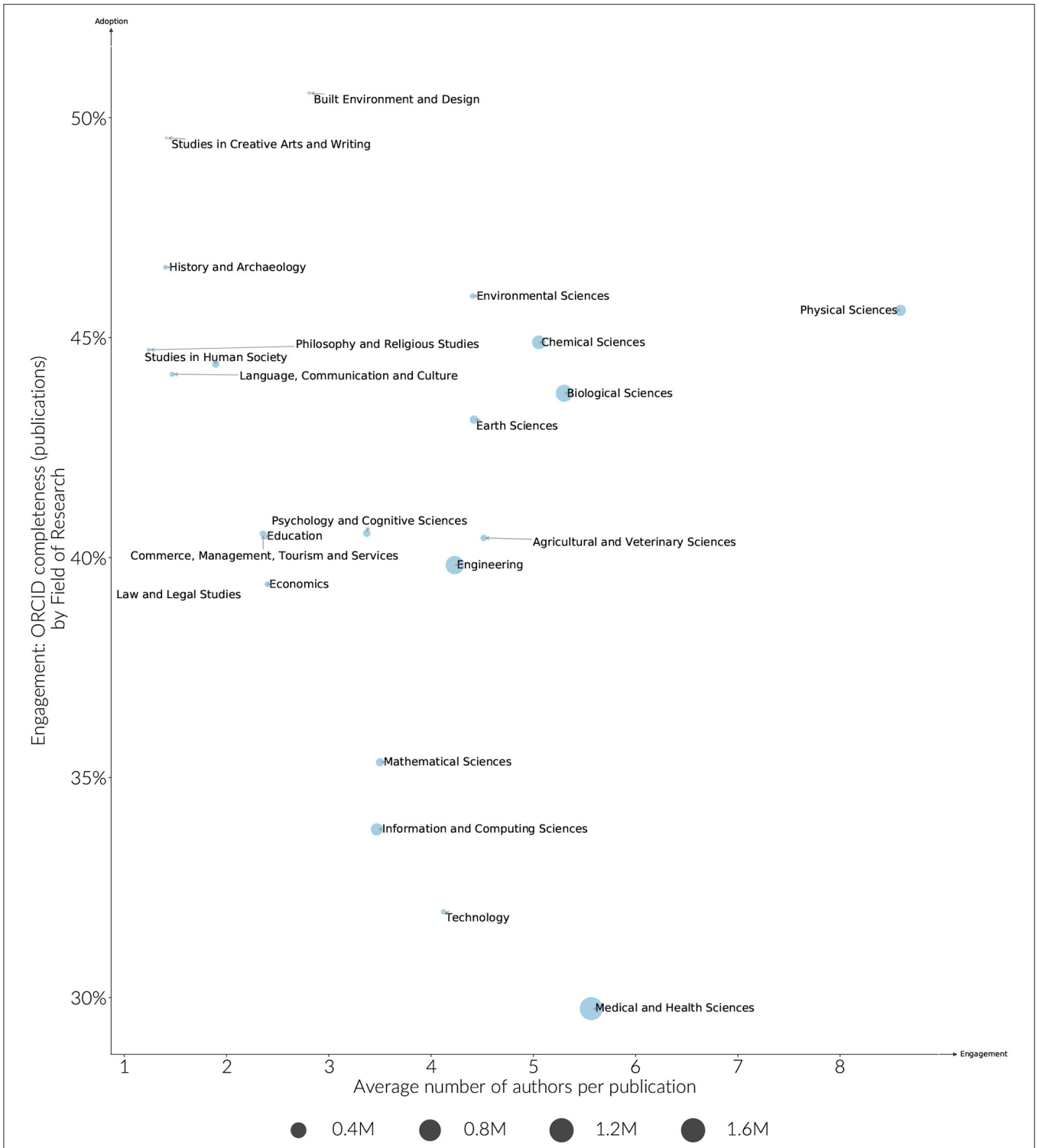


FIGURE 7 | Estimated engagement by Field of Research compared to the average number of authors per paper. Researchers are included if they have > 5 years publication history, publications (2015–2019). The size of the marker indicates the size of the identified researcher cohort.

