

Digital Research Reports

The Ascent of Open Access

An analysis of the Open Access landscape since the turn of the millennium

Daniel W Hook, Ian Calvert and Mark Hahnel

JANUARY 2019

About Digital Science

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


About Consultancy

Our consultancy team delivers custom reporting and analysis to help you make better decisions faster. With in-depth knowledge of the historical and current research ecosystem, our unique perspective helps get the most value from data on the research lifecycle. Our team of data scientists are experts in using innovative analytical techniques to develop revealing visualisations and powerful insights. We understand the changing research landscape, and we can help you develop an evidence base on which to build the best research management and policy decisions. Visit www.digital-science.com/consultancy

About Dimensions

Dimensions is an innovative research knowledge system that re-imagines discovery and access to research. Developed by Digital Science in collaboration with over 100 leading research organizations around the world, Dimensions links grants, publications, citations, alternative metrics, clinical trials and patents. It enables users to find and access the most relevant information faster, analyze the academic and broader outcomes of research, and gather insights to inform future strategy. Data and expertise that span the research life cycle were contributed by the teams at Digital Science portfolio companies ReadCube, Altmetric, Figshare, Symplectic, Digital Science Consultancy and ÜberResearch, who came together to realize their unique strengths and share their passion for building tools that benefit the research community. Find out more at www.dimensions.ai


About the authors

Daniel Hook is CEO of Digital Science. He has been involved in research information management and software development for more than a decade, as Director of Research Metrics at Digital Science, Founder and CEO of Symplectic and COO of Figshare. Daniel is a mathematical physicist specialising in quantum theory and holds visiting positions at Imperial College London and Washington University, St Louis and is a Fellow of the Institute of Physics.  <http://orcid.org/0000-0001-9746-1193>

Ian Calvert is Head of Data Science at Digital Science. He went to the University of Birmingham where he studied artificial intelligence and then worked at the BBC on a wide range of software for embedded devices from prototypes on development hardware to applications in millions of homes.

 <https://orcid.org/0000-0003-0035-0599>

Mark Hahnel is founder and CEO of Figshare. Mark created Figshare whilst completing his PhD in stem cell biology at Imperial College London. Figshare currently provides research data infrastructure for institutions, publishers and funders globally. He is passionate about open science and the potential it has to revolutionize the research community.

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

Acknowledgements

We would like to acknowledge the assistance of Juergen Wastl, Suze Kundu and Simon Porter in bringing this report together.

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

Digital Science, 4 Crinan Street, London N1 9XW
consultancy@digital-science.com

Copyright © 2019 Digital Science

Introduction

Open access has only gradually become part of collective academic consciousness. While it has been a key issue for the publisher, funder and librarian communities for more than 20 years, most academics have come to be aware of open access in a more measured way. The speed of uptake of open access has depended on field, with economics and physics being among the very earliest movers as they pioneered the concept of pre-prints – sharing pre-published work in an open manner with colleagues. The pre-print movement has itself had a long history.

In high-energy particle physics the arXiv was established in 1991 by Paul Ginsparg who was associated with Los Alamos National Laboratory at that time. One of the first subject repositories, arXiv built on the practices of an existing community of academics who were exchanging pre-published work through email. LaTeX was the enabling technology behind this movement since papers (including equations) could be sent via low-bandwidth connections using email. Early users of the arXiv may recall extremely strict limits on file sizes to ensure that papers could quickly be downloaded in all geographies. Many of the most prestigious journals in Physics were, and continue to be, published by scholarly societies such as the American Institute of Physics, American Physical Society and Institute of Physics. In the event of pre-prints these societies may have felt threatened. Instead, even though material that they were publishing was openly available, academics still needed an accredited independent peer review process along with the esteem associated with publishing in reputed journals for their careers. The prevailing model of a journal as a sole source of material became enhanced by a customer-centric approach with increased discoverability (via arXiv) and shorter publication workflows. As a result publishers in Physics have by and large been able to retain the subscription model for the majority of their journals, even though much of the content is now freely available.

Life Sciences (and many other areas) seem to have had a less easy relationship with pre-prints and hence less content has been made openly available through pre-print type approaches. This is surprising as in almost all other respects, the Life Sciences field has been a progressive one in scholarly communications. The landscape in the biological sciences is, however, quite different from much of physics, mathematics and computer science: LaTeX is a less-used tool, and funding profiles, collaboration profiles and other drivers tend to be somewhat different. As result of these differences there has been a bifurcation, with the mathematical, theoretical and computational sciences seeming to coalesce around arXiv and RePEc style subject repositories, whereas the biological and life sciences have moved toward open access journals and the APC (Author Processing Charge) business model.

