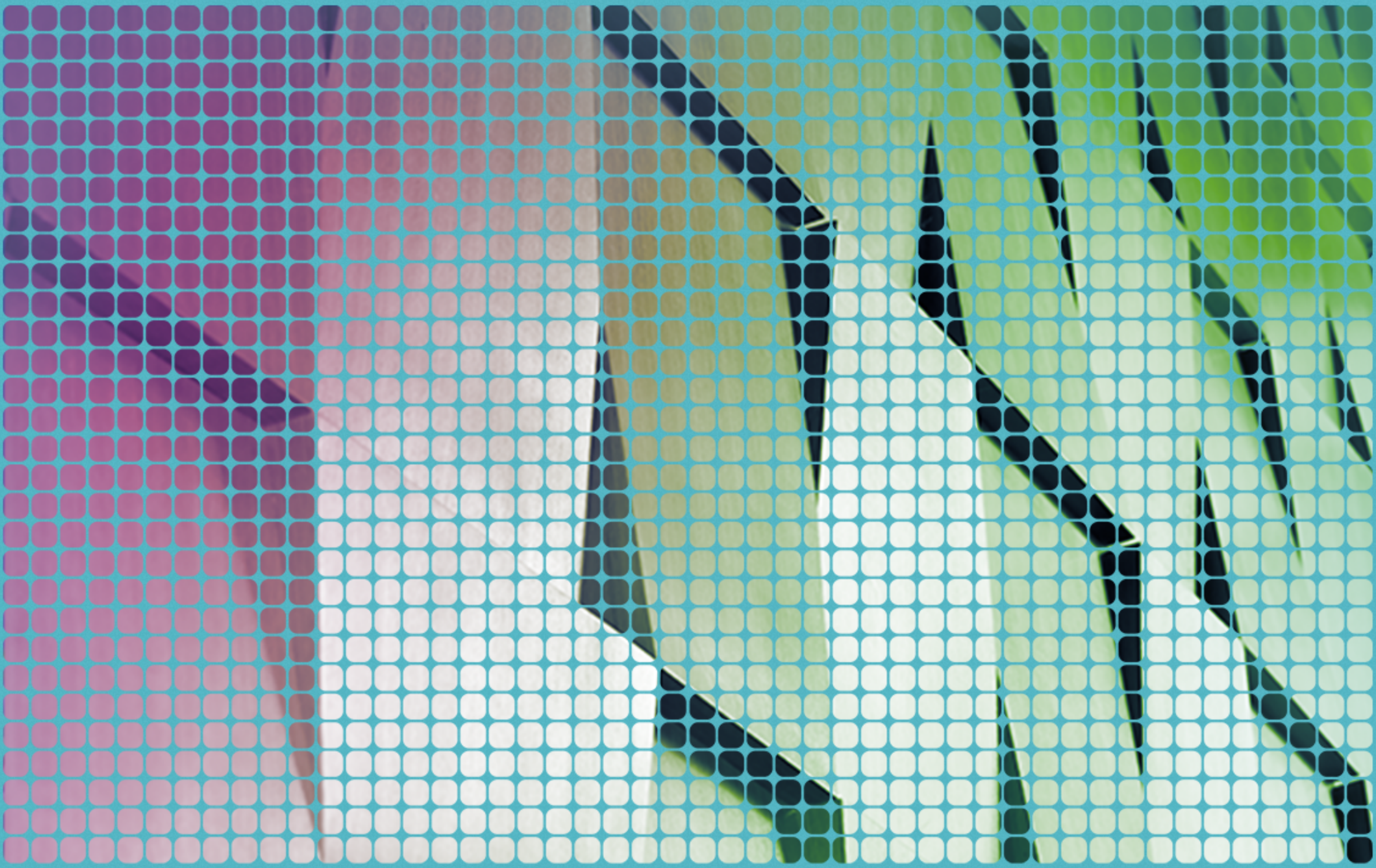


Digital Science Report

The State of Open Data 2022

The longest-running longitudinal survey and analysis on open data

Foreword by Mark Hahnel, Founder & CEO of Figshare

October 2022

DIGITAL science SPRINGER NATURE figshare

> “ *NIH is also aware that not just technological advancement, but behavioral change is also necessary to advance data science goals* ”

Ishwar Chandramouliswaran
Amy Hafez
Taunton Paine
Susan Gregurick
NIH

Contents

Foreword

Dr. Mark Hahnel

Founder and CEO

Figshare

2022 marks 10 years of Figshare and 7 years of 'The State of Open Data', the longest quantitative survey on researcher attitudes towards open academic data.

In those 10 years, we have seen data become a priority for academic stakeholders including Governments, Funders, Publishers and Institutions across the globe and across fields of research. This year's report highlights the global nature of this push, as well as the stakeholder and thematic viewpoints. We invited contributions from experts that represent all of these factors and variables in the articles in the report. We have a humanities and publisher viewpoint from Taylor and Francis, F1000 and Wiley. We have the Institutional Perspective from South Africa. We have the funder perspective from the National Institutes of Health (NIH) and the Computer Network Information Center of the Chinese Academy of Sciences (CNIC, CAS).

The increased globalization of open data is evident, notably with strong growth in Asia. Last year, survey responses from China made up 3% of the sample, whereas in 2022 they account for 11%. There has been an uptake in the Chinese national generalist repository, ScienceDB, with a 21x growth in the number of data depositors versus 2021. Experts from the Computer Network Information Center of the Chinese Academy of Sciences point out in their contribution to our report that training in the space has guided this change in researcher practices and will continue to be a focus. This is a recurrent theme throughout the report; evidence of successful uptake due to training, juxtaposed with a need for training to ensure the global uptake and benefit of open data runs in parallel with global mandates.

A further illustration of the progression of policies comes in the waning enthusiasm from researchers for open data mandates as the rubber hits the road and good intentions translate to more compliance. The goals of open data and more broadly open research are noble. They are essential for a more equitable society and level playing field for all. This does, however, mean more administrational process for researchers in making their outputs open. Concerns about ‘Who is going to fund all of this?!” seem at this point to be assuaged by funders stating they will cover “costs”, but the survey implies this message is not reaching researchers. All in all, two thirds of researchers consider funder mandates a necessary friction point for researchers to go forward towards the next paradigm of research. In the United States the Office of Science and Technology Policy (OSTP) was instrumental in setting expectations for federal funding agencies to require the planning and management of research data resulting from extramural research. As a result of this policy, and others worldwide, publishers began requiring Data Availability Statements (DAS) within research papers. These statements are designed to accelerate data sharing.

4/5 respondents are in favour of research data being made openly available as common practice

This next paradigm, the fourth paradigm of research, as coined by Jim Gray et al at Microsoft Research just over a decade ago, imagines knowledge discovery based on data-intensive science. Like the goals of FAIR (Findable, Accessible, Interoperable, Reusable) data, the fourth paradigm predicts that knowledge discovery can be accelerated by making use of the machines. Artificial intelligence algorithms and Machine Learning workflows are highlighting new patterns and predictions at a scale that the individual researcher’s brain cannot compute. The last 12 months have seen these aspirations begin to come to reality, not least in the success of AlphaFold, the Google AI company winning the Breakthrough Prize’ for their work predicting the 3D structure of proteins. For those working to provide a way for researchers to share their heterogeneous research outputs, this has always seemed like a distant goal, but an aspirational goal nonetheless.

In his commentary piece, Samuel Simango of Stellenbosch University highlights South Africa's push for 'Data for Good' - that principles promote the production and use of data for the advancement of the social good. Transparency, reproducibility and replicability are the short term goals of open data mandates that are being realized today, in all areas of research - whilst the step changes in biomedicine provide long term focus.

While most trends are encouraging around the adoption and acceptance of open data, the research community is now demanding more enforcement of the mandates that have been adopted by many governments and funders. We have seen many engaged funders and governments, most notably the recent memorandum from the Whitehouse Office of Science and Technology Policy, requiring that data that they fund be published. This has also led to national initiatives for **Research Data Management and Dissemination**. The NIH is not the first funder to tell the researchers they fund that they should be making their data openly available to all. 52 funders listed on Sherpa Juliet require data archiving as a condition of funding, while a further 34 encourage it. A push from publishers has also acted as a major motivator for researchers to share their data. This goes as far back as PLOS requiring all article authors to make their data publicly available back in 2014. Now, nearly all major science journals have an open data policy of some kind.