and Frontiers, near complete journal support for expressing at least a minimum amount of ORCID metadata has either been reached, or there is a clear trend toward it. Presence of ORCID metadata in the Crossref records is not only a measure of publisher support adoption of ORCID, it is also a measure of community participation in open metadata that can be further consumed by downstream systems—a commitment outlined in the ORCID Open letter for publishers (ORCID, 2016). In the

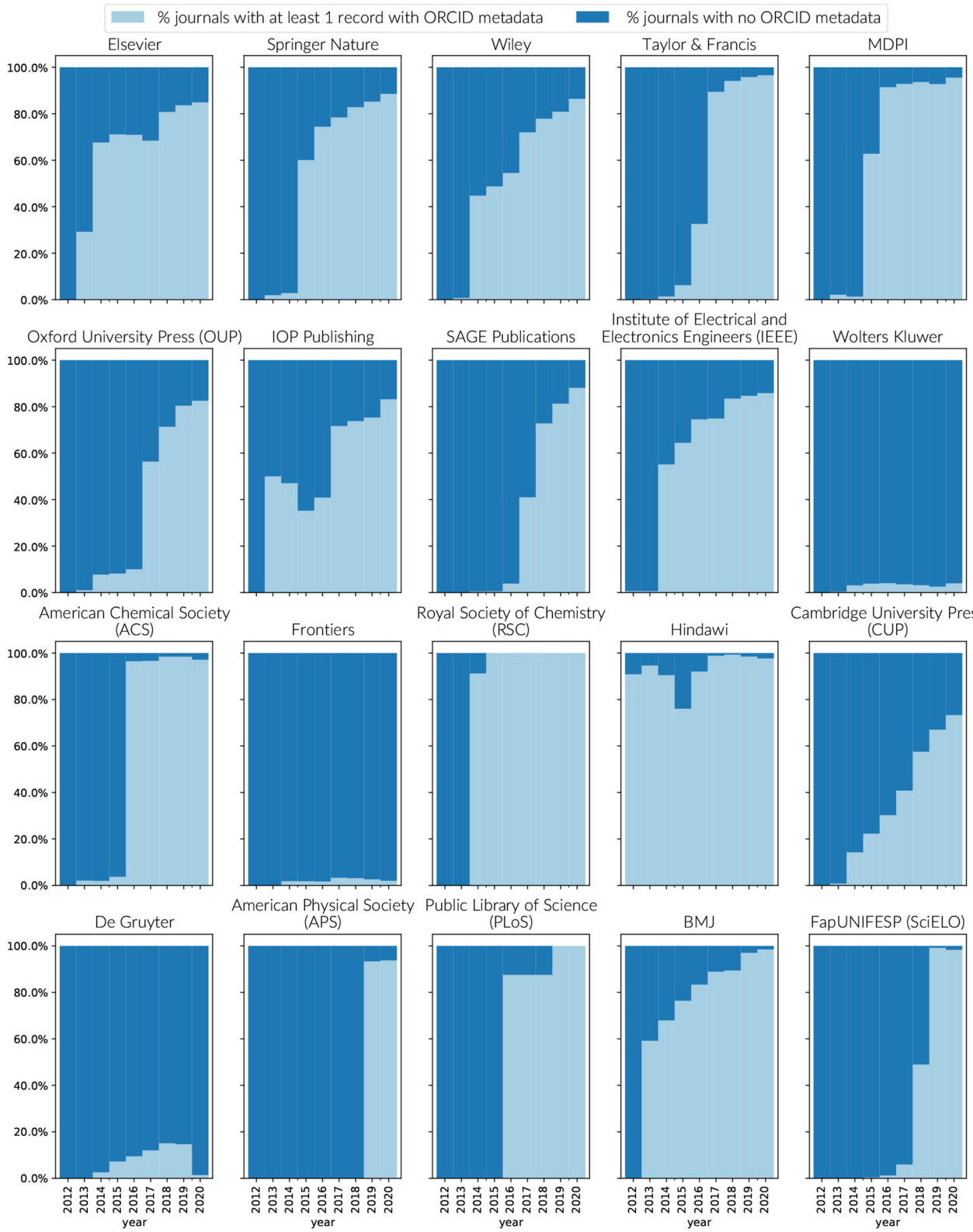


FIGURE 8 | Journals supporting the use of ORCID's within Crossref metadata by publisher by year. To be counted as supporting ORCID, a journal must have at least one record in Crossref that includes an ORCID iD in its metadata. Source: Crossref public file matched to Dimensions.

case of Frontiers, collection of ORCID iDs is a part of their workflow processes, however there was an oversight in passing the information across to Crossref [Internal Communication].