A mixture of these innovations such as subject repositories, the APC model, the emergence of institutional digital repositories and the rise of the Open Data movement, taken together with changes in the governance landscape such as funder mandates, institutional mandates, initiatives such

"The speed of uptake of open access has depended on field, with economics and physics being among the very earliest movers as they pioneered the concept of pre-prints"

"The prevailing model of a journal as a sole source of material became enhanced by a customer-centric approach with increased discoverability (via arXiv) and shorter publication workflows"

as SCOAP3, Project Deal and, most recently, Plan S, has led to a confusing environment that is not ideally placed to deliver on the aspirations that the open research movement wishes to see. Indeed, in many situations the incentives to share research data openly and to publish openly are far from aligned with the open research agenda¹.

In spite of this confused landscape, or perhaps because of it, sound science journals such as *Plos One* and *Scientific Reports* transformed into megajournals; collectively, just these two journals were responsible for publishing more than 1% of scholarly output in 2017. Open Access has unquestionably been an innovation driver for scholarly publishing.

¹ Mark Hahnel and Daniel Hook (2016) "Open by default" in *The State of Open Data Report*, Figshare, <https://doi.org/10.6084/m9.figshare.4036398.v1>

Open Access Initiatives

"Project Deal and Plan S are simply the most recent in a range of initiatives that renew the Open Access movement"

The march toward greater open access is not a single movement, nor it is propelled by a single agenda. Rather, its continued growth has relied on, and will continue to rely on, successive waves of innovation and a variety of initiatives that address different subjects, geographies, sensibilities and concerns. Project Deal and Plan S are simply the most recent in a range of initiatives that renew the Open Access movement. The positive part of this is that they give new energy to all parties involved and address difficult issues, the solutions to which gradually move us in a more open direction. The negative part is that each initiative does not necessarily follow seamlessly from the last, nor does it necessarily address the same constituencies, and hence we end up with a patchwork of legislation, licences and approaches that may even hold open access back.

The relationship that research has with data has fundamentally shifted in many disciplines over the last 30 years, and not just in the sciences. Digital Humanities is one of the fastest growing areas with some of the biggest challenges of data analysis. Again, in the early days, astrophysicists and particle physicists were the teams leading the need for faster, more distributed data infrastructures; the geneticists pioneered large scale analysis of data in the human genome project and now scholars in digital humanities push back the boundaries in machine learning to interpret, and bring meaning and context to, complex multifaceted data.

The combination of the Open Access and Open Data movements is perhaps the most enduring and compelling version of the Open movement. It is changing research communication, and with it the ecosystem of research itself. The reproducibility crisis that has been a recent focus in the scholarly and the popular press is the fruit of this new world and increases the need for openness, collaboration and efficiency. Negative research results are just as important as those that are positive, but until recently, with the development of Figshare, OSF, Dryad and other open data platforms there were few journals that were willing to publish these results. Of course, in this context openness is a double-edged sword – effects such as p-hacking both inside and outside research² can actually lead to a decrease in trust in research and researchers. In such a world context is everything.

² Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) *The Extent and Consequences of P-Hacking in Science*. *PLoS Biol* 13(3): e1002106. <https://doi.org/10.1371/journal.pbio.1002106>

Barriers to Open Research

Open Research is also a journey toward a more professional, well documented and collaborative research environment. It is a view of research that is gaining ground in spite of significant barriers. The system of incentives for academics is not well aligned to foster openness: Researchers continue to be incentivised to publish in volume rather than with the highest quality; they are incentivised to publish in top journals to support their job and promotion prospects; they are not recognised for being great creators, curators or sharers of data, nor are they recognised for their numerical and analytical work (just to pick a few examples). Scholarly communication of the future will certainly help to redress this balance, and the open access and open data efforts already bring some of the problems of the current structure of the academy into sharp relief.