72% of researchers said they would rely on an internal resource for help with managing or making their data openly available

Some may say there is no better motivator for a researcher to share their data than if a publication is at stake. When asked who they would be willing to receive support from, the most popular answer was publishers (41%) closely followed by those within their own institution (38%).

This makes sense as researchers see the publishers as those responsible for disseminating research, but given the ongoing battles around open access funding models, there is still healthy debate to ensure monopolies on data publishing are not established.

The Generalist Repository Ecosystem Initiative (GREI) from the NIH, providing funding for better interoperability and co-opetition between generalist repositories is paving the way for large scale improvements in data publishing. NIH is interested in establishing partnerships with communities, societies, and external programs to enhance the education, adoption, and implementation of FAIR practices through collaborative projects, workshops, and other activities.

Taken in whole, the survey points to a need to plug holes around training in open data, to remove yet more administrative burden from researchers. If the next 10 years can progress at the same rate as the last decade when it comes to open research, the priorities need to be:

- Better metadata
- More metadata
- FAIR metadata

To do this, we need support for training and support for human and machine based checks. I personally have commented in the past that funder policies need more support to realise. When I say support I mean money. The NIH is commendable in their 2021 GREI initiative, but more is needed from Funder, Publisher and Library budgets. The evidence that we can move further, faster in knowledge creation is upon us. These budgets should focus their attentions specifically on the following:

- Fund training, librarians, educators
- Fund curation of open research data

A future of ubiquitous research data publishing in academia is in reach. We have a great opportunity in the data space with a constant pressure on funders to require FAIR data publishing. There is no putting the genie back in the bottle when it comes to society’s demand for open research data. It may prove to be a step change in knowledge discovery if all stakeholders continue to push for unobstructed, equitable data publishing with high quality metadata for humans and machines.

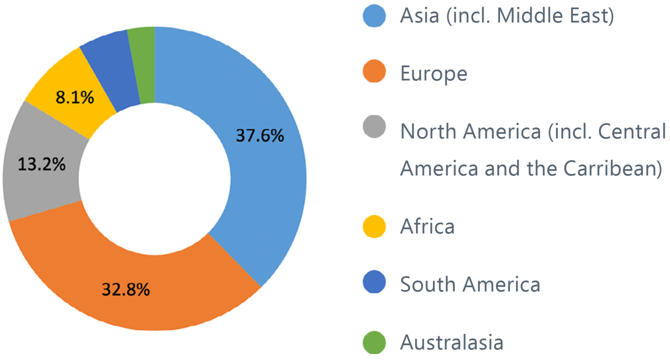
75%
of researchers said they receive too little credit for sharing their data openly

Key takeaways from the State of Open Data 2022

Laura Day
Product Marketing Manager
Figshare

Dr. Greg Goodey
Senior Research Analyst
Springer Nature

Now in its 7th year, the State of Open Data survey has had approximately 27,000 responses from 192 countries and continues to provide a detailed and sustained insight into the motivations, challenges, perceptions and behaviors of researchers towards open data. This year, the survey received the largest number of responses since 2019, with over 6,000 usable responses.



A diversification of voices

Open data and effective data management has been a strong focus for many researchers in Europe for a number of years and we've consistently seen high levels of engagement with the State of Open Data survey from those in the region. In 2022 there are some shifts that have provided us with more demographically diverse insights.

This year there was a significant increase in the number of respondents from China when compared with 2021 results. Last year, survey responses from China made up 3% of the sample, whereas in 2022 they account for 11%. By continent, the largest response was from Asia (38% including the Middle East) followed by Europe (33%). In their contribution to this report, Yuanchun Zhou and Lulu Jiang from the Computer Network Information Center of the Chinese Academy of Sciences noted that whilst there is still work to be done to make 'openness' the norm for Chinese Scholars, the increase in relevant legislative policy and training being made readily available, means that more researchers are turning their attention to data management and open data.

The two countries that account for the largest proportions of survey responses are China and the US and with an increase in national mandates, specifically in the US from the Office of Science and Technology Policy (OSTP) and the upcoming Data Management and Sharing Policy from the National Institutes for

Health (NIH), researchers in the US will need to be more engaged with open data than ever before. Whilst the US specifically still makes up 11% of the overall sample, it's important to note that since the survey's inception in 2016 - when looking at engagement on a continental level, North America has been steadily declining, whereas Asia has consistently increased.

Decision factors and motivations

In light of the increase in national mandates and top-down initiatives and legislature, it's key to remember that the responsibility and act of data sharing is more often than not directly in the hands of the individual researcher. When looking at the top three circumstances that would motivate respondents to share their data, the top responses are; citation of their research papers (67%), increased impact and visibility of their papers (61%) and either some form of public benefit or journal/publisher mandate (both 56%).

Whilst there is a strong awareness that open data contributes to 'some form of public benefit', it is the motivation of citations and increased visibility of the individual's research that appears to be paramount. This was a theme picked up by Holly Murray from Health Data Research UK in her contribution to the State of Open Data report, citing a potential 'misplaced motivation' for data sharing.

Building supportive communities

As publishers, libraries and institutions themselves are also the subjects of the aforementioned top down initiatives and mandates, they have an essential responsibility and role to play in the progression and increased adoption of open data practices and principles. In this year's survey, 72% of respondents indicated that they would rely on an internal resource (either colleagues, libraries or Research Offices) were they to require help with managing or making their data openly available. Furthermore, when asked who they would be willing to receive support from, the most popular answer was publishers (41%) closely followed by those within their own institution (38%).

Institutions supporting their researchers in light of top down initiatives, with proficient infrastructure and training is a prevalent theme throughout our report contributions. In particular, Stellenbosch University in South Africa has taken significant steps at the institutional level to ensure they are in a strong position to comply with the core aspects of the national South African open data strategy that were postulated in a proposed national policy.

It's encouraging to see that when compared with 2021 data, this year less researchers said that they would like more direction on how to comply with policies from their institutions, perhaps suggesting that there has been an improvement in the guidance, support and training provided by institutions for open data sharing.

When looking beyond the institutions themselves, it's also clear that the policy makers understand their responsibility to support the research community in complying with their mandates. In their submission to our report, the NIH highlighted the inception of the 'NIH Office of Data Science Strategy (ODSS)' which works to provide 'leadership, strategic guidance, and coordination' for those sharing data under the NIH's plans. It's encouraging to note that in light of the increase in top-down initiatives and mandates, there is a building commitment from policy setters to simultaneously be facilitators for compliance, by providing help and guidance. The establishment of support and guidance directly from funders to comply with their policies is beginning to make a difference to researchers, this year, 36% of respondents said that they were seeking more help on how to comply with funder policies, a not insignificant drop from the 2021 figures (41%).

Explore the full survey results
including the raw data and questionnaire

What circumstances would motivate you to share your data?

n=6,104

Citation of my research papers	67%
Increased impact and visibility of my research	61%
Journal/Publisher requirement	56%
Public benefit	56%
Full data citation	54%
Greater transparency and reuse	52%
Institution / Organisation requirement	49%
Funder requirement	47%
Co-authorship on papers	43%
Consideration in job reviews and funding applications	35%
Direct request from another research	31%
It was simple and easy to do	25%
Freedom of information request	25%
Financial reward	23%
Open data badge	19%
It was a field/industry expectation	16%
Other	2%
I never share my data	3%

The role of policy makers in China: **facilitating the move to open data for researchers and journals**

Yuanchun Zhou and **Lulu Jiang**

Computer Network Information Center, Chinese Academy of Sciences

China started the construction of the Scientific Sharing Project in 2001. In 2018, the General Office of the State Council issued the Measures for the Management of Scientific Data (MMDS), prompting the general deployment of scientific data management on a national level. Policy makers have effectively advanced data management by improving the national legal framework, expanding practice scales, and improving the public recognition of such policies in China.