The level of support for ORCID iDs within publications by publisher is less uniform. In 2016, many publishers signed up to the commitment to require at least the corresponding author to connect their ORCID iD, with the understanding that all authors should be provided the option to assert their relationship to the paper (ORCID, 2016). Most publishers began their implementations by implementing the first requirement with support for additional authors proceeding at different paces (Meadows and Haak, 2017).

By looking at papers published in 2019 with more than three authors, it is possible to observe how this trend has since moved. Examining the top 20 publishers by volume of ORCID assertions in 2019 (see **Figure 9**), the dominant publishing mode was still one ORCID iD per paper, however, clear differences in publishing practice can be observed. Nine publishers had at least one ORCID on over 90% of their publications in 2019. Of these JMIR, stands out both in the fact that it has the highest percentage of papers with two or more ORCID iDs, and that its overall discipline that it serves (Medicine and Health Services) does not have a high researcher engagement rate. Eight publishers had a percentage of greater than 60% papers with two or more ORCID iDs per paper, with a further band of seven between 20 and 40%. Elsevier and Springer Nature, the largest of the publishers have approximately 10% of their papers with two or more ORCID iDs, although their coverage of papers with one ORCID iD differs significantly at 18 and 38%, respectively. That there is such a difference in the spread of support for more than one ORCID suggests that the constraint still lies within individual publishing platform implementations, rather than a willingness for researchers to change behavior.

4. DISCUSSION

In the previous section, a scientometric analysis of ORCID behavior reveals a research information citizenry that is serious about their obligations to each other, albeit one still in transition to ORCID-centric workflows.

We have shown that:

- In contrast to the internationalization of research, ORCID adoption and engagement patterns are regional, with countries such as Portugal, Poland, Denmark, and Australia leading the way and research giants such as the United States, China and Japan falling behind. Researchers within countries with low ORCID adoption rates are also more likely to be disengaged with their profile.
- ORCID adoption rates for funded researchers are significantly higher than their country averages, reflecting the influence of both publisher and funder mandates
- Publisher mandates have played a key role in encouraging ORCID adoption, however the capacity for researchers to supply ORCID iDs is now significantly outstripping publisher ability to record them as part of the submission process
- Publishers are meeting their responsibilities for distributed metadata stewardship around ORCID, however there remain some challenges in retrofitting new ORCID processes to existing submission workflows. These challenges resulted in an error rate of ORCID to author assertions of about .5% in 2020. Continued data quality monitoring is essential to ensure that this error rate continues to fall.
- ORCID adoption and engagement profiles differ significantly by research discipline, with Chemical Sciences and Earth Sciences having the highest rates, and Medical and Health Sciences the lowest. Moving beyond mandates, innovation in ORCID engagement by discipline provides a sustainable path for ORCID adoption going forward.

4.1. Addressing Researcher Disengagement

Critically, as might be expected, ORCID's success looks different by region, funding regime, and subject area. Each of these factors plays intimately with the likelihood of success of ORCID for a given researcher. If the researcher works in an established research economy in a high-income country with a dual-funding structure and national evaluation in a STEM research area then they are most likely to have both drivers to use ORCID and the opportunity to benefit from infrastructure investments. All this to say that depending on where in the world a researcher is based, they will likely have a significant difference on how integral is ORCID to their daily workflows.

For ORCID, Research Information Citizenship is not just about having an ORCID iD, but using it in expected ways. For a researcher, a key responsibility is not only ensuring that their information is kept up to date, it is also about ensuring that information can flow into their ORCID record with as little latency as possible. That countries with low engagement and adoption rates also exhibit a higher rate of disconnection between Crossref and ORCID is of significant concern. As publisher support for ORCID increases, these researchers are likely to experience the administrative burden of ORCID (which typically impact article submission workflows), without benefiting from the administrative benefits (which typically accrue during national evaluation or funding applications). Strategic engagement of these researchers will not only increase the local benefits of ORCID to the researchers involved, it also offers a path toward reducing the number of 'empty' ORCID profiles.