The reproducibility crisis is just one manifestation of a broken credit system where researchers are incentivised to publish positive results and suppress or disregard negative results. This has highlighted deficiencies in the peer review process where academics do not have the time or resources to fully question and stress test the results that they are asked to review. Expecting a researcher to reproduce either entirely or in part the outcomes of a large or costly study or data collection exercise is not reasonable. We do not have a culture where peer review is paid for, and going to any of the lengths mentioned above may be prohibitively expensive, impractical or even impossible. The most frequently mentioned problem here is the increasingly unbalanced distribution of the review workload to particular countries, or people at particular stages in their career. Nevertheless, to achieve true reproducibility of results we need to look elsewhere. The peer review process alone is not fit for this purpose. Only by increasing openness of data, improved instrumentation and improved contextualisation of data can we increase reproducibility. Even then, there will still be some datasets that are linked to one-off events where reproduction is not possible, but we are talking in the broad generality of cases here.

Arguably research is always reinventing itself, but the changes taking place right now in scholarly communications are the greatest since the creation of the formal scientific publication by the Royal Society in 1665, and will have far reaching consequences in academia. While the foundation stone of new scholarly communication will be openness, it will be the reproducibility and collaboration that this new level of openness supports that will be the real win for research. But, in order to deliver these fundamental shifts at a systematic level, we need to reimagine the basic nature of a publication, and this will in turn lead to a re-evaluation of the incentives and drivers for professional academics.

"Open Research is also a journey toward a more professional, well documented and collaborative research environment"

"The system of incentives for academics is not well aligned to foster openness"

"The changes taking place right now in scholarly communications are the greatest since the creation of the formal scientific publication by the Royal Society in 1665"

The Role of Data

"From so many perspectives it is clear that research results need to be made openly available if we are to ensure that research retains its credibility for a broad audience"

The new “atom” of scholarly communication, beyond the publication, is profoundly driven by research’s new and emerging relationship with data. This new format for communication will require new infrastructure with reproducibility built-in. In this new world, peer review will remain central but will look completely different. With the rise of the post-truth or anti-expert era, it is critical that we develop mechanisms that make research open and reproducible that are beyond reproach. In years to come, the research world may look back on this period as one that has helped to positively define research communication in the future.

From so many perspectives it is clear that research results need to be made openly available if we are to ensure that research retains its credibility for a broad audience. Will that road be easy? No: There will be p-hackers who fail to understand the context of the data and who make the popular news in any case; there will be situations where data is shared, and the ethics of that sharing are questionable; there will be more “climategates”³, more retractions and more of the research conversation taking place in public than ever before. While the answer is unsettling, it needs to be that way for humanity to forge ahead and rebuild trust.

It is with this background in mind that we turn to an analysis of the development of Open Access since the turn of the millennium. We chart here, in a quantitative manner and at high level, the ascent of Open Access. There are many sociological and cultural, technological and financial challenges faced by Open Research, but new initiatives, technologies and policies continue to be the lifeblood of each successive wave of progress.

³ <https://www.nature.com/collections/synrzkgmlf>

Analysis

Data Provenance

We have used Digital Science's *Dimensions* and the data that it contains from Unpaywall⁴ to analyse open access trends between 2000 and 2016. The underlying dataset includes all publications with a DOI or PubMed identifier. At the time of writing *Dimensions* contains around 98m publications, more than 65m of which have a full-text record in *Dimensions* and from which funding acknowledgement data may be extracted from the body of the output. We include journal articles, monographs, book chapters and conference proceedings equally in our analysis below.

Funding references discussed below have been parsed out of acknowledgements sections wherever the data are available. This is an error-prone activity where some acknowledgements may have been missed. There are two levels at which funding acknowledgements are detected by *Dimensions* – at the level of the individual grant and the at the level of a funder. For the purposes of the analysis below, either a connection to one or more specific grants or a connection to one or more funders is interpreted as evidence of funding for the output.

The date range for the analysis was chosen with care: The year 2000 can reasonably be thought of as a good representative moment for when the Open Access movement became mainstream. Before 2000, Open Access had existed principally in the form of subject repositories such as those mentioned above. After 2000, the first open access journals began to explore different business models to support open publication: Plos was founded in 2000. The year 2016 has been chosen for more technical reasons. By our general definition of Open Access (i.e. where there exists a copy that can be freely accessed by a non-subscribing user) we need to be able to take embargo periods into account. These often only become clear in hindsight. Hence, 2016 is the “stable” year when embargo periods have now mostly expired and where articles are fully available.