Comprehensively improving the construction of laws and regulations

The promulgation of MMDS is a landmark event in China. It stipulates that scientific data supported by government budgetary funds should be shared, following the principles of openness as the standard and non-openness as the exception. It also stipulates that data management should follow the principles of 'hierarchical management, safety and controllability.' By the end of 2021, of 34 provinces in China, about 35% of those have issued their Official Detailed Rules for the MMDS. Following the MMDS, China promulgated a series of laws and regulations, involving human genetic resources, biosecurity, data security, etc., forming a legal system for data management and open sharing step by step.

Promoting the expansion of data sharing practices

On February 19, 2019, the Office of Chinese Academy of Sciences (CAS), issued *The Measures for the Management* and *Open Sharing of Scientific Data in CAS*. Of more than one hundred institutes in CAS, over 50% have formulated their institutional policies on scientific data management and that number is constantly increasing.

Apart from the practice in data policies, relevant standards have been released and have boosted the best practice of data sharing in China. As an example, GB/T 32843-2016 Science and Technology Resource Identification, a national standard, proposes a persistent identifier called the CSTR by which shared data can be identified. Equipped with supporting platforms(http://www.cstr.cn), it does not only act as an essential part of infrastructure construction in open data but facilitates scientific data to be shared, in compliance with FAIR principles in China. GB/T 35294-2017 Scientific Data Citation is another critical practice. Combined with GB/T 32843-2016, it means that scientific data can be formally cited and credited.

Furthermore, The Open Science Promotion Consortium (OSPC), initiated by the China Association for Science and Technology (CAST), listed 'Developing Open Data Standards and building an Interchange Mode' as one of primary tasks in its 2022 Work Plan. Predictably, the outputs of this work will further advance the open data practices among academic journals in China.

Rousing recognition of data sharing and awareness of data management

Sharing research data is a revolution that changes research habits, and this requires public recognition and awareness. Beyond policies and standards, data management training is critical. This is a long-time enterprise and policy makers in China undertake the essential responsibility of culture building in the scientific community. In 2022, both CAS and CAST arranged specific training courses for scholars and academic journals to foster data sharing awareness and enhance competency in data management. Through these courses, scholars and editors learned the benefits of data management as well as best practices.

According to the 2022 State of Open Data survey, respondents from China have more training requirements when it comes to developing a practical data management plan. However, the response size from China, accounting for 11% of the overall sample, is significantly larger than that in 2021. In addition, Science Data Bank (ScienceDB), a generalist data repository maintained by CAS, reports that compared with 2021 data there have been 21 times as many depositors from China during 2022 YTD, whilst the growth rate of scholars from CAS is about 2,910% YOY. It would appear that in light of new policies and relevant training being made readily available, more scholars in China are turning their attention to data management.

When compared with international publishers, Chinese academic journals are still in the initial stages of formulating data policies and across the board, there is a general lack of best practices. In the last two years, however, relevant guiding policies have been successfully issued by governments, CAS and CAST, aiming to rouse the recognition and support of Chinese academic journals for open data.

Improving the research evaluation system is another significant step being implemented by the government as part of its ongoing promotion of effective data management. In recent years, research evaluation frameworks have moved away from a sole focus on the published research paper, and there has been an increase in policies proposing non-traditional forms of research outputs are accounted for and considered in the evaluation mechanism.

The Government and the research institutions, at present, are the two primary policy makers for open data in China. They have a leading role in the construction of an open data environment, including legislative works, specific practices, and culture building. With the continued promotion of this work, there will certainly be more organizations in China participating in open data discussions as policy makers and subsequently contributing more to the global open science mission.

The US National Institutes of Health's policies, programs, and partnerships to enhance data discoverability and reuse

Ishwar Chandramouliswaran
Program Director in the NIH Office of Data Science Strategy

Taunton Paine
Director of the Scientific Data Sharing Policy Division, NIH

Amy Hafez
Health Science Policy Analyst at the NIH Office of Science Policy

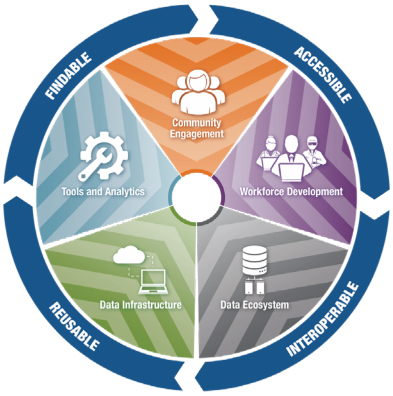
Susan Gregurick
Associate Director for Data Science and the Director of the Office of Data Science Strategy at the National Institutes of Health in the Division of Program Coordination Planning and Strategic Initiatives

Data sharing plays a critical role in the scientific enterprise by promoting data reuse, accelerating research, enabling rigorous testing and validation of research findings, and improving the quality and advancement of science. The National Institutes of Health (NIH), the world's largest funder of biomedical research, recognizes that storing, managing, and standardizing data to enable responsible data sharing facilitates greater public trust through accountability and transparency and ultimately serves to advance NIH's mission. As such, NIH has a long history of policies, programs, and partnerships to enhance data sharing.

> “ *NIH is committed to making results of NIH-funded research publicly available via the new Data Management and Sharing Policy that goes into effect January 25, 2023* ”

On January 25, 2023, NIH's new Data Management and Sharing (DMS) Policy will take effect, reinforcing NIH's commitment to making the results of NIH-funded research publicly available. The new Policy requires all NIH-supported research to 1) have a Data Management and Sharing Plan outlining how scientific data and accompanying metadata will be managed and shared, including any potential limitations on sharing, and 2) be compliant with the NIH-approved Plan. Data management and sharing costs may be requested as part of a budget request, including costs associated with curating data and preserving and sharing data through established data repositories. Through development of a prospective Data Management and Sharing Plan, NIH aims to foster a culture of data stewardship by promoting effective and responsible data management and sharing practices. Last year, NIH also sought public input on helpful resources as well as considerations for the NIH Genomic Data Sharing Policy and will continue to engage the public to inform new data sharing frontiers.

The complexity and volume of basic, translational, and clinical research data generated by NIH-supported research has exponentially increased in recent years. To take full advantage of these data, NIH is modernizing its strategy to better coordinate the collection, storage, analysis, use and equitable sharing of these data to ensure they are discoverable, interoperable, and (re) usable according to FAIR practices as outlined in its Strategic Plan for Data Science. The NIH Office of Data Science Strategy (ODSS) provides leadership, strategic guidance, and coordination for the implementation tactics associated with this plan.



> *“ The NIH has a vision of making NIH funded research data more discoverable, usable, and citable. ”*

These tactics address updating data infrastructure to optimize data storage and connect NIH data systems, including modernizing the data repository ecosystem to support storage, sharing, and use of datasets generated by NIH-supported research, adoption of data management best practices and development of generalizable research software development tools to broaden their utility and impact through improved discovery, as well as upskill the workforce to ensure effective stewardship and sustainability of the various research outputs for the greater good.

The NIH envisions a distributed ecosystem that interconnects data assets, allowing citable stewardship for all relevant research data, with the ability to measure scientific impact through metrics for usage and utility. The NIH strongly encourages use of open access data sharing repositories as a first choice for broad sharing with the research community. In addition to PubMed Central (PMC) for supplementary datasets directly associated with publications, NIH supports open domain-specific repositories and knowledgebases for structured data of a specific type or related to a specific discipline or area of science and cloud-based solutions via the STRIDES and related programs for petabyte-scale data. However, recognizing the fact that not all data have domain-specific repositories, generalist repositories serve to fill this gap through the Generalist Repository Ecosystem Initiative (GREI), allowing the self-publishing of data regardless of type, format or subject matter.