4.2. Emerging Strains Within Distributed Metadata Stewardship

On the other side of the relationship, it is remarkable that most publishers still publish more publications with only a single ORCID rather than multiple ORCIDs. Pressure to support ORCID assertions for all authors on a publication is mounting, with the capacity for researchers to supply their ORCID at the time of submission now outstripping functionality to support it.

Some journals are now choosing to implement ORCID policies that are beyond the current capacity of their publishing workflows Willighagen et al. (2019). For these journals, ORCID iDs will be supplied as part of the submission, however they will be unauthenticated by the researchers themselves, leaving

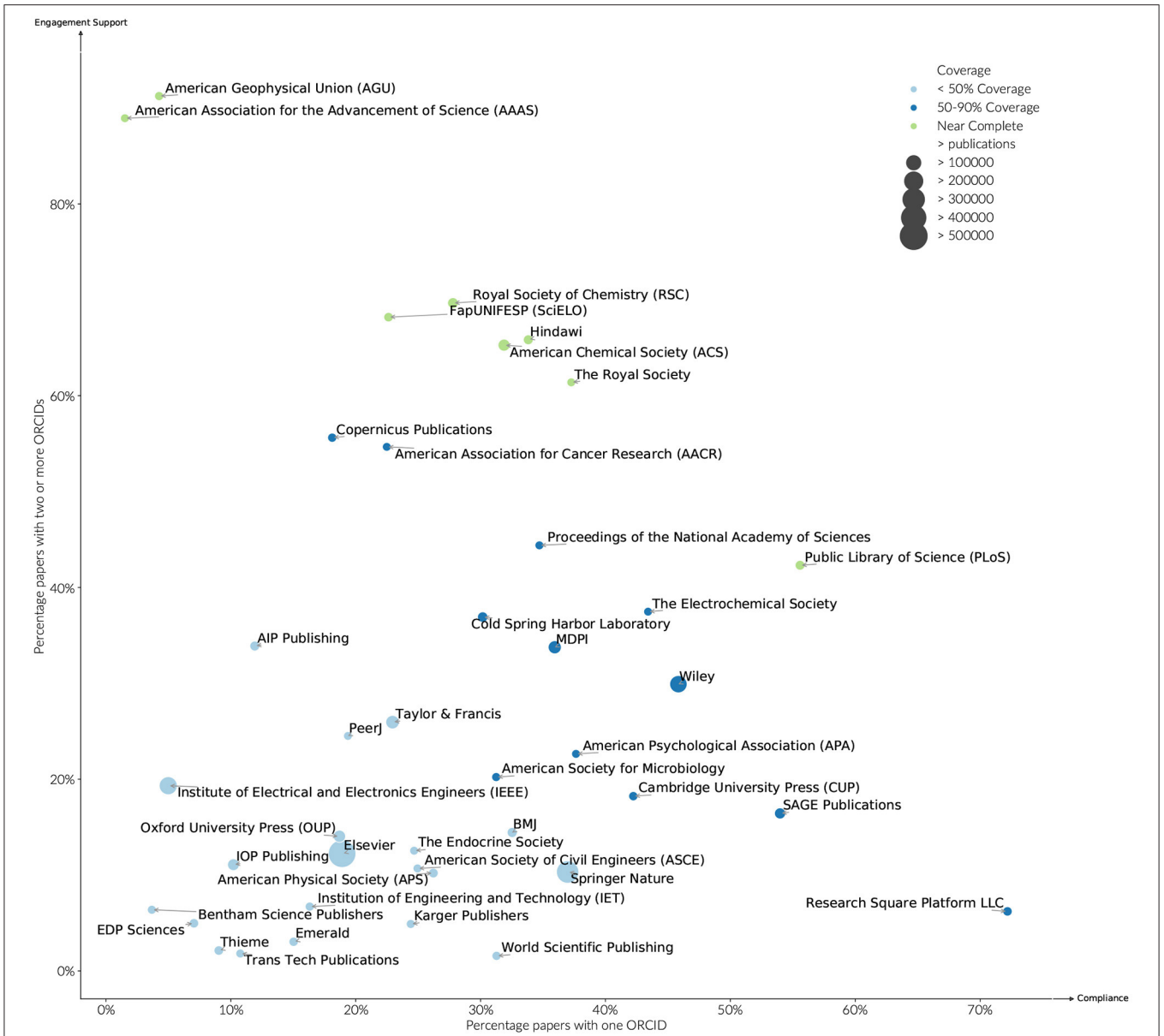


FIGURE 9 | ORCID adoption and engagement by publisher, publications (2019).

open the possibility that a researcher could be misidentified. It is possible that initiatives designed to increase ORCID engagement could also break community trust by introducing errors into the system.