Collaboration and Open Access

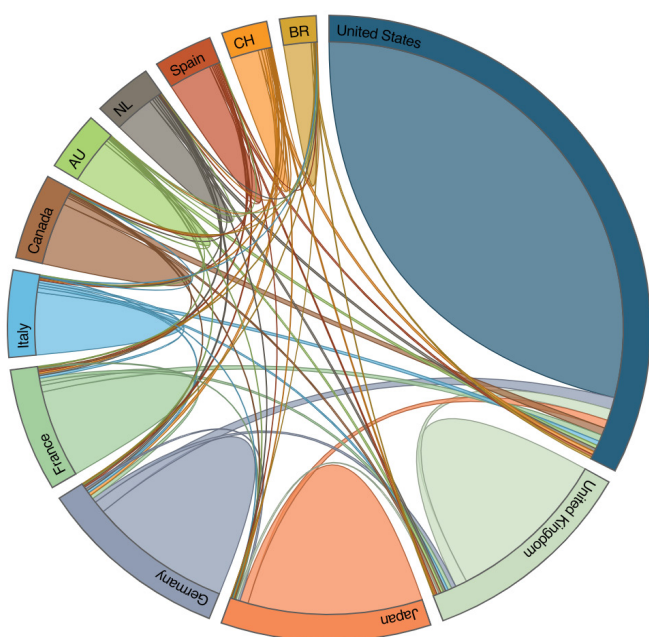
The volume of open access articles has clearly been rising in recent years. However, the overall volume of research has also been rising. In this context it is interesting to look at the individual trends that emerge based on the different strategies that countries have employed depending on their view of open access.

In Figure 1, we see collaboration diagrams for the top twelve research-publication producing countries. Interestingly, the top 12 research-publication producing countries are also the top OA-publication producing countries and remained broadly similar over the period, with only Switzerland and the Netherlands departing the group in favour of China and India as the research economies of Asia developed in the period.

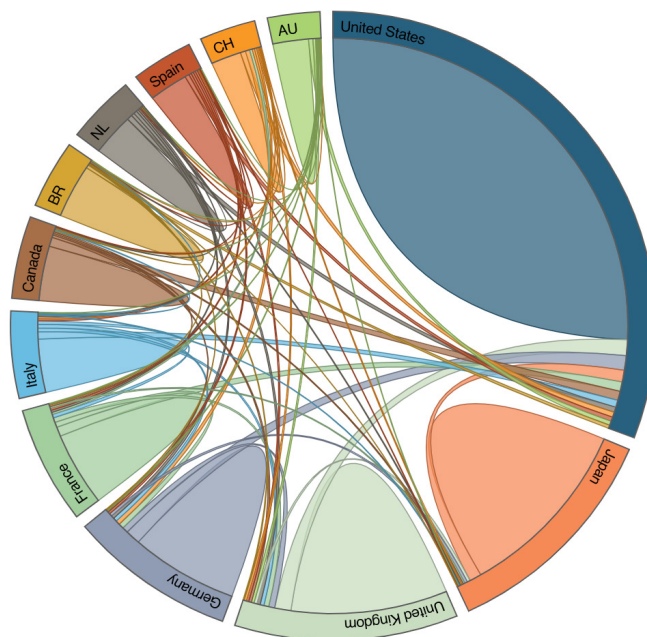
"We have used Digital Science's *Dimensions* and the data that it contains from Unpaywall to analyse open access trends between 2000 and 2016. The underlying dataset includes all publications with a DOI or PubMed identifier"

"The volume of open access articles has clearly been rising in recent years. However, the overall volume of research has also been rising"

⁴ <https://unpaywall.org/>



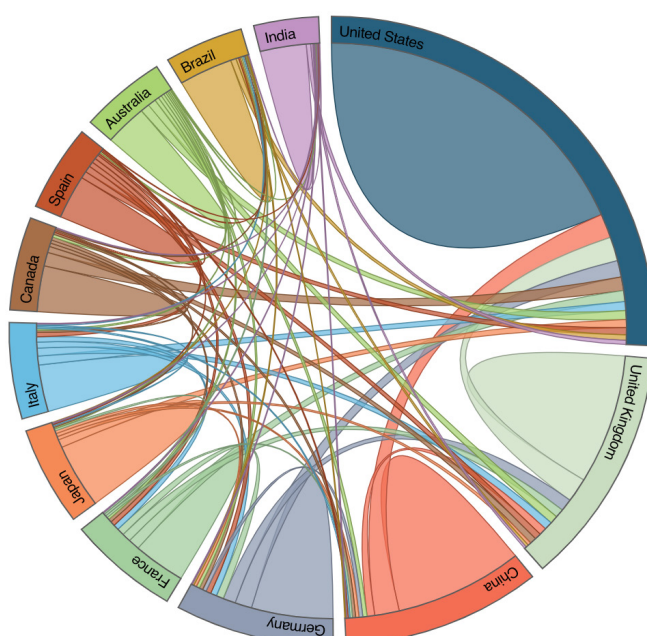
a: 2000: Collaboration profile for each country (including Open Access content)



b: 2000: Collaboration profile for just the Open Access segment



c: 2016: Collaboration profile for each country (including Open Access content)



d: 2016: Collaboration profile for just the Open Access segment

Figure 1: Chord diagrams showing the collaborative volumes between the top 12 Open Access publishing countries between 2000 and 2016. Each chord diagram shows a different filter of research output: In the left panels (a and c) we see the overall collaboration profile for each country (including Open Access content); in the right panels (b and d) we see the collaboration profile for just the Open Access segment of publications⁵

⁵ The chord diagrams in Figure 1 give an impression of the volume of papers for each country and the proportion of papers that are collaborative with each other of the main open access publishing countries. Each country lacks a segment to show the collaboration outside the countries directly involved in the diagram – this is a simplification that does not significantly affect the overall picture. Figure 1 is unnormalised for co-collaborations that involve many countries. Hence, there is n-fold counting where papers are collaborative between more than 2 countries. Given the volumes involved in these categories the effect is negligible.

In 2000, the US dominated both the overall publication landscape and the OA-publication landscape with a close collaboration relationship with the UK as the second-highest overall producer, and Japan with the second-highest OA producer. Collaboration on OA papers was significantly higher than collaboration on non-OA outputs across all 12 countries in 2000. The US was marginally less dominant in the OA sector in 2000 with slightly higher market shares for all other participants. It is an interesting and perhaps telling insight that Brazil, whilst 12 in overall production in 2000, was 8th in OA-production at the same time.

By 2016, the collaboration picture looks quite different. First of all, China has gone from not appearing in the top 12 producers in either category to being the second highest producer overall and the third highest in OA. India's more measured success is still no less remarkable. The combination of these two rising research powers in both overall production and OA production, together with the resultant displacement of established research economies in Europe, is an indicator of things to come. All countries appear to be increasing the proportion of their research that is collaborative, with OA seeming to be the vanguard to forward signal this type of collaboration. In this context, those countries that have invested in Open Access have managed to stay near the top of the output rankings – notably the UK has punched above its weight through successive initiatives to champion the cause of OA, whereas Japan, initially a great proponent of OA, has descended in the table as its capacity to collaborate internationally has waned versus the average.

Strategic Position and Open Access

Countries that have invested in Open Access have typically increased their level of international collaboration. The reasons for this can be manifold, however, it is possible that countries who have been able to afford to engage with Open Access and who have been able to take risks around the openness of their intellectual capital may also be the kinds of country who have travel funding and who, consequently, were always going to perform well. It is interesting to note the UK's substantial commitment to Open Access through successive waves of initiative: technological commitments through Jisc, institutional mandates and high-profile strategies such as the REF2021 and funder mandates, as well as more controversial approaches such as those established by the Finch Report have provided a continuous stream of actions that have continued to motivate the development of Open Access (Figure 2). Over 52% of the UK's output is available through Open Access channels, accounting for 7% of world output.

This has clearly been a significant strategic advantage of the UK and has allowed them to retain a disproportionately highly-ranked position in Open Access output, fending off China for several years while other countries progress at a more sedate pace. Brazil is another success story, second only to the UK, with 51.2% of its research output available through open access channels. Countries with slower rates of development such as Japan, Canada or France have descended in the table while countries with smaller research bases have also not been able to keep up with overall production rates in spite of their significant investments (Switzerland, Netherlands). A

"Countries that have invested in Open Access have typically increased their level of international collaboration"

"The UK's substantial commitment to Open Access through successive waves of initiative has clearly been a significant strategic advantage, and has allowed them to retain a highly-ranked position in Open Access output"

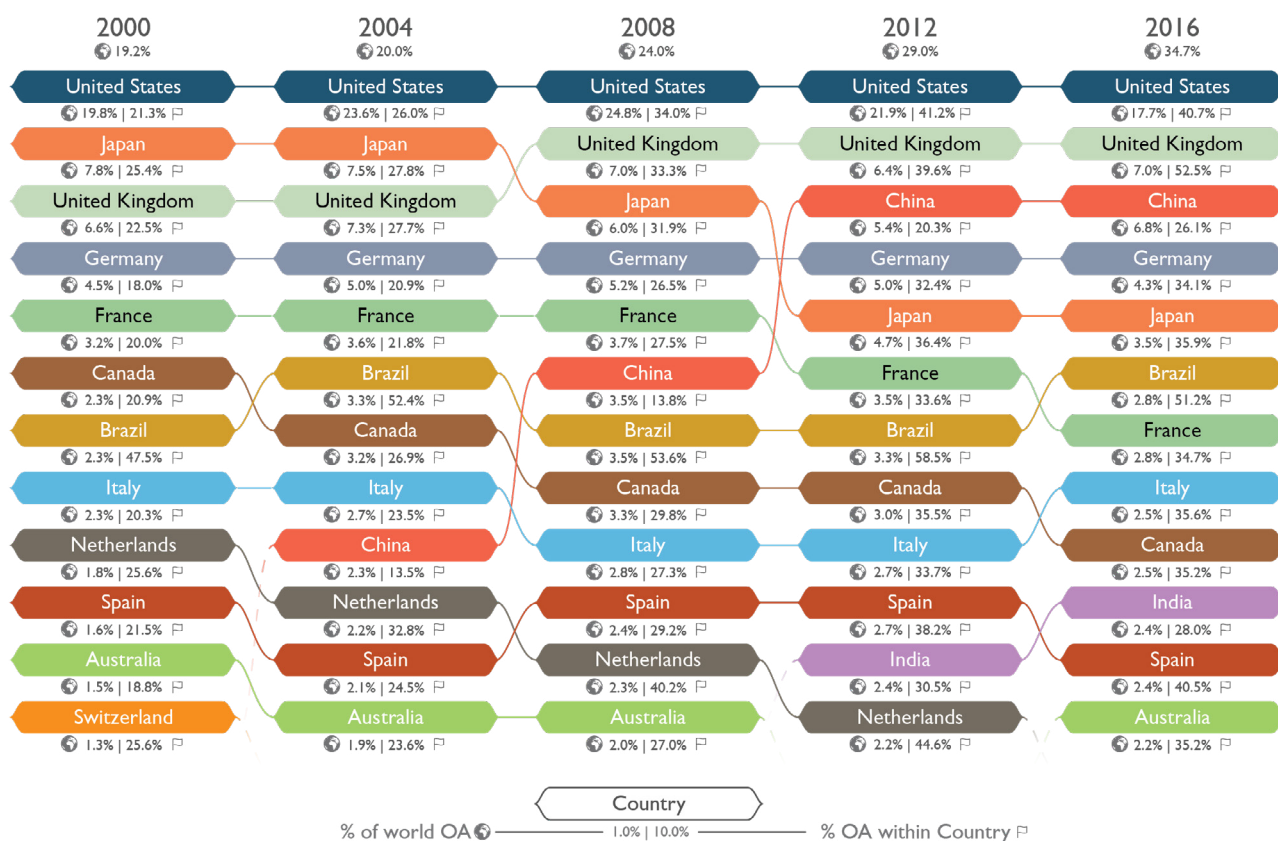


Figure 2: Placement of the top twelve OA publishing countries by volume in 4-year segments between 2000 and 2016. The world icon accompanies the overall percentage contribution to global open access publication while the flag icon denotes the percentage of OA relative to the overall output of the country in each case.

significant outlier is Australia, which has managed to maintain a position in the top 12 through aggressive growth in Open Access to offset the relative size of their university sector. At the top of the table, it is notable that the US remained fairly stagnant between 2012 and 2016 as OA rates peaked at around 41% and their world share of OA actually reduced by around 4% as the rest of the research world (and notably China) began to produce more, with more of it available through Open Access routes.

Translation and Open Access

Many have claimed that Open Access is a route to higher citation and hence it is clear that there is an alignment of academic incentives for publication. However, perhaps more interestingly, there is an alignment with Open Access and the garnering of Altmetric attention. This suggests that Open Access also serves to position a piece of academic work more positively for translation or for impact beyond the scholarly circles.

To illustrate the advantage that Open Access creates, we have taken the Open Access articles published by the top 12 OA-producing countries in 2016 as a basis for analysis. In Figure 3, the left side of the picture shows the proportion of articles published in either Open Access (pink) or non-Open Access (brown) channels in the year – 35.1% in OA compared to 64.9% non-OA. From the beginning, it is clear that Open Access makes a difference to Altmetric attention, with 53.2% of OA papers being tracked with an Altmetric mention versus 46.8% of non-OA papers being tracked

"There is an alignment with Open Access and the garnering of Altmetric attention. This suggests that Open Access also serves to position a piece of academic work more positively for translation or for impact beyond the scholarly circles"

"Open Access, Funded and Internationally Collaborative papers account for just 6.3% of all output but garner 15.2% of the citations"

"In general, we observe that it is better to publish in Open Access venues to optimise citation and Altmetric attention. Both measures are improved by being funded and collaborating internationally"

with at least some Altmetric attention. On the right side of the diagram we see the corresponding citation picture – the 35.1% of publications that were published through OA channels received 47.6% of the citations that have been made to date.

On both sides of the plot we move in one level and introduce two subdivisions: Funded (green) and Unfunded papers (light blue). For example, 16.5% of publications produced by the top 12 OA-producing countries in 2016 were published Open Access and were funded. 31.8% of them received Altmetric attention and they received 33.4% of the citations available. As another example, 44.7% of publications in the corpus were published through non-OA routes and were unfunded. 29% of them received Altmetric attention, and they received just 23.2% of the citations available. The clearly outperforming classification in the third level of the diagram is the golden coloured section at the bottom on the left: Open Access, Funded, Internationally Collaborative papers. These papers account for just 6.3% of all output but garner 15.2% of the citations (averaging 12.3 citations per paper since 2016). The least well-performing area is non-OA, unfunded, domestic papers, which make up 36.1% of all papers written in the cohort. These papers account for 16% of citations (averaging just 2.3 citation per paper since 2016) and just 22.5% of this class of papers received Altmetric attention.

In general, we observe that it is better to publish in Open Access venues to optimise citation and Altmetric attention. In addition, it is unsurprising to learn that both measures are improved by being funded and collaborating internationally.

Using the new classifications for Pure Gold, Hybrid, Bronze, Green (Submitted), Green (Published) and Green (Accepted) in Dimensions, we can get a high-level impression of which form of Open Access is the most advantageous. Using similar filters to those used for Figure 3 we are able to derive the table below.

Table 1: Comparison of citations and Altmetric attention by Open Access classification for the top 12 OA-producing countries in 2016

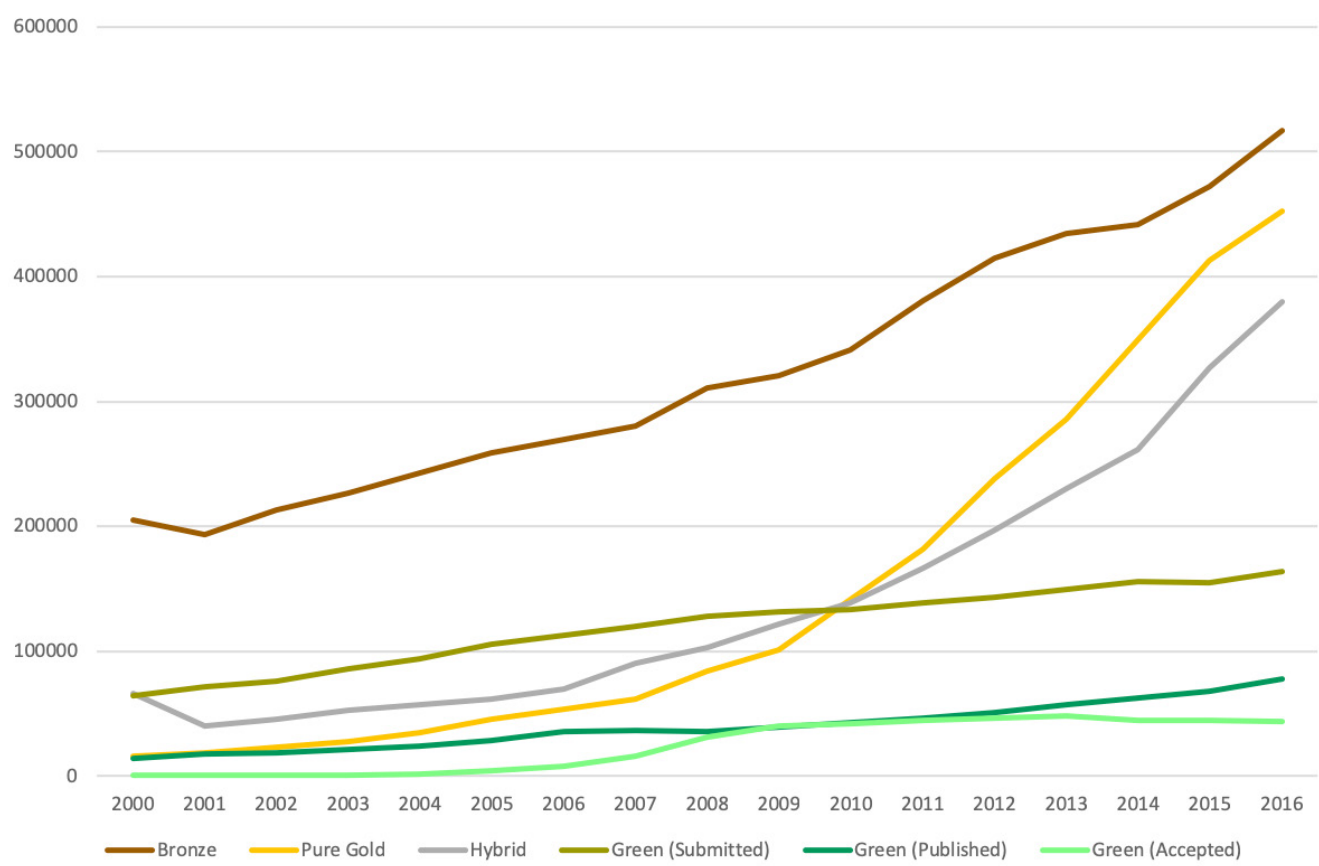
	Percentage of All Open Access Papers in this Channel	Cites / Paper	Percentage of Papers Receiving Altmetric Attention
Pure Gold	33.9	5.7	52.1
Hybrid	17.0	7.9	46.3
Bronze	23.4	6.6	51.1
Green (Submitted)	12.0	7.3	50.2
Green (Published)	7.7	6.4	53.2
Green (Accepted)	5.8	12.3	72.6

Green Open Access Classifications:

Green (Submitted)	Free copy of submitted version, or where version is unknown, in an Open Access repository.
Green (Published)	Free copy of published version in an Open Access repository.
Green (Accepted)	Free copy of accepted version in an Open Access repository.

While Pure Gold was the preferred route to publish in Open Access in 2016 among the top 12 OA-publishing countries (Table 1) the global trend was slightly different with Bronze being the channel of choice (Figure 4). There are clearly merits to the Green (Accepted) route with high attention in both citations and Altmetrics, even though Bronze and Gold routes are the fastest growing channels.

Figure 4: Trend in number of publications by Open Access type for all publications



"Dimensions as a data source gives us a new and inclusive way to understand the Open Access landscape"

"Continued waves of innovation in policy and technology will be needed to take us from the current state of Open Research to some future equilibrium"

Discussion

Dimensions as a data source gives us a new and inclusive way to understand the Open Access landscape, not simply in terms of citations and Altmetric attention, but also in the context of research funding, publisher engagement, institutional and country-based trends, the translation of research from Open Access pathways to patent, and the strengths of particular Open Access types or channels. This report explores the merest fraction of the potential of the *Dimensions* data source in the context of Open Access.

At this stage, it is beyond reasonable doubt that Open Research will form the basis of the future of global, publicly-funded research but the nature of outputs in the future and the nature of that openness is yet to be determined. The principal channels of open publication are still being formed and the sustainability of business models and the infrastructure that needs to exist to support research communication in the 21st century are still in the early stages of development. Continued waves of innovation in policy and technology will be needed to take us from the current state of Open Research to some future equilibrium. We hope that tools such as *Dimensions* can help to inform that future.

Part of **DIGITAL**science



digital-science.com