To support the longer term provenance, findability, and citation of NIH-supported data, NIH has recently joined DataCite as a consortium member. This will serve all

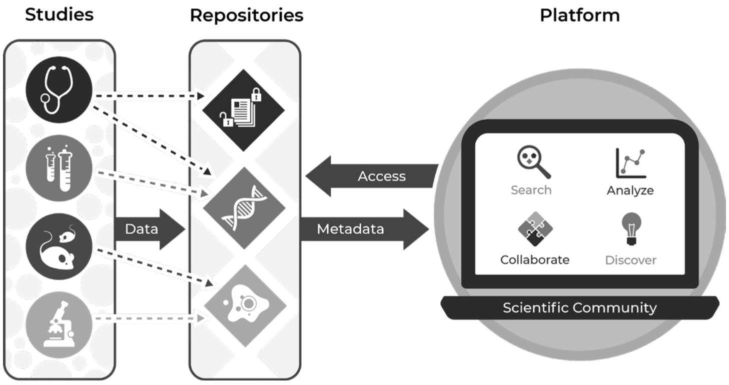


Figure created by Josh Terrell, Graphic Designer at the Renaissance Computing Institute (RENCI)

NIH Institutes, offering the assignment of persistent identifiers to research outputs (e.g., data, software, reports) while providing coordinated technical and administrative support for NIH-designated data repositories preserving these outputs. This partnership is expected to have significant impact in streamlining metadata collection and its use for search, developing metrics to measure value of data, and developing best practice guidelines in use of persistent identifiers by the research community in the preservation and citation of digital objects.

> *“ The NIH strongly encourages use of open access data sharing repositories as a first choice for broad sharing of NIH-funded data. ”*

NIH is interested in establishing partnerships with communities, societies, and external programs to enhance the education, adoption, and implementation of FAIR practices through collaborative projects, workshops, and other activities. Current examples include the network of the National Libraries of Medicine (NNLM) activities to enable data-driven research, partnership with the Data Curation Network (DCN) to better train stakeholders in data curation best practices via training events, partnership with the Federation of America Societies for Experimental Biology (FASEB) on the DataWorks! Prize, and other activities to engage and incentivize the research community in better data sharing and reuse practices. Other ideas for exploration include enhancing infrastructure resources to support a research 'data-mesh' that provides services for data access and use while reducing siloes, developing knowledge graphs to enable specialized search and "recommender" systems, dynamically respond to new and emerging data types/disciplines, and encouraging the adoption of open data and metadata formats to enable true discoverability of data and related digital assets.

Importantly, with the recent release of the White House Office of Science and Technology Policy (OSTP) memorandum, "Ensuring Free, Immediate, and Equitable Access to Federally Funded Research," NIH will continue working across federal departments and agencies to implement the updated US policy guidance and engage in global partnerships. Recognizing data stewardship is a global endeavor, NIH is an active member and one of seventeen agencies in the United States to embed a national data strategy from OSTP and the National Science and Technology Council (NSTC) Subcommittee on Networking and Information Technology Research and Development (NITRD) Program's revisions to the Federal Big Data Strategic Plan currently underway.

NIH continues to invest in a robust data ecosystem, one where its policies and activities keep pace with evolving scientific technologies, opportunities, and stakeholder expectations. NIH is also aware that not just technological advancement, but behavioral change is also necessary to advance data science goals and is open to establishing new partnerships to better enhance equitable data discoverability and reuse of NIH-funded research data and turning data-driven discovery into health at an accelerated pace.

For more information, please see https://datascience.nih.gov or contact datascience@nih.gov.

Preparing for South Africa’s proposed open data strategy — Lessons from Stellenbosch University

Samuel Simango
Manager: Research Data Services
Stellenbosch University

In this article, Samuel Simango discusses how Stellenbosch University ensures they are compliant with the core aspects of the national South African open data strategy that were postulated in the Proposed National Data and Cloud Policy.

Proposed National Data and Cloud Policy

In April 2021, the Department of Communications and Digital Technologies published an invitation for the public to submit written submissions on a Proposed National Data and Cloud Policy (hereafter referred to as the ‘proposed policy’). The purpose of this document is to enable South Africans to realize the socio-economic value of data through the alignment of existing policies, legislation and regulations.

The proposed policy acknowledges that during the earlier days of the digital economy, South Africa enacted several data-related statutes. However, since the realities of data in a digital economy were not sufficiently appreciated at that time, the legislation now falls short, in light of the multiple technological developments that we have observed. This has led to a lacuna in the system which the proposed policy seeks to address.

Among the different interventions proposed in the National policy, there are two that are particularly relevant to the future state of open data in South Africa:

1. The development of an open data strategy for the sharing of data that is informed by ‘Data for Good’ principles.
2. The application of the FAIR Data Principles to South Africa’s open data.

‘Data for Good’ principles

Broadly speaking, ‘Data for Good’ principles promote the production and use of data for the advancement of the social good. This implies the production and use of data to advance society or social causes without having regard to financial gain. This interpretation aligns with the use of the term in the proposed policy.

Within the context of the proposed policy, ‘Data for Good’ principles are understood as access to data that is provided to non-governmental organizations without requiring payment for such access.

FAIR Data Principles

The FAIR Data Principles, which were first published in 2016, stipulate certain criteria for generated data, namely that such data should be findable, accessible, interoperable, and reusable. The overall aim of the FAIR Data Principles is to facilitate knowledge discovery by assisting humans and machines in their discovery of, access to, integration and analysis of, task-appropriate scientific data and their associated algorithms and workflows.

Stellenbosch University and the Proposed Policy

Although the aims of the proposed policy are still aspirational at the moment, it is worth reflecting on the manner in which an institution such as Stellenbosch University already strives to promote both the 'Data for Good' as well as the FAIR Data Principles.

Stellenbosch University has already taken steps that address both the 'Data for Good' and the FAIR Data Principles. The university has achieved this outcome through its Figshare-powered repository - better known as SUNScholarData. SUNScholarData is a multidisciplinary institutional research data repository that was launched in August 2019. The repository is used for the registration, archival storage, sharing and dissemination of research data produced or collected in relation to research conducted under the auspices of Stellenbosch University. SUNScholarData creates a medium through which Stellenbosch University's research data can be made findable and accessible. It also facilitates the interoperability and re-usability of the university's research data.

Overview of SUNScholarData

Research data are not automatically published on SUNScholarData. After submissions for the publication of research data have been made by researchers the submissions are then subjected to a curation process that is managed by dedicated staff members at Stellenbosch University's Library and Information Service. During the curation process submissions are appraised and then subjected to metadata enrichment.

Data appraisal

The appraisal of research data serves the purpose of assessing the suitability of submissions for publication on SUNScholarData. This entails reviewing several aspects relating to submissions such as: the nature of the data, security of the files, file formatting, file organization, the presence of appropriate data documentation and the accuracy of the metadata fields completed by users in order to describe their research data. In addition to this, the data appraisal helps to verify whether or not the publication of research data would give rise to disclosure risks.

Assigning and managing metadata

Following their appraisal, the research data are assigned administrative metadata. In addition to this, the descriptive metadata are inspected for their appropriateness. The descriptive metadata are also enriched further through the use of a controlled vocabulary and the application of the widely used domain-agnostic DataCite Metadata Schema.

Facilitating access to research data

Research data are published and made available via a publicly accessible information retrieval interface that facilitates browsing and searching. The research data are assigned an open access setting thereby making them openly accessible without restriction. Furthermore, each of the data files are assigned a digital object identifier (DOI) that is used to uniquely and persistently identify the data files as well as to resolve to a digital landing page for the respective files.

Reuse of content

Different forms of reuse pertaining to SUNScholarData's content are permissible subject to attribution. This is ensured through the use of open licenses such as the Creative Commons Attribution license (CC BY) license or the Open Data Commons Attribution license (ODC-By). Open software files are made available under one of a variety of standard Open Source software licenses such as the following: MIT, GPL, GPL 2.0+, GPL 3.0+ and Apache 2.0.

To conclude, prior to the publication of the Proposed National Data and Cloud Policy, Stellenbosch University had already taken steps to make its research data openly accessible where possible. The resource that the university relies on for such purposes – SUNScholarData – ensures that the university adheres to the 'Data for Good' principles as well as the FAIR Data Principles. As such, when South Africa's open data strategy is eventually finalized, Stellenbosch University will already be in a relatively good position to support the implementation of the strategy.