This pressure on publishers will increase still further with an evolution of funder requirements around open access publishing. UKRI now require all authors to be uniquely identified by their ORCID iD on papers published after April 2022 (UKRI, 2021). Notably, the policy does not specifically require ORCID iDs to be authenticated, raising the risk that the number of unauthenticated ORCIDS will rise significantly. This level of funder activism is interesting in that it imposes a mandate on coauthors from other countries to add their authenticated ORCID to UKRI funded publications.

Overall, publishers can be seen to be meeting their research information citizenship obligations by passing on metadata through to Crossref. The problem of author shuffling as identified in Section 2 reflects a persistent inherent difficulty with these workflows. At the core of the issue is the task of assigning ORCIDs to the individual author statements made through the manuscript submission process. Workflows that begin with the free text author statement on a manuscript and require a decision to be made on which author belongs to which ORCID. These decisions introduce name matching errors that are difficult to completely overcome, particularly when retrofitting ORCID to

fit over legacy submission workflows. Continued monitoring of author shuffling with feedback to publishers to correct them should be considered an important activity to continue to maintain trust in the ORCID ecosystem.

4.3. The Importance of the Crossref Dataset When Measuring ORCID Adoption

When assessing the success of ORCID adoption and usage, we believe that we have demonstrated that it is not enough to assess the completeness of the ORCID registry in isolation. There have been many studies that compare the completeness and reach to research profiles such as ResearchGate (for example Boudry and Durand-Barthez, 2020). Because there are so many ORCID assertions in Crossref that have not made it back to the ORCID registry, this approach will almost certainly underestimate researcher ORCID engagement. By comparing ORCID to other profiling systems, such studies also risk incorrectly characterizing challenges with ORCID adoption as a choice between profile systems. This perspective leaves the research information citizenship that publishers and other actors exhibit in establishing the ORCID research graph unexamined.

4.4. Reflections on the Role of Scientometric Monitoring of ORCID Practices Going Forward

Scientometric monitoring of ORCID adoption and usage can offer insight into how ORCID practice is taking place in the community. Although providing an imperfect lens, by extending the known research graph through the use of natural language processing and algorithmic approaches, tools like Dimensions provide a way to observe these shifting dynamics, as well as make decisions about which interventions are likely to have the most impact in moving the research community forward. As illustrated by the ORCID journey, establishing new research practices centered around persistent identifiers require interconnected efforts to build new research infrastructure and change research practices. At different points of the journey, different approaches become possible. A first round of technical implementations for publishers focusing on connecting the first or corresponding author to their ORCID is now under pressure to accommodate all authors on a paper. What began as a push from publishers to make researchers supply their ORCID iDs is now reversing (in some disciplines) to be an expectation that all authors on a paper should be able to supply their ORCID. Through scientometric monitoring, we are able to identify these changes as they occur.

Scientometric monitoring can also play a role in selecting the most effective areas of research in which to innovate. Publisher support and innovation around ORCID may only just be beginning. Within disciplines where ORCID adoption and engagement levels are already high, it might also be possible to turn the relationship between author and ORCID on its head by adopting ORCID first approach to author assertions. Beginning with an ordered set of ORCID iDs, it would then be possible to derive the authorship statements on a paper. ORCID iDs could then be authenticated as part of the submission process (or as part of the document authoring process) without the additional requirement for author statement matching. Uncoupling ORCID author assertions from the submission process would also open up opportunities for greater collaboration between publishers and research authoring tools. Based on the observations made in this paper, it is more likely that innovations such as this would be more likely to take hold in fields with high ORCID adoption and completeness levels such as Earth Sciences or Chemical Sciences.

Of course, to be useful, the insights provided by scientometric analysis must be also by sufficiently accurate. Although aspects of this study, particularly the completeness calculations, would benefit from replication using data sources other than Dimensions, a level of calibration can be observed in the results themselves. For instance, the impact of ORCID interventions at the country level can be clearly recognized in the analysis above.

Finally, whilst this analysis has only measured funder contributions to ORCID adoption and engagement rates indirectly, funder interventions can be seen to correlate with high ORCID and engagement rates - particularly amongst countries with well established networks of current research information systems. Until recently, funders have expressed their role as a researcher information citizen as a consumer of ORCID information. More recently (ORCID Funder Working Group, 2019), in 2019 a move analogous to the publisher open letter (ORCID, 2016) a consortia of funders has proposed extending their role to also be a creator of ORCID assertions for grants, by creating both a public record of the grant with a DOI, and an ORCID assertion to it. As these new information pathways establish, and the known research graph continues to expand (Cousijn et al., 2021), scientometric approaches such as the one showcased here will provide an important methodology for charting its progress.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: doi: 10.6084/m9.figshare.16638337.

AUTHOR CONTRIBUTIONS

All the work contained in this article was conceptualized, carried out, and written by SP.

REFERENCES

Allik, J., Lauk, K., and Realo, A. (2020). Factors predicting the scientific wealth of nations. *Cross Cult. Res.* 54, 364–397. doi: 10.1177/1069397120910982

Australian Bureau of Statistics. (2020). *Australian and New zealand Standard Research Classification (anzsrc)*. Available online at: https://www.abs.gov.au/statistics/classifications/australian-and-new-zealand-standard-research-classification-anzsrc/latest-release

Baglioni, M., Mannocci, A., Manghi, P., Atzori, C., Bardi, A., and Bruzzo, S. L. (2021). "Reflections on the misuses of orcid ids," in *Proceedings of the 17th Italian Research Conference on Digital Libraries*, eds D. Dosso, S. Ferilli, P. Manghi, A. Poggi, G. Serra, and G. Silvello (Padua: CEUR Workshop Proceedings), 117–125.

Blackburn, R., Cabral, T., Cardoso, A., Cheng, E., Costa, P., Demain, P., et al. (2020). *ORCID Public Data File 2020*. ORCID. Dataset. doi: 10.23640/07243.13066970.v1

Boudry, C., and Durand-Barthez, M. (2020). Use of author identifier services (orcid, researcherid) and academic social networks (academia.edu, researchgate) by the researchers of the university of caen normandy (france): a case study. *PLoS ONE* 15, e0238583. doi: 10.1371/journal.pone.0238583

Brown, J., Demeranville, T., and Meadows, A. (2016). Open access in context: connecting authors, publications and workflows using ORCID Identifiers. *Publications* 4, 30. doi: 10.3390/publications4040030

Clark, R. (2020). *Metadata Deposit Schema-Crossref*. Available online at: https://www.crossref.org/documentation/content-registration/metadata-deposit-schema/

Clark, R. (2021). *New Public Data File: 120+ Million Metadata Records-Crossref*. Available online at: https://www.crossref.org/blog/new-public-data-file-120-million-metadata-records/

Cohen, A. (2015). *fuzzywuzzy*. Available online at: https://github.com/seatgeek/fuzzywuzzy

Cousijn, H., Braukmann, R., Fenner, M., Ferguson, C., van Horik, R., Lammey, R., et al. (2021). Connected research: the potential of the pid graph. *Patterns* 2, 100180. doi: 10.1016/j.patter.2020.100180

Dasler, R., Deane-Pratt, A., Lavasa, A., Rueda, L., and Dallmeier-Tiessen, S. (2017). Study of ORCID adoption across disciplines and locations.

Haak, L. L., Fenner, M., Paglione, L., Pentz, E., and Ratner, H. (2012). ORCID: a system to uniquely identify researchers. *Learned Publish.* 25, 259–264. doi: 10.1087/20120404

Hook, D. W., and Porter, S. J. (2021). Scaling scientometrics: dimensions on google bigquery as an infrastructure for large-scale analysis. *Front. Res. Metr. Anal.* 6, 656233. doi: 10.3389/frma.2021.656233

Hook, D. W., Porter, S. J., and Herzog, C. (2018). Dimensions: building context for search and evaluation. *Front. Res. Metr. Anal.* 3, 23. doi: 10.3389/frma.2018.00023

Leydesdorff, L. (1995). *The Challenge of Scientometrics: The Development, Measurement, and Self-Organization of Scientific Communications.* Leiden: DSWO Press, Leiden University. doi: 10.2139/ssrn.3512486

Meadows, A., and Haak, L. (2017). Orcid open letter-one year on report.