70% of respondents said they were required to follow a policy on data sharing for their most recent piece of research

Infrastructure needs of researchers for open data
a Latin American perspective

Juan Miguel Palma Peña

Academic Librarian and Lecturer at National Autonomous University of Mexico (UNAM); Doctor in Library and Information Studies, National Autonomous University of Mexico (UNAM)

Scholarly Communications and Infrastructure for Open Data: Context

Scholarly communications and the pathways towards open science demand infrastructures that enable free access to the data and the methodology used to generate the research and, later, its publication.

Therefore, the infrastructure is a core component of opening access to data – crucial to the dissemination, visibility and open access to research processes, data and outputs. Currently there is a lot of activity towards the development and implementation of open access data platforms that are findable, accessible, interoperable and reusable.

The “Budapest Open Access Initiative” (2022) has recommended to host and publish data, metadata and other digital research results in open infrastructures managed by academic communities. It also suggests the use of infrastructures shared on the basis that those citizens who use repositories as readers will see the point of depositing as authors.

International studies and surveys which have proved that infrastructure is necessary to open data have also evidenced advances in the development of platforms and metadata for availability, visibility and open access. Likewise, many academic communities have expressed interest and familiarity with open science, the FAIR principles and the need to promote and expand the visibility of their scholarly outputs for attribution and academic contribution, among other reasons.

The "UNESCO Recommendation for Open Science" (2021) suggests that infrastructures for open data need to be sustainable, shareable, interoperable and harmonize with FAIR principles to locate, have permanent, barrier-free and non-profit access to data and information to support the needs of communities.

In this context, it becomes relevant to study which open data infrastructures are being implemented by institutions within the Latin American region.

Infrastructure for Open Data from Latin America: Analysis and Findings

Latin America has a long tradition of action towards open access. Open data project initiatives in different disciplines have gradually been undertaken. This is because both the entities that finance open access (such as governments, higher education institutions and research centers) and the academic community are interested in free access to open data - openness provides increased visibility of academic outputs and supports teaching, research and dissemination activities, among other benefits.

Current projects on open data in the Latin American region are scattered all over the web. A mixed research study which included searching on libraries official web portals, the Registry of Research Data Repositories and Dataverse Project was carried out to identify the main actions regarding infrastructures for open data developed and implemented in the region, The study was based on defined variables such as: actions, data repositories; infrastructure; library collaboration; repositories in re3data and repositories in Dataverse. The findings obtained are presented below.

Findings

The results retrieved came from nine countries that have repository initiatives defined within the study, in official institutional web portals. These are presented below by country, in alphabetical order.

For those countries from which information has not been retrieved the assumption is that they have not developed, implemented or documented their actions.

Argentina. Argentina has seven data repositories registered in re3data. These include the National System of Digital Repositories, Genomic Data Portal, and Biodiversity Data Portal, among others. The technological infrastructures used are DSpace and Eprints. Open data actions are supported by the university libraries that developed these repositories.

Brazil. Brazil has sixteen data repositories registered in re3data and five in Dataverse. These include Research Data Repository of the University of Campinas, Scientific Database of the Federal University of Parana, and Biodiversity Research Program Data Repository, among others. The main technological infrastructures used are DSpace and Dataverse. Collaboration of libraries is with SciELO Data.

Colombia. Colombia has ten data repositories registered in re3data and one in Dataverse. These include Intellectum of the University of La Sabana; Colombian Biodiversity Information Facility; and Research Data Repository of Universidad del Rosario (2016), among others. The technological infrastructures implemented are DSpace and Dataverse. Collaboration is within Colombian library networks.

Costa Rica. The National Council of Rectors signed a collaborative agreement with the Research Data Alliance in 2021 for develop research data actions (Solano, 2021). The country does not refer to data repositories. A library collaboration manages "Kimuk: national repository" (Arturo Argüello Chaves Library, 2021).

Chile. Chile has two data repositories registered in re3data and one in Dataverse. The Digital Repository of the National Research and Development Agency (ANID, 2020) and the Research Data Repository of the University of Chile. The technological infrastructure is Dataverse. Collaboration of libraries is via SciELO.

Ecuador. Ecuador has a one data repository registered in Dataverse - InData, the Ecuadorian Corporation for the Development of Research and Academia. The technological infrastructure used is Dataverse. No collaboration with libraries could be retrieved.

Mexico. Mexico has thirty data repositories registered in re3data and one in Dataverse. The main open data actions come from the National Autonomous University of Mexico (UNAM): Open Data Portal: University Collections (UNAM, 2015 and the Center for Data and High-Performance Computing of the Institute of Nuclear Sciences, among others. The technological infrastructures used are DSpace, Dataverse, Eprint and MySQL. Collaboration between libraries promotes free access to information.

Panama. Panama has two data repositories registered in re3data, from the Center for Tropical Forest Science; and Smithsonian Tropical Research. The technological infrastructure used is locally owned and no information related to collaboration between libraries could be retrieved.

Peru. Peru has three data repositories registered in re3data and one in Dataverse. These include Open Data Portal of Pontiphal Catholic University of Peru (PUPC, 2019) and the Repository of Open Data of the Ministry of Education. The technological infrastructures used are DSpace and Dataverse. Collaboration of libraries is based on the portal implemented within the library system of the university.

Challenges and opportunities

The findings allow us to conclude that open data progress in Latin America is gradual. This might be because the implementation of infrastructures requires a set of measures to guarantee harmonization of regulations.

Open data initiatives and projects in the Latin America region are scattered on the Internet, and it may be relevant to undertake a “Latin American Open Data Repository Infrastructures” research project to compile actions, initiatives, regulations, services and tools on open data from / and for the region.

References

ANID. (2020). Propuesta de Política de acceso abierto a la información científica y a datos de investigación financiados con fondos públicos de la ANID. Chile: Agencia Nacional de Investigación y Desarrollo (ANID) https://s3.amazonaws.com/documentos.anid.cl/estudios/Politica_acceso_a_informacion_cientifica_version_final_26-05-2020.pdf

Biblioteca Arturo Agüero Chaves. (8 julio 2021). Kimuk Repositorio Nacional de Costa Rica. Obtenido de archivo de video. https://www.youtube.com/watch?v=nysQCy9AGsI

BOAI (2022). The Budapest Open Access Initiative: 20th Anniversary Recommendations. https://www.budapestopenaccessinitiative.org/boai20/

Dataverse Project. https://dataverse.org/

PUCP. (2019). Portal de Datos Abiertos. Perú: Pontificia Universidad Católica del Perú (PUCP). https://datos.pucp.edu.pe/

Registry of Research Data Repositories. https://www.re3data.org/

Solano, V. (2021). Costa Rica fortalece la gestión de datos de investigación, las visiones de acceso abierto y ciencia abierta. Costa Rica: Consejo Nacional de Rectores. https://www.conare.ac.cr/noticias/282-costa-rica-fortalece-la-gestion-de-datos-de-investigacion-las-visiones-de-acceso-abierto-y-ciencia-abierta

UNAM. (2015). Lineamientos para la Integración y Publicación de las Colecciones Universitarias Digitales en el Portal de Datos Abiertos UNAM Colecciones Universitarias. Gaceta UNAM, 4727, 24-28. https://datosabiertos.unam.mx/informacion/docs/GA_D_LI_CCUD_20150924_Integracion_Publicacion_Colecciones_PDA_UNAM.pdf

UNESCO. (2021). Recommendation on Open Science. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en

Universidad del Rosario. (2016). Portal Institucional de Ciencia Abierta. https://cienciaabierta.urosario.edu.co/?_ga=2.6183541.73949541.1625717635-999329532.1625717635

Data saves lives

Holly Murray
Research Manager, Health Data Research UK

Health data research has the capacity to transform healthcare and improve lives, both now and in the future. As the UK's national institute for health data science (an emerging discipline, combining mathematics, statistics, epidemiology and informatics), Health Data Research UK's (HDR UK) 20-year vision is for large scale data and advanced analytics to benefit every patient interaction, clinical trial, biomedical discovery and enhance public health.

The COVID-19 pandemic has demonstrated that data does indeed save lives. Expedient and trustworthy access to linked health, social and care data underpinned the UK's policy response and enabled the research into the virus, vaccines and treatments that have ultimately helped control the pandemic.

Clinical trials including RECOVERY and PRINCIPLE, enabled by HDR UK, both applied innovative use of routinely collected health data to discover the first effective treatments in record time; rapid linking of genomic and population data enabled near real-time monitoring of virus variants; and large-scale population-wide datasets made rapid evaluation of the vaccine effectiveness possible.

Of course, these benefits do not (and must not) stop with COVID-19. The ability to collect, link, access and ultimately use health data for research is critical and yet complex.

The balancing act

In many ways, health data is similar to research data generated in other disciplines, and involves the same complex issues around interoperability, curation, processing, analysis and long-term storage. And ideally needs to be findable, accessible, interoperable, and reusable (FAIR).

Unlike other data, however, health data reflects real people; where each data point collected from electronic health records, group studies, blood or tissue samples, imaging, wearable devices and more is a moment in a person's life – from birth to death.

It's not surprising to see that 'concerns about their data containing sensitive information or requiring specific permissions' was a particular consideration for the State of Open Data survey respondents in Medicine & Health Sciences. As such, health data research must balance potential conflicts between sharing and patient privacy.

Whilst the often-used mantra in research data management 'as open as possible, as closed as necessary' goes some way to address this and lends space to anonymisation and controlled access (for example, as facilitated by the HDR UK Innovation Gateway and trusted research environments), what's missing – and what I would like to focus this conversation on - is the need to balance research, patient, and broader societal interests.

Working together

Involvement and engagement of patients and the public is vital for achieving this balance, through building trustworthiness in health data research and understanding public priorities.

Evidence shows that involving patients and members of the public in research and service development results in an increase in the quality and relevance of research studies, and helps in securing funding, designing study protocols and choosing relevant outcomes (Blackburn et al 2018).

Alongside governance and security, patient and public involvement and engagement (PPIE) is critical to each of HDR UK's activities spanning health data science including data access and analysis. From representation on executive and scientific committees, to citizen deliberation on data access requests, and involvement in the shaping of research projects, PPIE is changing the culture of health data science and influencing how data is requested, accessed, communicated, and used.

No one left behind

Beyond involving the public and patients, the data at the core of health data research must reflect the diverse range of people it intends to benefit. Unfortunately, most health datasets do not adequately represent minority groups – often because sampling is not inclusive or the quality of the data about individual characteristics is poor (that is, inconsistent, incomplete, or inaccurate). This in turn has the potential to exacerbate inequalities. For example, almost 80% of people in genome wide association studies are of European ancestry – despite people of European ancestry making up only 16% of the global population (Martin et al 2019). The bias means polygenic risk scores, for example, are less accurate in non-European populations.

Whilst alarming, recent initiatives to raise awareness and begin tackling the fundamental problem of diversity in health data provide some promise. The HDR UK community has made progress: convening data custodians to explore challenges in ethnicity coding; developing standards that ensure datasets for training and testing AI systems are diverse, inclusive, and promote AI generalisability; and creating informatics tools to promote ethnic and gender equality in genomic medicine. Other steps forward include (but are not limited to) Benevolent AI's data diversity analysis tool and the Data Science for Health Equity community.

The end-goal is research that benefits as many people as possible - based on data that reflects diversity of culture, healthcare conditions and aspects such as race, ethnicity, gender and age that provides fairer and more equal access to the latest treatments and medical technologies.

Misplaced motivation?

A key takeaway from this year's Report on Open Data is that researchers' top motivations for data sharing are citation of their research paper (20%), co-authorship on papers and public benefit (both 12%).

Whilst it is promising to see public benefit as a key motivator, one is led to question why public benefit is less motivational than citation? Is this a reflection of an evaluation culture centered on metrics which promote competition? Self-interest vs altruism? And/or a call for further work to be done to ensure that the public and patients are centered at the heart of research?

The HDR UK community includes a range of researchers within both health and computation roles from research organizations across the UK, and recognising the discipline-specific academic pressures on researchers is itself an important factor when thinking about the drivers of FAIR data and open science.

Still, public and patient benefit for all should be seen as the preeminent goal of health data research, FAIR data, and open science more broadly.

Understanding and supporting data sharing in the humanities: new insights from a publisher survey

Matthew Cannon Head of Open Research, Taylor & Francis
Dr. Rebecca Grant Head of Data and Software Publishing, F1000
Kate McKellar Publisher, Social Sciences and Humanities, Wiley

Introduction

In this chapter, we report on the results of a survey of over 400 humanities researchers, which assessed data sharing practices as well as the experiences and attitudes of humanities scholars in relation to data sharing. The results complement the findings of the 2022 State of Open Data survey, focusing on the specific challenges and opportunities associated with humanities research data including terminology, methods of sharing, and the availability of targeted support.

The survey was developed, distributed, and analyzed by members of the STM Association's Research Data Program Humanities Sub-group, comprised of representatives from academic publishers Taylor & Francis, Routledge, Wiley, F1000, SAGE, Brill, Oxford University Press, and Cambridge University Press. The STM Association's Research Data Program is a publishing industry initiative aiming to better align data sharing policies across stakeholders, and to encourage the uptake of stronger data sharing policies and data availability statements by academic publishers and journals.

Early in the formation of the Humanities Sub-group, it became apparent that although researchers in medicine, life, earth, and natural sciences are relatively well served by existing journal research data policies and associated guidance, additional consideration would be needed for the humanities. As representatives of academic publishers, the group members were aware that humanities journals are currently less likely to have a policy requiring authors to share data openly, or to include a data availability statement with their article. It was also clear that reusing existing scientific data policy wording would not be appropriate: while replicability of results is at the heart of many research data sharing policies, this is not necessarily relevant to humanities research practice. Nevertheless, as publishers we believe that data sharing is relevant for the humanities, and that it can support transparency of research methods, allow others to build on published work, and provide credit to the data creator in the form of citations.

Humanities scholars work with a variety of sources, including documents, moving images, audio files, maps, photographs, and other physical or digital artifacts. There is evidence that they have a preference for analogue sources over digital, accessed via libraries, archives, and museums, and that they read printed versions of materials when possible (De Gruyter Report, "A Day in The Life—Insight into the six phases of the HSS researcher workflow in Germany, Austria and Switzerland," 2022). Academic publishing in the humanities has reflected these trends with generally a slower transition to electronic manuscript handling systems to conduct peer review, and on-screen proofing and editing tools than science counterparts. These differences in research methods and publishing practices further demonstrate why standard scientific policy wording is not relevant or appropriate for humanities scholars.

To ensure the policy and guidance that is developed to support humanities data sharing is based on input from humanities researchers and appropriate for their needs, the STM Humanities Sub-group designed a survey aiming to capture humanities scholars' current knowledge, attitudes, and experiences of working with, generating, and sharing "data." The survey ran for four weeks in Spring 2022, circulated via social media and targeted email, and received 422 responses. A selection of the quantitative responses captured by the survey is described below, and analysis of the qualitative responses is planned for late 2022.

Survey results

The results of our survey complement the findings of the larger scale State of Open Data survey, delving deeper into the specific experiences of researchers in the humanities. The survey respondents were predominantly located in Europe (56%) and North America (20%), with smaller numbers from Asia (9%), Australasia (5%), the Middle East (4%) and Africa (2%). Over three quarters stated that they had been active researchers in the humanities for more than 10 years, and 49% had published more than 20 academic articles. Given the self-selecting nature of the humanities survey respondents, the results cannot necessarily be generalized across all humanities researchers in all regions, but provide a snapshot of attitudes and practices which have not previously been captured.

In analyzing the responses, a key finding is the extent to which humanities scholars believe that "research data" is a term which is applicable to them and their research practice; while 52% find the term appropriate, nearly half believe that it does not apply to their work either some or all of the time. Other preferred terms suggested by respondents include "research materials," "information," "evidence," and "sources" (figure 1). This lack of consensus indicates that the development of data sharing policies and guidance (or future surveys) aiming to address humanities researchers should consider the most appropriate terminology to be used.

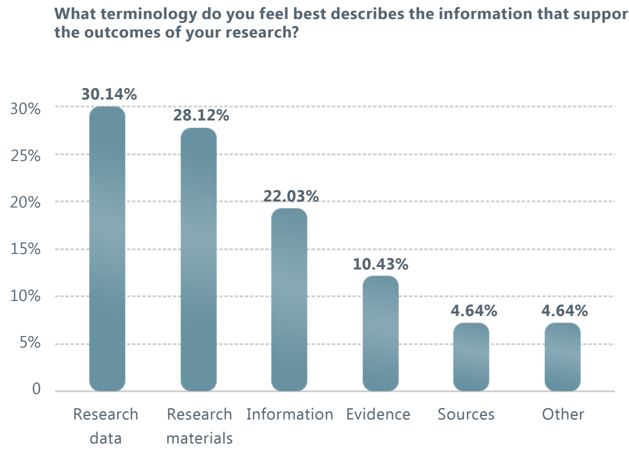
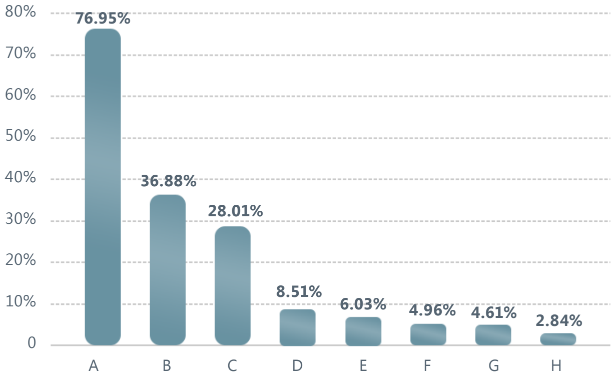


Figure 1: Preferred terminology to refer to outcomes of humanities research (n=345)

In spite of the variance in terminology, a very high proportion of respondents (88%) believed that humanities research data should be shared with others. This is higher than the cross-disciplinary sample from the State of Open Data survey, where 79% agreed that making data openly available should be common practice. Although the humanities respondents were keen to share data, the majority rely on peer-to-peer sharing methods (76% shared data by email) with only 36% sharing via a data repository, which would provide long term preservation and persistent identifiers for citation, representing best practice for data sharing (figure 2). The State of Open Data survey indicates that only 19% of cross-disciplinary respondents felt that they received sufficient credit for having shared their data. Based on low uptake of repositories in the humanities, there is a risk that the percentage of humanities researchers receiving appropriate credit would be even lower, as there is no formal mechanism for measuring and rewarding the ad-hoc sharing methods they use.

Have you ever shared your research data with others?
Select all that apply



A. Yes - by personal/email transfer

B. Yes - in a data repository or online database

C. Yes - on a website

D. No - I didn't have permission

E. No - it didn't occur to me

F. No - I didn't think anyone would find it useful

G. No - I didn't want to

H. I'm not sure

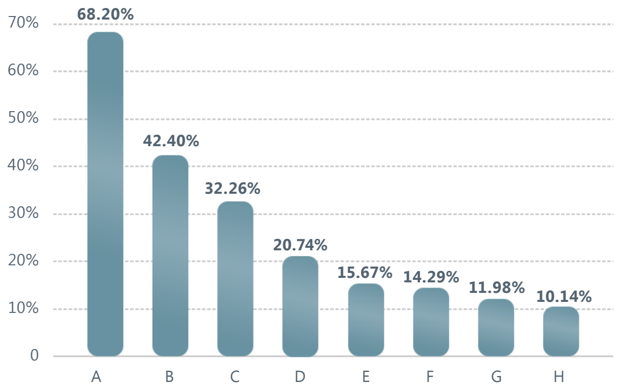
Figure 2: Reported methods of sharing humanities data (n=282)

When it comes to motivation for sharing, only 15% of respondents believed that journal publishers should mandate data sharing as a condition of publication, with 68% believing that data sharing should be a decision made by researchers themselves. In contrast, 56% of the cross-disciplinary State of Open Data respondents reported being motivated by publishers' data sharing requirements. An issue here could be familiarity, as a mandate to share data as a condition of publication is rare for humanities journals but more common in STM disciplines, as previously noted.

When asked for additional feedback on why publishers should not mandate data sharing a number of concerns were stated, reflecting those reported in the State of Open Data survey in previous years.

Humanities researchers are concerned that their data may be misused by others, and are unsure about copyright and licensing, as well as having broader objections around the relevance of data sharing to their research practices (figure 3). If these concerns are widespread in the community, it could explain why private, peer-to-peer sharing practices are still the most prevalent.

Why should publishers not require authors to share data?
Please select all that apply



A. It's not relevant to my field

B. It should be up to authors to choose

C. I don't wish to share that part of my research

D. Concerns about misuse of data

E. Unsure about copyright/data licensing

F. Unsure of the permission requirements from my funder/institution

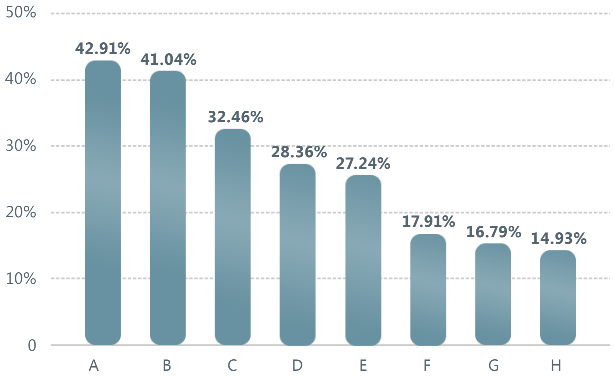
G. Unsure about which repository to use

H. Other (please specify)

Figure 3: Reasons why publishers should not mandate data sharing as a condition of journal publication (n=217)

To address data sharing concerns and provide support, more training might be necessary: 80% of respondents stated that they had never received training on data sharing. When asked what additional support publishers could provide, guidance on selecting a suitable data repository was the most popular response (43%); in addition, 41% of respondents would like publishers to collaborate more closely with other stakeholders (institutions and libraries) to provide support for data management and sharing (figure 4).

What additional support from academic journals and publishers would you find most useful as a researcher of humanities research?
Please select all that apply



A. I think that publishers should introduce stricter Data Sharing Policies to ensure that researchers in the humanities are making their data available to others

B. I would like to see publishers collaborating more with institutions and libraries to offer more support to researchers in managing and sharing their data

C. I would like publishers to be more involved in Research Data Management and data sharing from an early stage in the research

D. I would like to receive written information on the benefits of Research Data Management and data sharing (e.g. a handbook, online toolkit resources)

E. I would like to receive information on which repositories I should use for my subject area

F. I would like to receive data sharing training from a publisher

G. I would like to receive Research Data Management training from a publisher

H. I would like more information on what my "research data" consists of

Figure 4: Suggested additional support for data management and sharing which could be provided by academic publishers, excluding "other" (n=268)

A high proportion of respondents reported reusing data shared by other humanities researchers - 61% had done so. This is comparable to cross-disciplinary data reuse reported prior to the pandemic, with 58% of surveyed researchers stating that they reuse data shared by others. The main method of accessing the data was by contacting the data creator directly (27% of respondents had done this), although some accessed data by searching online (14%), through a data repository (13%) or via links found in a research paper (10%). While attitudes to data sharing and reuse are clearly positive overall, there is potentially a gap in information, education and support to encourage humanities researchers to use data repositories, and to access the potential rewards in the form of data citations and the additional benefits of long-term preservation and enhanced accessibility.

Conclusions and next steps

In reviewing the survey responses, it is clear there are a number of relevant factors when looking to increase the availability of humanities research data and strengthen journals' data sharing policies. Terminology is key, as "research data" is not a universally recognised term: over half of the respondents felt it did not apply some or all of the time. There is evidence that humanities data is being shared frequently between peers and colleagues, but using informal sharing mechanisms like email and file transfer. Concerns around controlling the reuse of data, or confusion around copyright and licensing, are potentially reducing open sharing via data repositories.

Also evident in the responses is a lack of access to support for data sharing or guidance on where to find more information, and potentially also a lack of understanding of why and how data sharing could be beneficial to both for humanities research and the research community. In considering what guidance might be appropriate, key issues to address include copyright and licensing, the benefits of sharing in repositories, and how physical sources or collections can be described transparently in the context of data availability.

As publishers, we are committed to supporting researchers in the handling and sharing of their research data, and ensuring they get appropriate reward and credit for their work. We plan to use the information we have gathered to create a set of resources to help publishers support humanities authors and develop appropriate data sharing policies, and to demonstrate the benefits of proactive sharing practices. The findings shared here will be enhanced through the analysis of the qualitative responses received via this survey, and we intend to publish further findings in due course.

An anonymised version of the survey data is available from https://doi.org/10.6084/m9.figshare.21207239 under a CC-BY license.

Author biographies

Matt Cannon

Matt Cannon is the Head of Open Research for Taylor & Francis. Matt has been working at Taylor & Francis for almost 15 years in a variety of editorial roles, in both science and social science areas. In 2019 Matt moved to the Open Research team where he sets policies and practices to improve the reproducibility and transparency of research. Matt is a member of the Research Data Alliance and sits on the publisher advisory board of FAIRSharing.org and the Centre for Open Science.

Ishwar Chandramouliswaran

Ishwar Chandramouliswaran is a Program Director in the NIH Office of Data Science Strategy and technical lead for the strategy, planning, coordination, and oversight of trans-NIH programs to establish a FAIR, modernized, integrated data ecosystem at the NIH. In his role to foster partnerships and develop a global data strategy, he is a NIH representative in the Big Data Interagency Working Group of the National Science and Technology Council's (NSTC) Subcommittee on Networking and Information Technology Research and Development (NITRD) and also serves on the Editorials Boards of ELIXIR Europe's RDMKit and FAIR CookBook, both resources for FAIR data management and stewardship practices .

iD https://orcid.org/0000-0001-9697-9599

Laura Day

Laura is the Product Marketing Manager at Figshare, having previously worked across Digital Science at Altmetric and Dimensions and was previously working in Academic Publishing. Laura has a keen interest in taking academic research beyond academia and reaching wider audiences.

Dr. Rebecca Grant

Dr. Rebecca Grant is Head of Data & Software Publishing at F1000, where she supports the development of policy, guidance and publishing methods to encourage researchers to share open and FAIR research data. She has a background in data management for the humanities and social sciences and was previously based at the Digital Repository of Ireland and the National Library of Ireland. She is co-chair of the STM Association's Research Data Program Humanities Data Sub-group, and the Research Data Alliance Data policy standardization and implementation Interest Group. She is a qualified Open Data trainer certified by the Open Data Institute. Her doctoral thesis explored the connections between archival theory and research data management practice, using Ireland as a case study.

Dr. Greg Goodey

Greg Goodey is a Data Analyst at Springer Nature. He is responsible for managing projects to collect and analyse information on customers, markets, products and communications within the STM market. He plays a major role in coordinating the annual State of Open Data survey where he manages development of the survey and the analysis of the outcomes. Greg completed his Ph.D. in Physiology at UCL and has since held research roles in a number of industries joining Springer Nature in 2016.

https://orcid.org/0000-0002-1541-6805

Susan Gregurick

Susan Gregurick is the Associate Director for Data Science and the Director of the Office of Data Science Strategy at the National Institutes of Health in the Division of Program Coordination Planning and Strategic Initiatives. Under Dr. Gregurick's leadership and guided by the NIH's Strategic Plan for Data Science, ODSS leads and coordinates catalytic data science activities in scientific, technical, and operational programs in collaboration with the NIH institutes, centers, and offices. She also serves as the NIH representative to the OSTP NITRD subcommittee and the representative to RDA, as well as a member of advisory committees to DOE and NSF and other agencies in Canada. Dr. Gregurick received the 2020 Leadership in Biological Sciences Award from the Washington Academy of Sciences.

https://orcid.org/0000-0002-1074-7576

Amy Hafez

Amy Hafez is a Health Science Policy Analyst at the National Institutes of Health in the Office of Science Policy. She works with the Scientific Data Sharing Policy Division to monitor research and the science policy ecosystem to develop biomedical research policy in the scientific data management and genomics and health areas. She recently served as a Science and Technology Policy Fellow through the American Association for the Advancement of Science and a Congressional Science Policy Fellow through the American Chemical Society. Dr. Hafez earned a Ph.D. in Molecular Genetics and Microbiology from Duke University.

iD https://orcid.org/0000-0002-2529-0177

Dr. Mark Hahnel

Mark 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.

iD https://orcid.org/0000-0003-4741-0309

Lulu Jiang

Lulu Jiang, the product manager of Science Data Bank, is working in the Computer Network Information Center, Chinese Academy of Sciences. Her current research interests include scientific data management and data publishing. She has engaged with ScienceDB since 2018 and has a research background in digital publishing.

Kate McKellar

Kate McKellar is a Publisher at Wiley for Social Sciences and Humanities, specializing in Humanities and Linguistics. She has a background in Humanities scholarship, with an MA in Contemporary Literary Studies from Lancaster University, and worked previously at Oxford University Press. She has recently launched the Gold Open Access journal Future Humanities, which highlights the rise and convergence of new and critical Humanities by publishing trans- and inter-disciplinary research focused on their diverse subjects and methodologies, including data reports and null findings. Along with Dr Rebecca Grant and Matt Cannon, she is co-chair of the STM Association’s Research Data Program Humanities Data Sub-group.

Holly Murray

Holly currently works as Research Manager at Health Data Research UK where she leads on research culture, open science, and impact initiatives. She is an open research advocate and is involved in several professional organizations, working groups, and boards dedicated to making science more transparent and ensuring all outputs are valued. Holly holds a PhD from the National University of Ireland Galway, and previously was Head of Data and Software Publishing at F1000.

Taunton Paine

Taunton Paine is the Director of the Scientific Data Sharing Policy Division. Taunton leads a team responsible for a broad array of issues relating to data sharing and access, including the NIH Policy for Data Management and Sharing, the NIH Genomic Data Sharing Policy. Taunton also works with NIH senior leaders in identifying scientific data sharing issues that require trans-NIH coordination. Before becoming director of the division, Taunton led the clinical research policy team within the OSP Division of Clinical Research and Healthcare Policy. Taunton began his OSP career in 2011 and first worked on issues relating to biosecurity and dual use research of concern. He holds a dual master's degree from Columbia University and London School of Economics and Political Science, where he studied the history of international relations.

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

Juan Miguel Palma Peña

Juan Miguel Palma Peña is an Academic Librarian and Lecturer at National Autonomous University of Mexico (UNAM) and Doctor in Library and Information Studies, National Autonomous University of Mexico (UNAM). He is a member of the National System of Researchers of National Council of Science and Technology (CONACyT), Mexico; He is a Standing Committee Member of the Serials and Other Continuous Resources Section (SOCRS) of the International Federation of Library Associations and Institutions (IFLA).

iD https://orcid.org/0000-0002-6292-4511

Samuel Simango

Samuel Simango is the Manager: Research Data Services at Stellenbosch University's Library and Information Service. In his role, he has been involved in the conceptualisation and implementation of several research data management initiatives at the University such as: drafting governance documents related to research data, managing the institutional research data repository and providing data management planning as well as research data support services. In addition to this, he was involved in the Open Science component of the ilifu big data infrastructure for data-intensive research and served as one of the developers of the Research Data Management Adventure Game.

iD https://orcid.org/0000-0003-2342-6787

Yuanchun Zhou

Yuanchun Zhou is Research Professor of the Computer Network Information Center, Chinese Academy of Sciences. His work focuses on scientific data management and data sharing and data Intelligence. He has over 100 scientific publications in international conferences and journals.

Part of **DIGITAL** science

Altmetric

Dimensions

figshare

ifi CLAIMS

Overleaf

readcube

ripeta

scismic

SYMPLECTIC
Elements

SYMPLECTIC
Grant Tracker

writefull

DIGITAL science
Consultancy

digital-science.com

10 years of figshare

Available online also:

knowledge.figshare.com/state-of-open-data