Mejias, G. (2020). *Collect & Connect-Improved and Updated! ORCID*. Available online at: https://info.orcid.org/collect-connect-improved-and-updated/

ORCID (2016). Orcid in publications.

ORCID (2020). *Funders' Orcid Policies*. Available online at: https://info.orcid.org/funders-orcid-policies/.

ORCID (2021a). Internal ORCID weekly statistic report. *ORCID*.

ORCID (2021b). ORCID record schema. *ORCID*. Available online at: https://info.orcid.org/documentation/integration-guide/orcid-record/

ORCID Funder Working Group (2019). *Orcid and Grant Dois: Engaging the Community to Ensure Openness and Transparency of Funding Information.* Available online at: https://orcid.figshare.com/articles/online_resource/ORCID_and_Grant_DOIs_Engaging_the_Community_to_Ensure_Openness_and_Transparency_of_Funding_Information/9105101/1

Porter, S. (2016). *Digital Science White Paper: A new 'Research Data Mechanics'*. Available online at: https://digitalscience.figshare.com/articles/report/Digital_Science_White_Paper_A_New_Research_Data_Mechanics_/3514859/1

Porter, S. (2021). *Bringing Narrative to Research Collaboration Networks In 3d-Digital Science*. London: Digital Science.

Puuska, H.-M. (2020). *Orcid in Publications*. Available online at: https://info.orcid.org/the-new-finnish-research-information-hub-provides-a-comprehensive-view-of-finnish-research/

UKRI (2021). *Ukri Open Access Policy*. Available online at: https://www.ukri.org/wp-content/uploads/2021/08/UKRI-060821-UKRIOpenAccessPolicy-FINAL.pdf

UNESCO (2021). *UNESCO Science Report: The Race Against Time for Smarter Development*. Paris: UNESCO.

Visser, M., Eck, N. J., v., and Waltman, L. (2021). Large-scale comparison of bibliographic data sources: scopus, web of science, dimensions, crossref, and microsoft academic. *Quant. Sci. Stud.* 2, 1–22. doi: 10.1162/qss_a_00112

Willighagen, E., Jeliazkova, N., and Guha, R. (2019). Journal of cheminformatics, ORCID, and GitHub. *J. Cheminform.* 11, 44. doi: 10.1186/s13321-019-0365-4

Conflict of Interest: SP was employed by Digital Science.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

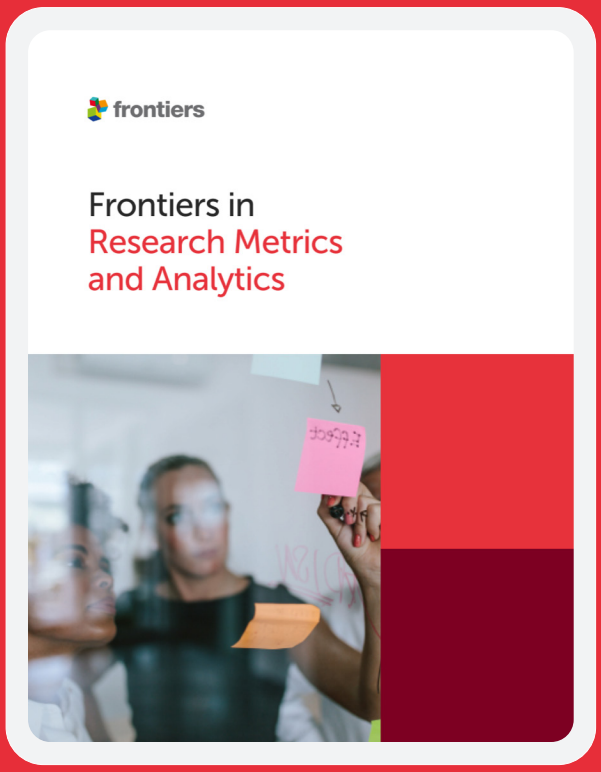
Frontiers in
Aging Neuroscience

Explores what drives excellence in methods of analysis

Provides a forum for the study of measuring, evaluating, and improving the efficiency, reliability, and transparency of research and innovation in all areas of scientific inquiry and applications.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact