



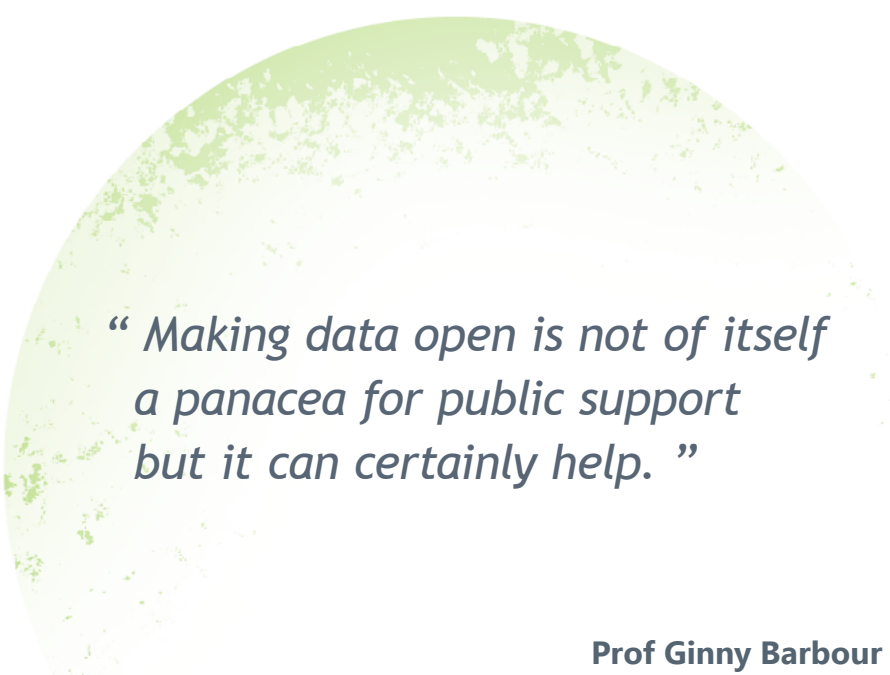
Digital Science Report

The State of Open Data 2021

The longest-running longitudinal survey and analysis on open data

Foreword by Natasha Simons, Australian Research Data Commons (ARDC)

November 2021



*“ Making data open is not of itself
a panacea for public support
but it can certainly help. ”*

Prof Ginny Barbour

Queensland University of Technology

Contents

Foreword	4
Natasha Simons — Australian Research Data Commons (ARDC)	
Three key findings from this year's State of Open Data survey	9
Dr Greg Goodey — Springer Nature Megan Hardeman — Figshare	
The role of data curation in enhancing data and metadata quality.	12
A day in the life of a data curator: the steps, challenges, and rewards of the data review process	12
Dr Connie Clare — 4TU.ResearchData	
Open source and open data: collaboration is key.	16
Sara Gonzales — Galter Health Sciences Library & Learning Center, Northwestern University	
Consolidating research data management infrastructure: a vital piece of the FAIR jigsaw & (meta)data quality improvements.	18
Damon Strange — University of Oxford	
How publishers can uphold research quality through embedded data support.	21
Graham Smith — Springer Nature	
Open data and the life sciences: the turning point.	24
Daniel Kipnis — Rowan University	
J-STAGE Data: evidence data platform for Japan's learned society publishing	26
Keisuke Iida — Japan Science and Technology Agency Nobuko Miyairi — Scholarly Communications Consultant	
Tips for engaging your researchers in open data sharing practices: practical guidance from the University of Pretoria.	29
Veliswa Tshetsha, Rosina Ramokgola, and Pfano Makhera — University of Pretoria	
How open data can help validate research and combat scientific misinformation	33
Prof Ginny Barbour — Queensland University of Technology	
Contributor biographies	36

Foreword

Natasha Simons

Associate Director, Data & Services

Australian Research Data Commons (ARDC)

Open data saves lives. The global pandemic has highlighted beyond anything that came before it the importance of data sharing in solving the big challenges of our time. COVID-19 data may be the most visualized data in history and it was made publicly available on a daily basis to people all over the world. The urgent need to better understand and treat the virus in 2020 brought unprecedented collective and collaborative action from all research stakeholders on an international scale to bring down barriers to research and speed up analysis and testing. These efforts, combined with support from governments and industry, resulted in not one but many vaccines made available by the end of the year. This gives us a glimpse of what incredible research outcomes are possible when we start with collaboration to address a common threat. Imagine how much more we could do, how many more lives we could save, if research data was routinely made open and shared. So, why isn't data sharing the norm? The answers lie in the harmony needed between policies, infrastructure, and practices.

Despite the increasing number and strength of data sharing policies from publishers, funders, and institutions — along with significant improvements in the technical infrastructure required to support data sharing — why is “data available on request” still the most common data availability statement in journals today? Why do researchers hesitate to share data and make it FAIR (findable, accessible, interoperable and reusable)? The reasons are complex and in this sixth year of the State of Open Data report, we have the data to reflect on these reasons. The data underpinning this report is

based on the largest longitudinal survey of researcher motivations, challenges, perceptions and behaviors toward open data with over 21,000 responses from researchers in 192 different countries over the six year period. The State of Open Data report from Figshare, Digital Science, Springer Nature and other leading industry and academic representatives is a critical piece of information that enables us to identify the barriers to open data from a researcher perspective, laying the foundation for future action in addressing these barriers.

“ The urgent need to better understand and treat the virus in 2020 brought unprecedented collective and collaborative action. ”

Enormous strides have been made in policy over the past decade as highlighted in the 2021 [UNESCO Recommendation on Open Science](#). This landmark document defines shared values and principles for open science and identifies concrete measures for enabling open access and open data for adoption by the 193 member states. The recommendation includes making an effort to contribute at least 1% of their national GDP to Research and Development, to set up regional and international funding mechanisms for open science, and to ensure that all publicly-funded research is in line with the core values and principles of open science.

What is most striking about this year’s State of Open Data report is that while researchers’ familiarity and compliance with the FAIR data principles is greater than ever before, there is also more concern about sharing datasets than ever before. In their article on the three key findings of this year’s State of Open Data report, Dr. Greg Goodey and Megan Hardeman stress that concern has risen in several key areas, one of which is not receiving enough credit or acknowledgement for data sharing. This points to the uncomfortable tension between the increasing ubiquity of data management and data availability policies and the rareness of rewards and recognition for data sharing. Clearly, the reward and recognition structures of academia are misaligned with the increasing demands for openness and transparency of research from publishers, funders, and institutions.

Professor Ginny Barbour reflects the sentiments expressed by many of this year's survey participants in calling for a change in the rewards system. In her article examining how open data can help validate research and combat scientific misinformation, Barbour asks: how can we ensure that the research done and published is of the highest quality and invokes trust? Open data, she argues, has overlapping roles to play in increasing the credibility of research and combating scientific misinformation so that wider society can trust it. Barbour challenges us to strengthen confidence in research as we seek to address the looming global challenge of climate change.

The principles of open science and open data are globally applicable across all research disciplines and this year's report contains perspectives from contributors in Africa, Asia, North America, Europe, and Australia. Daniel Kipnis draws out the State of Open Data trends in researchers' attitudes, behaviors, and practices in the life sciences. Almost half of the life sciences researchers responding to this year's survey share their research with the public using institutional repositories while almost 40% use external repositories such as Figshare or Zenodo. This is a significant finding as repository choices vary between disciplines and this is evidence that institutional and general repositories are the preferred option for many researchers.

This year's report found that repositories, publishers, and institutional libraries in almost equal measure have a key role to play in helping make data openly available. There is a shared responsibility between those who provide assistance to researchers that is not widely acknowledged and a corresponding lack of coordination between them. Regardless of the data sharing platform selected, researchers need help in making data open yet support for the effort required is rarely factored into the funding for research projects. Researchers must carry out this activity themselves and they seek help from those who may be able to offer it. What kind of help do researchers need to make data open and how is it offered?

Dr Connie Clare introduces us to a day in the life of Jan van der Heul, a curator for 4TU.ResearchData in the Netherlands. He describes scenarios whereby researchers need assistance to improve the quality and FAIRness of their data. Aside from assessing data files, he helps researchers improve the quality and richness of their metadata to

improve the discoverability, reusability, and reproducibility of their research.

Veliswa Tshetsha, Rosina Ramokgola, and Pfano Makhera from the University of Pretoria provide tips for engaging researchers in open data practices. They suggest that while research data management is still new at the university, the institutional library will continue to grow support for data sharing particularly in key areas such as copyright and licensing which, according to this year's report, continue to be the area that researchers require most help.

While the report shows that researchers are seeking help from institutional libraries, institutional support for data sharing is not the sole responsibility of the library. Data sharing at the institutional level is a cross-cutting activity because it is a significant undertaking that involves support across the whole research lifecycle. To streamline the process, over half of Australia's universities are collaborating to develop and trial a national research data management framework through the [Australian Research Data Commons' Institutional Underpinnings program](#). While still in progress, it is a promising model for institutional support.

“ Hurdles to data sharing in the area of policy and cultural change will fall short if we do not have underpinning research infrastructure and the experts needed to run the infrastructure. ”

This year's State of Open Data report contains a surprising insight about researchers' attitudes to policy mandates. Of the survey participants based in Asia, 42% believe funders should withhold funding or penalize researchers for not sharing their data if the funder has mandated that they do so at the grant application stage. This sentiment puts the onus on funders to check compliance yet the STM Association's 2021 research on funders with data policies found that less than one quarter actually checked compliance. The large variation in the content and strength of data policies continues to

be a challenge to researchers' understanding and compliance. While solid progress has been made in the area of publisher policies, we need to standardize and harmonize data sharing policies within and between publishers and funders. The funder-publisher alignment project currently underway through the Research Data Alliance offers promising progress in this area.

Hurdles to data sharing in the area of policy and cultural change will fall short if we do not have underpinning research infrastructure and the experts needed to run the infrastructure. We need world class data repositories, virtual research environments, facilities, supercomputers and the like to support open and FAIR data in all disciplines. We need information infrastructure on a global scale that enables interoperable human and machine readability of metadata, standards, and persistent identifiers to support data sharing and these need to be well established in research communities and embedded into research workflows. Nobuko Miyairi's interview with Keisuke Iida from the Japan Science and Technology Agency shares insights into the development of J-STAGE Data, an evidence data platform for Japan's learned society publishing. Iida outlines the challenges of building a data platform needed to match rapid changes in the scholarly publishing and technology landscape.

There have been vast improvements in data infrastructure with the development of national and regional cloud services such as the European Open Science Cloud and the China Science and Technology Cloud. Alignment, cooperation, and interoperability between open science clouds is important as research is global, and initiatives such as CODATA's Global Open Science Cloud aim to make progress in this area. Indeed, international collaboration through forums such as the Research Data Alliance, CODATA, GO FAIR and FORCE11 play a key role in identifying the challenges of policy, infrastructure, and culture change in open data and open science and putting forward solutions to these.

The State of Open Data report provides insights and commentary to the progress and challenges in researchers' attitudes and behaviors in open data. I hope you are as excited as I am to read the report to reflect on how far we have come in open data and where we need to go if we are to address the big challenges of our time and save lives.

Three key findings from this year's State of Open Data survey

Dr Greg Goodey

Research Analyst
Springer Nature

Megan Hardeman

Product Marketing Manager
Figshare

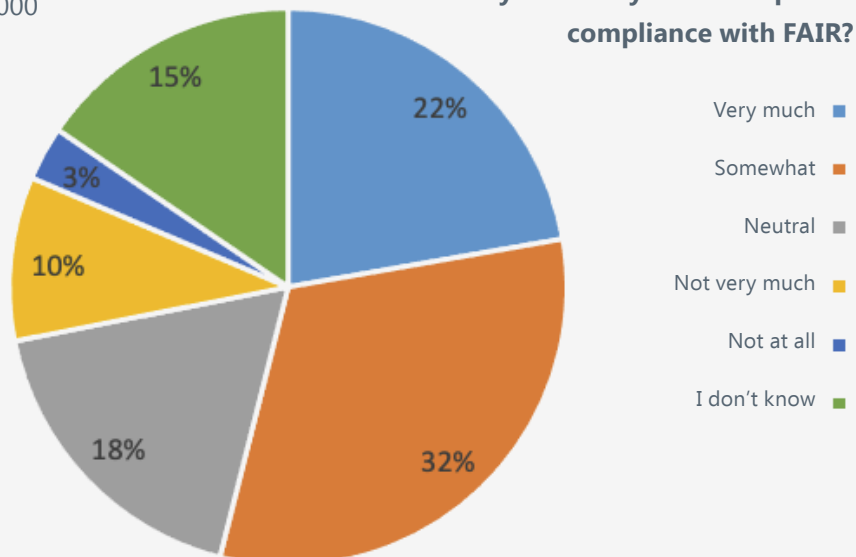
Over the course of the six years we've been running the State of Open Data survey, we've had over 21,000 responses from researchers from 192 countries, providing detailed and prolonged insight into their motivations, challenges, perceptions, and behaviors toward open data.

This year, the survey set out to continue monitoring the levels of data sharing and usage as done since the outset in 2016, and also focuses on a few key topics including what motivates researchers to share data and the perceived discoverability and credibility of data shared openly.

There is more concern about sharing datasets than ever before

In this year's survey, the proportion of respondents indicating they have concerns about misuse of data, don't receive enough credit or acknowledgement for sharing data, or are unsure about copyright and licensing has gone up compared to previous years. Given that 65% of respondents have never received credit or acknowledgement for sharing data, it comes as no surprise that this is an area of concern.

To what extent do you think you make your data open in compliance with FAIR?



Respondents indicated that their primary motivations for sharing their data are: citation of their research papers (19%), co-authorship on papers (14%), increased impact and visibility of their research (11%), and public benefit (11%). These motivations are tied to more traditional institutional measurements of impact and credit. There are calls for credit systems to be put in place for data sharing like the [Credit for Data Sharing](#) initiative developed by the Association of American Medical Colleges, the Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and

Harvard, and the New England Journal of Medicine. Initiatives such as this, however, have yet to be widely implemented.

Concerns over misuse of data and licensing are closely tied to ensuring data are as FAIR as possible; the more thoroughly documented the data are, the less likely they are to be misinterpreted or misused.

There is more familiarity and compliance with the FAIR data principles than ever before

It has now been five years since the FAIR (findable, accessible, interoperable, and reusable) data principles were established. Yet despite concerns over misuse of data and licensing, 66% of respondents had heard of the FAIR (findable, accessible, interoperable, and reusable) data principles. Of that, 28% were familiar with them, the highest number since this question was first asked in 2018. In addition, 54% of respondents thought their data was very much or somewhat compliant with the FAIR data principles; this was also the highest number since this question was first asked in 2018. These numbers are hugely positive and indicate that there could be a lessening of concern over sharing data in the long run if data are as accessible and reusable as possible.

There's also a correlation between respondents who are familiar with the FAIR data principles and

respondents who reuse their own or others' data. Of those who were familiar with the FAIR data principles, 58% had reused their own data and 44% had reused openly accessible data shared by other research groups. This suggests that data that meets the FAIR data principles are likely to be reused.

“ About a third of respondents indicated that they have reused their own or someone else's openly accessible data more during the pandemic than before. ”

Repositories, publishers, and institutional libraries have a key role to play in helping make data openly available

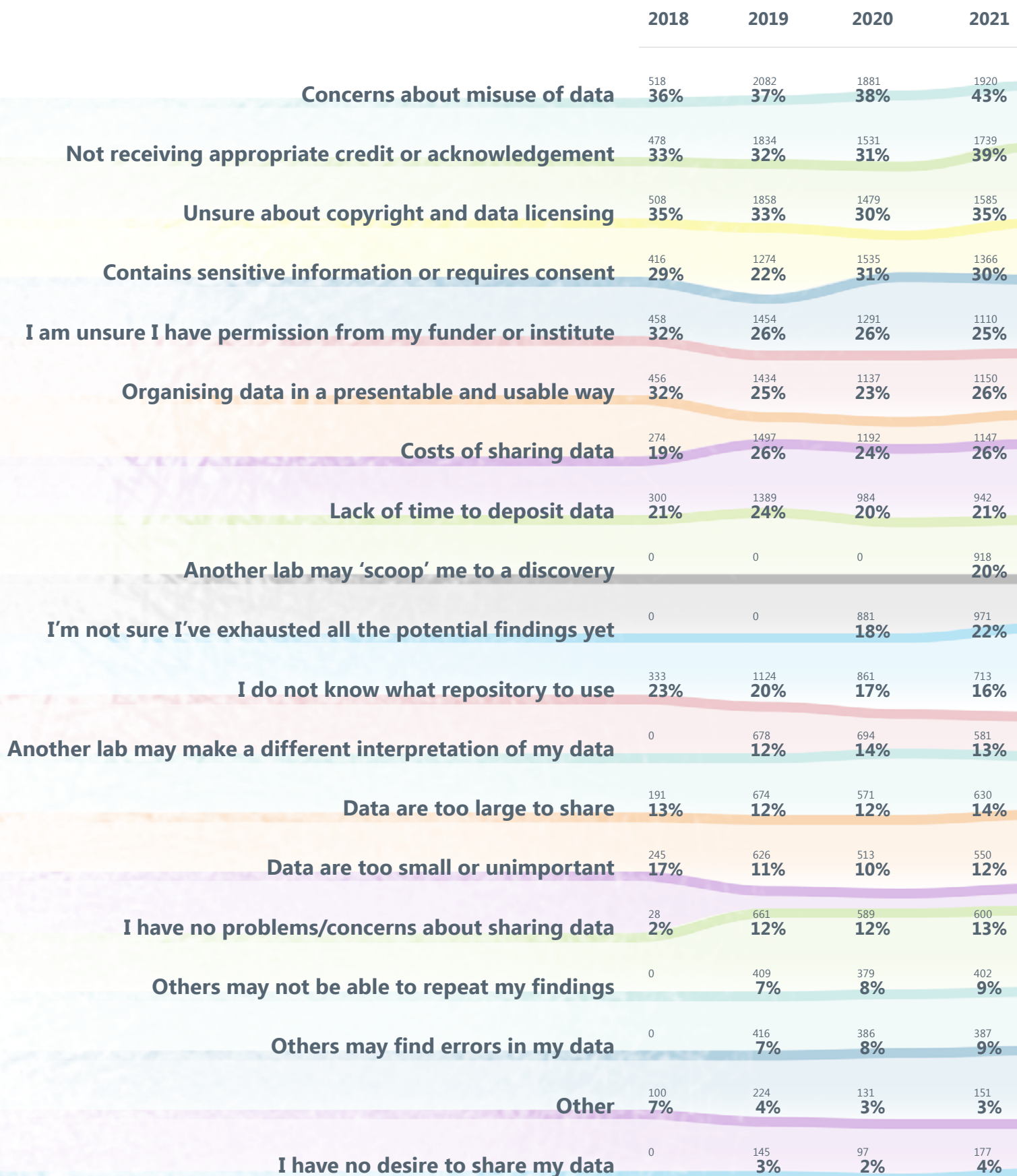
If respondents required help in making research data openly available, 35% relied upon repositories, 34% upon publishers, and 30% upon institutional libraries. Therefore, it's imperative that these organizations are able to provide the required support and resources for making data open and FAIR. Areas such as copyright and licensing (55%), finding appropriate repositories (46%), and data management policies (43%) were where respondents needed the most help. Copyright and licenses continue to be the area requiring the most help (55%) and have been so since the question was first asked in 2018. Institutions can also provide more guidance on how to comply with their policies on open data with 58% of respondents indicating they would like more direction from institutions.

[Check out the full survey results including the raw data and questionnaire](#)



Problems/concerns with sharing data

over the last 4 years



The role of data curation in enhancing data and metadata quality

A day in the life of a data curator: the steps, challenges, and rewards of the data review process

Dr Connie Clare

Community Manager

Contributors:

Marta Teperek

Head, Research Data Services

Jan van der Heul

Data Curator

4TU.ResearchData

[4TU.ResearchData](#) is an international data and software repository composed of 8,000+ science, engineering and design datasets that is run by a consortium of technical universities in the Netherlands.

Whilst the technology underpinning 4TU.ResearchData is provided by [Figshare](#), a team of dedicated staff members are responsible for managing and maintaining various aspects of the data repository, highlighting the importance of human infrastructure to support researchers with data publication.

upload

review

feedback

revisions

acceptance

publication



Meet our data curator

[Jan van der Heul](#) is one of 4TU.ResearchData's data curators. His role advances the organization's mission and vision of making research datasets published in the repository as [findable, accessible, interoperable and reusable \(FAIR\)](#) as possible.

"The data review process provides an essential service to our community by supporting researchers with the curation, sharing, access, and long-term preservation of their data," says Jan.

"Every data and software submission is thoroughly reviewed to check the validity of the [meta]data and to ensure quality requirements of the repository are met."

He explains that proper data curation enables datasets to be more easily found, understood and reused to benefit wider society.

"We're not just a 'Dropbox' for data," says Jan. "But rather our repository provides an intuitive infrastructure that allows researchers to discover, download and reuse data to avoid duplication of time and effort spent unnecessarily creating new datasets."

Quality control checks on data

Jan conducts quality control checks on data and software code submissions according to 4TU.ResearchData's [review guidelines](#). He provides researchers with detailed feedback via email before their submission is accepted and completed.

File formats

Checks are first carried out on the data to make sure that files are completely and correctly uploaded and that they adhere to 4TU.ResearchData's guidance on [preferred file formats](#).

Jan describes scenarios whereby researchers need assistance to improve the quality and FAIRness of their data.

“Sometimes, researchers don’t upload their data files but provide links to data stored on their personal computer which we can’t access and could easily be lost. In this case, we request that researchers upload the relevant data files.”

He adds that the choice of file format is also critical to ensure that the data can be reused in the future.

“In the event that researchers upload data in unconventional or proprietary file formats, I ask them to convert them to standard, interoperable, open formats to guarantee their long-term sustainability and reuse.”

Jan also mentions that a large amount of data published in the repository is [NetCDF \(Network Common Data Form\)](#) data, a file format for storing large multidimensional array data and embedded metadata.

He recommends that researchers transfer their NetCDF data to 4TU.ResearchData’s [OPeNDAP](#) server.

“The OPeNDAP protocol allows access and analysis of NetCDF data from a remote server without the need to download the data files. This helps to promote data reuse as researchers can inspect the embedded metadata as well as specific ranges, slices, and subsamples of the data,” explains Jan.

File contents

The data file contents and structure are checked to make sure information is clear, understandable and aligns with 4TU.ResearchData’s [data collection policy](#).

“I advise that datasets are deposited in English as the universal language and that they don’t contain ambiguous keyboard characters. I also ensure tabular

datasets are formatted with legible headers and labels,” says Jan.

Another essential aspect of Jan’s work is to prevent researchers from publishing data containing personally identifiable, sensitive, or inappropriate information.

“In the past, I’ve reviewed medical datasets that contain highly sensitive patient data, including patient photographs, names, and diagnoses. In cases such as this, I advise that researchers anonymize or pseudonymize their data and have informed consent to share their data before openly publishing in our repository.”

Metadata review

Aside from assessing data files, Jan makes suggestions to help researchers improve the quality and richness of their metadata to improve the discoverability, reusability, and reproducibility of their research.

“I look for peer-reviewed journal publications that accompany the dataset, check if the researcher has previously published datasets, and explore online resources, such as [Scopus](#) and [Web of Science](#) to collect relevant metadata. From this, I can suggest a more descriptive title, subject categories and keywords to describe the dataset. Sometimes it’s possible to add information about the organization that contributed to the creation of the dataset, the funding organization, and authorship,” he says.

As part of the metadata curation process, Jan also advises that authors and co-authors assign their respective [ORCID](#) ID: a unique, persistent identifier that distinguishes researchers with the same name and ensures the correct attribution of the dataset.

To improve reproducibility, the metadata record should contain a description detailing the context and contents of the dataset.

“A good description provides information about the

purpose and type of study, data collection methods, and any legal and ethical requirements. I recommend that researchers upload a [README file](#) for each dataset. This is a text or PDF file that provides data-specific information such as parameters, variables, column headings, units, codes, and symbols used,” explains Jan.

4TU.ResearchData offers researchers the option of linking additional resources to their dataset, such as peer-reviewed journal publications, supporting datasets, and [GitHub](#) accounts for software development. Jan dedicates time to validating these additional resources by checking the links have been inserted as full valid URLs that resolve to the desired location.

Finally, he checks that a license has been selected to specify the reuse requirements of data and software and suggests suitable open licenses when necessary. In addition, if a dataset is published under embargo, he confirms this choice with researchers and advises that they provide a rationale for their choice.

Challenges and rewards

Jan reveals that the main challenge of the review process is the time required to review datasets when metadata fields are only partially completed.

“Usually, I review datasets within 24 hours of submission but incomplete submissions take more time. Then, once we’ve made suggestions we have to wait for researchers to make amendments to their submission before we can publish.”

Despite this difficulty, Jan explains that the process is highly rewarding.

“My personal contact with researchers guides them through the process and helps them learn how to publish better quality FAIR data. It’s gratifying to receive their positive feedback once I’ve helped them succeed in publishing their data.”

Read more about Jan and his colleagues’ efforts on [4TU.ResearchData’s testimonials page](#).



Meme created by VU Amsterdam PhD researcher, Nadia Bloemendaal, following receipt of support from data curator, Jan van der Heul.

Open source and open data: collaboration is key

Sara Gonzales

Data Librarian

Galter Health Sciences Library & Learning Center, Northwestern University

As the world has risen to the challenge of the COVID-19 pandemic, researchers and the public alike have developed a greater appreciation for accurate and reliable open data sources. From the [National Institutes of Health's Open-Access Data and Computational Resources to Address COVID-19](#) to the local data sources that inform our nightly news updates, open data have become a more important force in our lives than ever before. People have a stake in data and, increasingly, people are contributing their time and getting involved in developing the tools that help researchers, and the world at large, interact with that data. One way we are achieving this locally at Northwestern University is through participation in open source data repository development.

InvenioRDM: an open source platform

The open source coding community is responsible for dozens of software solutions crucial to our daily lives including web browsers, content platforms, and operating systems. The open source Python programming language, with its structured, general purpose, object-oriented base, serves as the basis for the development of the new open source, turn-key repository InvenioRDM, currently being developed by an international team of highly-engaged collaborators coordinated by CERN, the European Organization for Nuclear Research. While a version of the Invenio framework has existed for over 20 years, its modernization started began in 2018, with the goal of making the institutional repository modular, scalable,

customizable, and ultimately more accessible.

From the beginning of this process, the InvenioRDM product managers have worked closely with an international team of partners including:

- Northwestern, Caltech, and NYU in the US
- Various European universities and organizations
- Eko Konnect — a cluster of the Nigerian Research and Education Network (NgREN)
- The Turkish Academic Network and Information Center
- The National Institute of Informatics of Japan.

In addition to the partners, dozens of users from around the globe have independently installed versions of the repository software and launched them at their own institutions. Both in terms of daily development and distributed user support, the open source InvenioRDM team has worked boots-on-the-ground and collaboratively to support their peers in standing up the software and supporting open resources and data at their institutions.

Metadata, DOIs, and controlled access

Though the coding team is distributed, we have prioritized agreement on a base metadata model that is compliant with data sharing mandates from the European Union and increasing mandates from US funding agencies, while simultaneously maximizing findability of data for users of the repository. Inspired by the open and participatory nature of the project, we

instituted community-based project meetings tailored for non-technical but highly involved users of the repository at the partner institutions. These users have provided significant subject matter expertise as the key users of metadata while either cataloging their own deposits or searching for deposited data from other researchers.

Through these conversations, and bolstered by the project's use of DataCite to mint unique digital object identifiers (DOIs), the partners agreed upon the use of the DataCite schema for InvenioRDM's data model. DataCite also supports data discoverability through hosting the [DataCite Commons](#), a free online tool through which users can discover the minimum required metadata that is provided with each resource that registers for a DataCite DOI. Taking these curation and accessibility conversations a step further, the InvenioRDM community's Metadata Interest Group committed to the use of the [COAR Access Rights Controlled Vocabulary](#) which has allowed us to tag data records with clear designations of either Open Access, Embargoed, Metadata Only, or Restricted.

Hosting and disseminating institutional repository records designated as Metadata Only was a key motivating factor in Northwestern's commitment to the InvenioRDM open source repository as this feature helps to serve the needs of local researchers who wish to make their datasets discoverable, regardless of the file deposit location. Librarians at Galter Health Sciences Library & Learning Center work to preserve and disseminate the scholarly output and data of biomedical researchers while respecting the privacy restrictions that must be upheld for datasets containing personally identifiable information (PII), a common occurrence in medical datasets. The Metadata Only record serves this need as it enables robust description and active curation of medical datasets through a vetted standard that maps well to Dublin Core and Schema.org, among others, while not requiring deposits of the datasets themselves. These Metadata

Only records are compliant with funders' data sharing requirements, such as the recently updated requirements of the [National Institutes of Health](#), going into full effect in 2023, while enabling data sharing upon request through Data Use Agreements, thus protecting patient privacy.

Collaborations and repository best practices

The collaborative nature of the repository work has inspired and motivated the team and continues to do so as we explore additional metadata and other enhancements. Open source tools have a critical role to play in the data sharing ecosystem, encouraging collaborations between developers, librarians, and subject matter experts. Through sharing ideas and working together to design system improvements, experts from each of these professions learn from their peers and find new skills and perspectives to bring to their own work. As librarians and researchers work to test repository improvements made by developers, each group learns from the others about workflows, usability, controlled vocabularies, and data and metadata standards. Each group comes away with a greater appreciation of their role in the lifecycle of data preservation and with a clearer idea of what they can do to make data accessible and discoverable.

Repositories of all types have served as a guide and inspiration in this process, demonstrating how data can be effectively curated and preserved for any field research. Repository best practices these tools have incorporated such as vetted schemas and controlled vocabularies, embedded file viewers, comprehensive deposit agreements, and adoption of Creative Commons licenses, have set standards toward which all new development efforts strive. The open source InvenioRDM project continues to work towards these goals while acknowledging and supporting our project partners across the globe, supporting open data cataloging and discoverability every step of the way.

Consolidating research data management infrastructure: a vital piece of the FAIR jigsaw & (meta)data quality improvements

Damon Strange

Digital Humanities Sustainability Project Manager
University of Oxford

“Putting all of your eggs in one basket” is an idiom with negative connotations, for example, when you’re referring to personal finance or data storage practice. But in our case, for the University of Oxford’s [Sustainable Digital Scholarship \(SDS\) service](#), this is exactly what we are trying to do for digital research, offering digital research projects guidance, support, and a long-term home for their digital outputs. The opportunities to converge and consolidate research data management infrastructures onto managed, shared services (e.g., Figshare) are vast, but are also not without their challenges.

We have some exceptional, world leading “eggs” at Oxford and it is only right that we have “baskets” fitting to store and showcase them. However, many researchers are often wedded to their current (or until now have had little choice but to use), often aging, “baskets” which they have had

for many years and it’s only when the “basket” finally gives out and “eggs” fall and are broken are they forced to consider an alternative. Let’s dispense with this metaphor and discuss the Sustainable Digital Scholarship service’s approach to rationalizing research data storage.

The Sustainable Digital Scholarship service: how do we ensure the content is accurate and as FAIR as possible?

The Sustainable Digital Scholarship service was launched at Oxford in February 2021 to offer support and guidance to researchers and provide access to a managed repository for storing research outputs and to showcase digital research projects. Projects are predominantly connected with the field of Digital Humanities; however, our support is by no means limited to one discipline. The primary aim of service, as the name suggests, is to ensure research data is sustainable. What we mean by that very much aligns with the FAIR principles.

Findable – A very simplistic view of meeting this principle could be the simple act of hosting research data on a platform like Figshare to make it more findable (and more accessible, interoperable and reusable) than some current hosting arrangements due to native features of the platform. However, the SDS team do offer support and guidance to researchers when it comes to metadata mapping and field creation for their projects to ensure items are well-described and custom metadata is used (where relevant) to make research more discoverable.

Accessible – It is quite often the case with some research outputs that not all the data can be made fully open for reasons ranging from personal data to copyright concerns. It has been very useful to have the feature to gate certain data items behind Single Sign-On for our repository and offer varying levels of restricted access or embargo.

Interoperable – Given the fact that many of the research projects the SDS service supports are from a Humanities-leaning discipline, the range in topics and required metadata categories have been extensive. However, we continue to work toward encouraging and promoting the use of commonly used controlled vocabularies and standardizing where a standardized approach is applicable.

Reusable – Given we are currently only 9 months into our journey as a new service at the University only time will tell. However, our hope is that as we work with and onboard more projects, we can look to reuse metadata standards and techniques to yield not only efficiencies but improved clarity & quality of (meta)data.

A green letter 'F' inside a light green rounded square.A teal letter 'A' inside a light teal rounded square.A red letter 'I' inside a light red rounded square.A grey letter 'R' inside a light grey rounded square.

Research data “resurrection” of legacy collections: can we make version 2.0 better?

We predict that over the coming years, the SDS service will continue to work with researchers whose data collections or research projects have fallen offline or have experienced a level of diminishing functionality as part of its historical technical arrangements. Although this is potentially a worrying time for the researcher, out of the uncertainty of hosting on failing (or failed!) infrastructure, there are potential opportunities to reinvigorate and refresh the research project as part of its next iteration.

One current and relevant example we have been working with is a project called the Novum Inventorium Sepulchrale - Kentish and Anglo Saxon Grave Goods in the Sonia Hawkes Archive. It's a fascinating database that published records of c. 1,000 graves and the objects found within them, including images and diary entries. However, since the project went offline indefinitely in 2018, all that remained was access to 2 metadata spreadsheets on a single [project webpage](#). With the support of the project's Principal Investigator, [Professor Helena Hamerow](#) at the School of Archaeology, the SDS team has brought this [project back to life](#).

Clearly, there is a very binary way of looking at the improvements here for Novum Inventorium Sepulchrale in the sense it wasn't a project database online and now it is back online once more. But also, we have had the opportunity to take a very hands-on and curatorial approach to cleaning the project's metadata before we rebuilt it on our repository. Naturally, the addition of mandated Figshare fields to allow DOI creation for each record is an excellent improvement and a necessary process for ingest. We were also able to rationalize some of the metadata fields whether this be omission, merging, or adding new fields; the hope is that the quality of metadata

attached to the collection will be improved by undergoing this process.

Our hope is that the number of research projects falling into the category of “resurrection” will diminish over time by virtue of the good work we are doing as part of the SDS service and encouraging the practice of building in “sustainability by design” for new research grant applications. This is something we are aiming to achieve by working closely with Research Facilitation & IT Support teams at the University.

Final Thoughts

If, at the University of Oxford, we can continue to amass digital research project “eggs” within our “SDS service basket,” this will hopefully improve data sustainability and make research as FAIR as possible. There will always be the odd “egg” that needs to be stored in a less than ideal “basket” that needs regular maintenance and updates or a custom-built feature-rich “basket” with all the technical ‘bells and whistles’ deemed relevant for a particular research use case. With the pursuit of research innovation this is perhaps inevitable, but where we can standardize and consolidate, we must do so as the benefits of doing so are significant.

How publishers can uphold research quality through embedded data support

Graham Smith

Research Data Manager

Springer Nature

Scholarly publishers have a fundamental duty in upholding research quality, from editorial expertise to managing the peer review process. Research data is a growing part of Springer Nature's policies, systems and workflows and a key component of the ambition that research outputs should be openly available and reproducible. In order to uphold the quality of data alongside that of the related literature, we are building on the specialist support developed for data articles, developing processes more widely applicable across our journals.

Previous [State of Open Data reports](#) have highlighted the key role that publishers play in helping researchers share their data. The COVID-19 pandemic put a particular spotlight on data quality and, moreover, research quality. The scientific community's initial response focused on making research outputs rapidly and openly available; [funders, journals and researchers combined their efforts](#) to ensure this happened. [Preprints saw considerable growth](#) from these initiatives and peer review times dropped. Up-front release of data was specifically included in these measures and some publishers, [including Springer Nature](#), provided additional support for data curation and sharing.

However, there have been [doubts raised](#) over the quality of such "rapidly published" research. The [Surgisphere scandal](#) was a notable example of extremely rapid data release with major question marks

over the quality, provenance, and veracity of said data, despite the fact that it swiftly formed a basis of public health decision making.

The role of research data and specialist support

So, what do we learn from such scandals? Research data has a clear part to play, ensuring there is evidence behind the claims in peer-reviewed literature. We at Springer Nature have [championed FAIR data since its inception](#) while [supporting transparency and reinforcing community expectation](#) through the rollout of standardized data policies. By focusing on the findable and accessible aspects, simply making data available is a first step in improving the quality of published research, allowing greater scrutiny of reported findings. Along this theme of transparency, much of the backbone of FAIR data is good metadata,

with detail provided (or not) enabling an assessment of how much a dataset can be trusted. As the Surgisphere example demonstrates, however, data quality doesn't end with making data available and potentially reusable.

Specialist data support (also known as data curation or stewardship) is a growing field enabling FAIR compliance and checks on the robustness and reliability of data. This is ideally provided as early as possible in a research project, for example when producing a data management plan. Some research [institutions](#) and [repositories](#) provide this service, but as the [2020 State of Open Data report](#) outlines, researchers usually look to publishers for help sharing data related to their papers. While a researcher is the expert in their own data, a general data specialist supplements this expertise with support in areas

the researcher may not know about like selecting the right repository, adding useful metadata, long-term preservation, data rights, and linking. Working alongside editors, who often bridge the gap in disciplinary and data-specific expertise, these specialist roles provide researchers with assurances about their data, minimise risk, and promote data quality.

Springer Nature supports data sharing both through improving data availability across our research journals and publishing data-specific journals and articles. Two prime examples of this data publishing are [Scientific Data](#), Springer Nature's flagship data journal, and the briefer data notes article type at [BMC Research Notes](#) and [BMC Genomic Data](#). All have embedded support from research data specialists to safeguard data quality working alongside peer review of data and manuscript itself.

Like the FAIR data principles, the checking process considers three areas:

- ❑ the data themselves
- ❑ metadata describing these data
- ❑ infrastructure e.g. hosting, linking, and preservation

Expanding this support to a wide range of journals and disciplines, standardized checklists can form the basis of data quality assessment. The resulting action, however, might be something a specialist can apply an immediate fix to or that requires a closer look with the author and or/editors. Such issues include:

- Are the data shared in the right repository? Is there a more suitable discipline-specific venue available? Have the right standards been used?
- Are the data provided complete, consistent, and accurate alongside the reported manuscript or metadata?
- Do the data contain sensitive elements that should be removed or anonymized?
- Are the data licensed appropriately to maximise reuse?
- Do the metadata provide sufficient context for another researcher? Are the files organized in a way that supports access and reuse?

These checks may supplement or even overlap with peer review which will incorporate considerations such as methodology to produce the data. In this

context, another main consideration and challenge for publishers is effectively getting data in front of reviewers.

What's next?

Of particular relevance to data quality — and, therefore, research quality — is the strength of link between published literature and underlying data. Data journals like Scientific Data have this at their core, as reflected by embedded support and the expectation of data peer review. The reality is that for much of research publishing, reviewers might not look at underlying data at all. With the growth of wider data policies, data sharing mandates, and community expectations, there is a growing awareness of data in the publishing process that can and should be supplemented by standardized checks, workflows, and supporting tools.

The [development and implementation of standardized data policies across Springer Nature's journals](#) has provided a strong foundation to promote and improve data quality. The checks and balances outlined above lead the way in particular journals and article types; our next steps are to apply a suitable level of data expertise more widely throughout our publications and processes. It is encouraging to see in this year's State of Open Data that researchers highly rate the quality factors of clear descriptors, classifications and coding of data, as well as links to peer reviewed literature. These are all areas that embedded editorial support can improve.

We also acknowledge that researchers come to us relatively late in the research lifecycle and that this is a community effort with other actors such as funders and institutions playing a growing role in data management, support, and sharing. Putting the researcher's needs front and center is paramount, whether as data producers or users, and we all have a role to play.

Open data and the life sciences: the turning point



Daniel Kipnis

Life Sciences Librarian, Rowan University

It may have finally happened. The catastrophic COVID-19 pandemic had you hearing conversations such as: "But what does the data say?", "Did you read the [Israeli study](#) analyzing real world safety data from the Pfizer vaccine?", or "Was the sample size large enough?" Scientific research may never be the same and COVID-19 could be the historic inflection point [where using open data](#) transformed how researchers collaborate. Life sciences research is global and COVID-19 has proven this [with researchers from all around the world](#) coming together to seek solutions.

The Open Data movement has slowly grown with 2,700+ repositories available in [Re3data](#). Approximately [1500 are in the life sciences](#) and [68 are COVID-related](#) entries. So, where do we stand on sharing data since the COVID-19 pandemic arrived in 2020? This is the sixth State of Open Data report that [Digital Science](#) has published and this year's survey results and analysis for 2021 reveal shifts in how life science researchers are viewing open data.

Key findings and guidance on next steps

Respondents from within the life sciences (20%; n=820), from North America (26%; n=743) and from larger institutions (27%; n=246) were significantly more likely to indicate that their funder was encouraging them to develop a data management plan than average (16%; n=4,491).

Seek out a librarian on your campus to help with working on a data management plan and working on your data's metadata. Many universities have data librarians on staff to help researchers manage their data. Creating a README file is helpful for future researchers and only 20% of 271 life science researchers were able to access a README file from survey results. Placing energies at the start of data collection will help down the road when it comes time to share, discover, and reuse the data. A good place to start is [DMP Tool](#) where you can access ready-to-use data management templates.

Almost half (46%) of 779 life sciences researchers responded that they share their research with the public using institutional repositories, followed by external repositories (e.g. Figshare, Zenodo) at 39%, cloud file sharing (e.g. Dropbox, Google Drive) at 20%, funder repositories at 19%, blogs/websites at 14% and other at 13%.

Consider archiving data in a discipline-specific or a general repository. This helps with consolidating subject-specific data and will address the different and siloed approaches in how various publishers are handling data. Archiving data is another method to [increase citation rates](#). If increased visibility to research and impact factors continue to be models for promotion in the academy, then archiving data and making it readily available should help with elevating researchers.

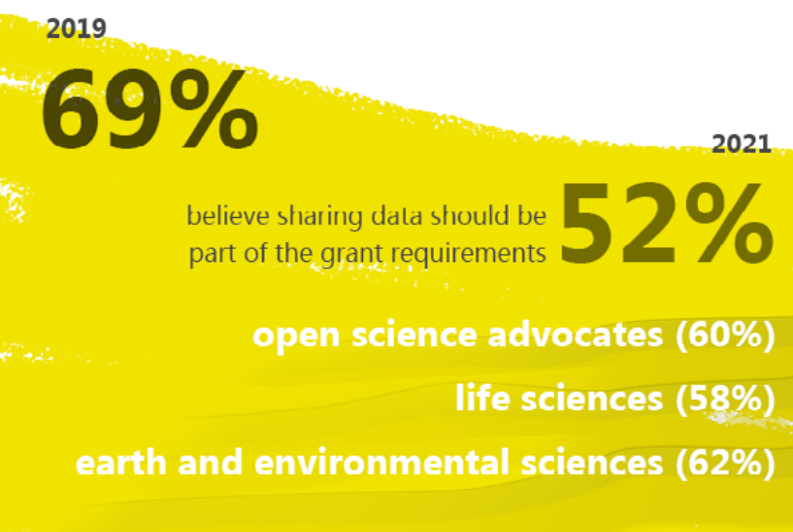
shared data, 41% had concerns about not receiving appropriate credit or acknowledgement, and 34% were unsure about copyright and data licensing.

Data can be as relevant as an article citation. One could even argue that an article citation only happens with data. Researchers should advocate for tenure committees to see the value of open data and rethink what “counts” in the academy. Many prizes are given for scholarly papers, why not prizes for data or other vital research content? For example, the importance of open data is being elevated thanks to awards that demonstrate the importance of open data including the [University of Bristol Open Research Prize](#) and [The University of Groningen Library Open Research Award](#).

In addition, here is another opportunity for librarians to help with understanding and teaching copyright and licensing issues. Education efforts teaching about FAIR data principles continue to be an opportunity for librarians and data curators. 29% of 820 life science researchers had never heard of [FAIR data principles](#) before taking the survey. 30% of respondents indicated familiarity and 41% had previously heard of the FAIR data principles, but were not familiar with them.

Many complex issues involving open data continue to exist including interoperability between dataset, discoverability of datasets, misuse of pre-published research and long term storage, and data management strategies. The findings in the survey show how researchers are working with open data and the work that needs to continue to help with research innovations that save money and address the global problems such as climate change and food security.

Issac Newton is credited with the expression “standing on the shoulders of giants” to exemplify that truths can be discovered by building on previous discoveries. In order for this to happen, a transparent process of sharing data is imperative to help with reproducing studies and creating new shoulders to stand on.



When asked whether researchers felt that sharing data should be a part of the requirements for awarding grants 52% agreed. **This proportion was significantly higher for open science advocates (60%) and those in the life (58%) and earth and environmental sciences (62%).** The level of support does appear to be waning since the question’s first introduction in 2019 when support was at 69%.

According to the 820 life science researchers who responded, 42% had concerns about misuse of

J-STAGE Data: evidence data platform for Japan's learned society publishing

Keisuke Iida

Department for Information Infrastructure
Japan Science and Technology Agency

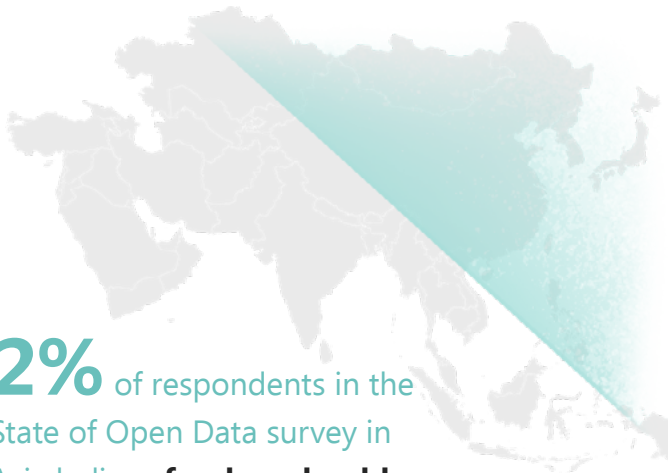
Interviewed and translated by: **Nobuko Miyairi**

Scholarly Communications Consultant

Japan Science and Technology Agency (hereafter JST) promotes research and development in Japan through funding basic research, commercialization of new technology, and promoting international collaboration. JST also provides a variety of information platforms and services, including J-STAGE, an electronic journal platform. In October 2019, JST commemorated the 20th anniversary of J-STAGE, which now hosts more than 3,000 journals, conference proceedings and other academic contents published in Japan. J-STAGE Data is a new data repository to make underlying data available for J-STAGE publications.

How did you come up with the idea of building J-STAGE Data?

Since its launch in 1999, we have invested significant resources in J-STAGE to keep it up to global e-journal standards and good practice, by adding new features from manuscript submission to peer review process to the dissemination of contents. Over these 20 years, however, the scholarly publishing environment has so rapidly evolved that we felt the need to revisit J-STAGE policies and operations in order to adapt to the changing standards and practice. We established our advisory committee in March 2018 to deliberate our mid- to long-term strategies. Their final report [available in Japanese only] boiled down strategic actions into three areas: updating the e-journal platform in response to new demands, strengthening



42% of respondents in the State of Open Data survey in Asia believe funders should withhold funding from or penalize researchers for not sharing their data if the funder has mandated that they do so at the grant application stage

the collaboration mechanism with Japanese learned society publishers, and optimizing the means for service quality improvements. Creating a data repository was part of the action plans reflecting recommendations in these three areas.

What were the changes that J-STAGE had to make?

The world of scholarly publishing has evolved in its technology with diversified contents over the years. The open access publishing model is widespread, and the emergence of preprints and other non-traditional research outputs no longer warrant a single platform just for peer-reviewed journals. Sharing underlying data for publications is a prevailing trend among scholarly publishers, underscored by increasing awareness of research ethics, transparency, and access needs for publicly funded research. In addition, long-term preservation of all the research products, as well as standardization of metadata and infrastructure, require us to constantly optimize our choice of technology and platforms.

How does J-STAGE Data meet those new demands?

J-STAGE already offered a service called “electronic supplement” to allow publishers to upload supplementary data for a journal article, but the number of files and their size offered was quite limited. Creating J-STAGE Data as a new, separate platform allowed us to leave technical legacies behind and incorporate new practices into our workflow. We adapted Figshare for Publishers as our base platform, which met our basic requirements such as DOIs for datasets, Creative Commons licenses, and a user interface to browse, search, and download. Since each dataset had to be associated with a corresponding J-STAGE publication, we also needed a review workflow in place as a data curation mechanism and also for the

peer review process. Having a separate data platform made it more flexible to publish a wide range of file types in large sizes — including different contributors associated with each — and link everything back to the main J-STAGE publication. Basic usage data like views, downloads, and citations are important indicators to determine the success of J-STAGE Data; so far, we are seeing increased usage of these datasets compared to the supplementary data in the older format.

What were the challenges in creating J-STAGE Data and how did you overcome those challenges?

The initial challenge was how to accommodate Japan-specific requirements. [Japan Link Center \(JaLC\)](#) DOI was our default choice and the DOI minting process had to be customized in the Figshare platform. We added a number of metadata fields to allow both English and Japanese information for title, authors, descriptions, etc. Perhaps bigger challenges came outside of the system. We developed extensive user manuals for our J-STAGE users, who are quite familiar with the journal publishing process but not necessarily with data publishing. There are certain things you can do on the platform that are not in accordance with the J-STAGE publishing policies, so we had to come up with a standard data publishing workflow that does comply. This included the timing of data release, embargo setting, support communications, among other things. We developed our policies based on user feedback and after some trial and error.

Do you have any research data policies in Japan and how do those policies relate to J-STAGE Data?

Japan’s Cabinet Office has assembled the expert panel on open science since 2014 and published several recommendations. More broadly to cover the science and innovation policy, the national [Basic](#)

[Plan](#) is renewed every five years and we are in the beginning of its 6th cycle (2021-2025). The latest plan calls for actions by government agencies and research organizations to refurbish research systems that form the backdrop of open science and data-driven research, where research data management and reuse is strongly encouraged.

JST established our open access policy in 2013, which has been replaced by the [Policy on Open Access to Research Publications and Research Data Management](#) that covers both OA publications and research data management. J-STAGE Data is an extension of an existing e-journal publishing platform and does not directly support our policy as a funding agency; however, since J-STAGE is widely used by Japanese society publishers, we believe J-STAGE Data can serve as a vital tool for those who are in need of a data publishing platform conforming to the national recommendations.

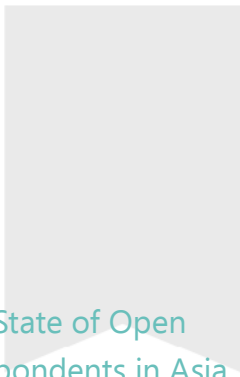
How was J-STAGE Data received by society publishers?

We spent the first year on a pilot basis with a small number of publishers on board. This soft launch was useful to gather feedback from early adopters and optimize our policies and operations. After J-STAGE Data was officially launched in March 2021, we started organizing hearing sessions for those publishers considering data publishing. We received mixed feedback partly due to the familiarity with data sharing practice in different fields and perhaps due to different levels of personal incentives, too. Some research disciplines may have a longer history of data sharing while other fields may have reservations due to the sensitivity of the datasets. As they come on board, the data curation process reveals different metadata needs by research fields and their practice. Overall, however, the response is positive and we are starting to receive more applications than we expected.

How do you plan to develop J-STAGE Data in the next few years?

Enriching the metadata is our next priority. For example, most datasets list “authors” of corresponding publications as data creators, which may not always be the case. We could more correctly capture each author’s (and others’) contributions if more granular metadata allow us to do so. Most datasets are labeled with a CC license and are openly published, but some datasets may require tighter access controls and an explicit copyright statement when necessary. Multi-language support is something we need to consider, too, as we expand our user base in Japan.

Finally, as we see more journals make use of J-STAGE Data for sharing evidence data for their publications, we hope to see a clear synergy between the two platforms. Usage increase is an obvious one, but it may also be possible for published data to inform new research more directly, in which case data citation will be a clear indicator and something we are keen to closely keep track of.



23% of the State of Open Data survey respondents in Asia indicated that **citations** of their research papers would **motivate** them to **share their research data**

awareness

trainings

recognition

guidance on licensing

correct errors

enhancing profiles

descriptive metadata

data management

Tips for engaging your researchers in open data sharing practices

— practical guidance from —
the University of Pretoria

Veliswa Tshetsha

Senior Coordinator:
Open Scholarship
University of Pretoria

Rosina Ramokgola

Data Curation Officer
University of Pretoria

Pfano Makhera

Metadata Specialist: Scholarly
Communications
University of Pretoria

Background on data management and engagement practices at the University of Pretoria

The library at the University of Pretoria started engaging in research data management activities in 2009. We conducted an initial research data management (RDM) survey from October 2009 to March 2010. A second survey involved interviewing the Deputy Deans of Research from Faculties to determine the essential research data that the University should manage. Two pilot projects aimed at gaining insights and understanding of researchers' RDM needs took place in 2013 and 2014 with the Institute for Cellular and Molecular Medicine (ICMM) and the Neuro Physiotherapy Group. These projects used open source document management systems, Alfresco and Islandora, and were customized to manage data. After the pilot projects, further developments took place including identifying a campus-wide database or repository for the publishing of open access datasets. Further investigation took place until 2018 where Figshare was introduced to the DLS and was approved and implemented as a data repository solution in July 2018.

The library developed RDM resources such as a LibGuide and implemented an RDM readiness Training Toolkit. The [toolkit](#) contains videos on how to upload datasets, how to be responsible with your research data, how research data management is quick and easy to implement with access to data remaining under your control, and why effective research data management matters today, tomorrow, and in years to come.

Researchers are supported with meeting funder requirements on data management plans and data sharing practices. The library has recently established a strong partnership with the institutional research grant office and is working toward integrating data management plans in the grant application process. Advocacy and support for researchers is required particularly because with the free availability of the university's open data repository, researchers should rest assured that their datasets will be securely curated and accessed when needed.

Future plans for RDM

RDM is still new at the university. We have just started and we will continue tracking and harvesting University-affiliated datasets and engaging our users by consulting them on further use cases and how we can provide support. We will also look at developing an integration where postgraduate students are required to submit datasets for their thesis submission before they graduate. We will also create a requirement to submit a data management plan as part of the grant application process.

Tips for how to engage your researchers in open data sharing practices

The following are a few common problems or challenges that survey participants said they faced with sharing datasets in this year's State of Open Data survey. We have provided some tips and examples of how to overcome these challenges based on our experiences at the University of Pretoria.

Misuse for commercial use or misinterpretation

Institutions should create awareness (training, advocacy) for researchers in areas pertaining to reuse of data — for example, the Creative Commons licenses. The library worked on a roll-out strategy by hosting RDM Repository roadshows, workshops, creating awareness across faculties.

The library hosted webinars on RDM for early career and well-established researchers on how to discover, manage and share data, how to upload data, how to create Data Management Plans, and RDM in general.

Researchers and postgraduate students received training and were guided on how to secure their data by generating DOIs to enable attribution and discoverability. Our repository has features to protect datasets either privately or publicly. User guidelines and data dictionaries are provided. Researchers should also provide as much information as possible in the metadata; this will make it easy for other researchers to understand and interpret data.

Unsure about copyright and data licensing

Libraries should play an active role in providing training and guidance on copyright ownership and data licenses. Our library provides copyright services such as training researchers on copyright and fair use.

The library also offers copyright compliance awareness lunch hour sessions for lecturers and students. The aim is to engage and educate users on copyright and the importance of complying with the legislation.

Not receiving appropriate credit or acknowledgement

Institutions can implement research data recognition grant awards for researchers who are not receiving appropriate credit or acknowledgement or who do not have the desire to share data. This reward tool can be expanded to include postgraduate students, as well. This can also form part of the research(er) data performance evaluation.

Researchers can be recognized in many ways; for instance, have Researcher of the Month through university websites, social platforms, or have University Researcher Month where researchers will be acknowledged and prizes won or certificates of merit issued.

In South Africa, researchers are rewarded for generating, preserving, sharing and/or re-using research data by the National Science and Technology Forum (NSTF-South32 Awards). The call for nominations recently came out and we will nominate researchers who are sharing data in our research data repository.

Organizing data in a presentable and usable way

Researchers should use data management plans (DMPs) and archive their data in trusted repositories. An RDM policy, as well as funders, encourages researchers to create DMPs; this can be done using something such as DMPTool. The library assisted in the establishment of a national Data Management Plan (DMP) tool in South Africa. Recently, the library had a stakeholder engagement and a DMPTool roll-out strategy.

Another lab may make a different interpretation of my data

If the data is not described properly, others are likely to misinterpret it. Descriptive metadata plays an integral part in ensuring that data is interpreted correctly. Data sharing fosters collaboration both locally and internationally.

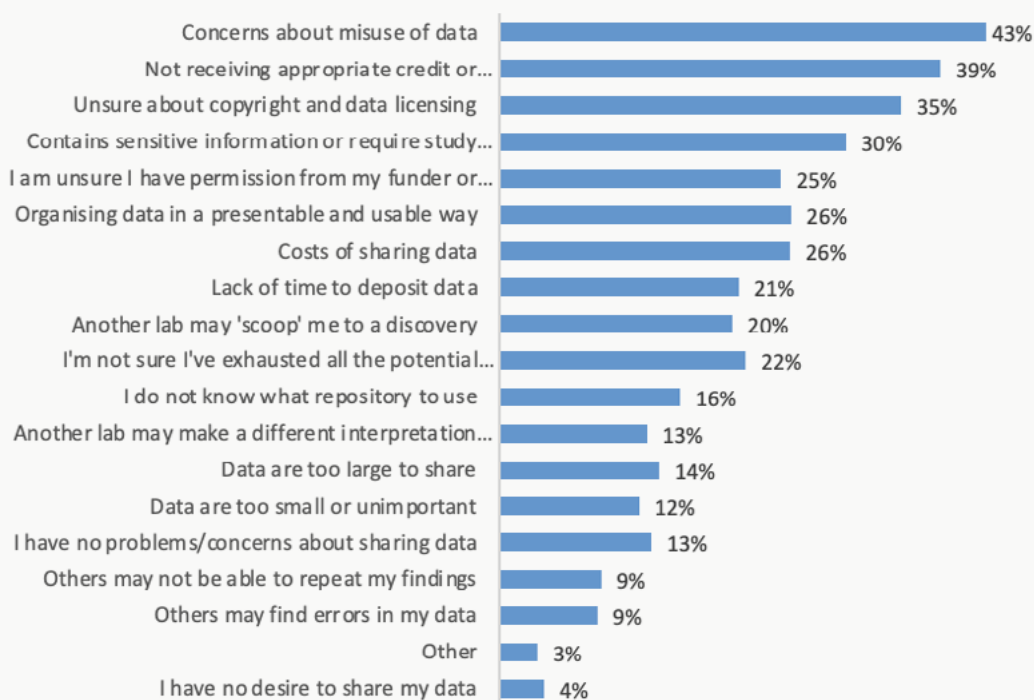
Others may find errors in my data

Data is for reuse and sharing data allows others to correct those errors and collaborate with the researcher. When data is publicly open it fosters collaboration with other researchers in the same field or in adjacent fields. Lack of complete information may result in errors and libraries should guide researchers on the use of good data management tools and data quality standards.

I have no problems/concerns about sharing data

Positive researchers can educate other researchers; this is the practice that institutions should adopt. During workshops, we use existing researcher profiles to educate others. During International Open Access Week, we showcase some of our institution's researchers whose works are open so as to encourage others. We have recently started a research(er) visibility and impact project where we support researchers to enhance their profiles on our institutional data repository.

What problems/concerns, if any, do you have with sharing datasets?



How open data can help validate research and combat scientific misinformation

Prof Ginny Barbour

Co-lead, Office for Scholarly Communication,
Queensland University of Technology
and Director,
Open Access Australasia

The 2021 State of Open Data survey provides valuable insights into data sharing globally. Though it can't capture what researchers everywhere think of data sharing, this survey of nearly 4,500 researchers offers helpful perspectives, some reasons to be hopeful, and some key takeaways that can support discussions on how open data can help validate research and combat scientific misinformation.

The decision to share data and the mechanisms necessary to support sharing don't exist in a vacuum. In many ways, the problems of how to share data are reflective of both the culture of science and of current logistical challenges playing out across research globally. How can we move to a more open world? How can we ensure that the research done and published is of the highest quality? How do we increase trust in research? How do we shape an incentive system that addresses these challenges? The survey has insights to offer on each of these key questions.

It is worth noting up front that anyone answering a survey on data sharing is likely to have an interest in the topic as well as time and sufficient access to the technology to respond. Not surprisingly, therefore, the country with the largest individual responses was the US; the sample was tilted towards researchers from large institutions and only a handful of countries had more than 100 respondents. The respondents were largely supportive of open access to all research outputs (over three-quarters) — a higher number than would be found in a random sampling of researchers. The responses should therefore be viewed as being slanted to the technically well-supported, relatively well-funded end of the researcher spectrum, who have been exposed to discussions and information on open science.

What do we learn from the survey that can inform the debate on how we can promote trust in research? First, at an individual level, many respondents are putting in the hard work of data sharing. They are making data management plans (74%), doing work to curate their data for sharing (76%) and 66% of respondents are familiar with the FAIR principles that underpin data sharing. These are hopeful indicators, though digging further into the results, there are gaps in the support for data sharing. At the most basic level, 30% of respondents are unclear on who will pay to make data open, and more than half the researchers need support in understanding copyright and licensing of data. The results show evidence of a policy vacuum around data sharing, with respondents looking to both institutions and national funders for leadership: 52% of researchers believe that funders should make data sharing a requirement and 48% feel that if such a mandate is in place, funders should hold researchers to it.

When we turn to researchers' efforts to reuse data, difficulties become apparent, and the importance of key infrastructure is highlighted. For example, respondents who sought to reuse others' data were

more likely to be able to get the full dataset from an institutional repository compared with a journal. But even when datasets were accessible, more than 50% lacked clear licensing information, and the quality of other descriptors were variable, pointing to gaps in key metadata and contributing to researchers' perceptions of the quality of the datasets. When considering challenges to making their own data open, researchers expressed a series of concerns, ranging from being scooped, having sensitive data misused and getting insufficient credit for making data open.

“ Open data has two important, overlapping roles to play in increasing the credibility of research: validating research, so that researchers can trust it, and combating scientific misinformation, so that wider society can trust it. ”

Underpinning all of research has to be the concept of reproducibility. For too long, we have had a publishing system that rewards the publishing of new and exciting findings in specific journals more than publishing of confirmatory (let alone so called “negative”) findings. Largely, these findings are still published in a prescribed format that allocates more time and resources on typesetting and branding than it does in providing access to the underlying data, code, and other materials that allow research to be verified. The survey shows that researchers are only too aware of the limitations of the system that they are required to work within but that they understand

the need for change — and want this change. 80% of researchers thought that a research article that had data openly available was more credible but when we wonder why this practice is not more widespread, the finding that only 18% of respondents believe that researchers currently get sufficient credit for sharing data, offers an explanation. Researchers know how they want to be credited: 61% want to get credit for the data they share through citations of papers with data. Interestingly, in the absence of such credit, researchers are cooperating among themselves to share credit by including the generators of datasets as authors on papers that reuse these data. This practice indicates, yet again, how critical it is for researchers' careers to get sufficient credit for their work and in the absence of other mechanisms — such as specific support for open science practices, as championed by DORA and through the Hong Kong Principles — researchers will attempt to get credit through the current system of journal publications.

What about wider public trust in research? The COVID-19 pandemic has shown us, yet again, how critical it is that research is trustworthy. Making data open is not of itself a panacea for public support but it can certainly help. High profile retractions of papers during the pandemic because of concerns over underlying data show how far we have to go. In less highly-scrutinized research, it's unlikely that the problems with underlying data would have come out so quickly, if at all. Contrast this state of affairs with

what the public expects for other products that they consume: there would be outrage at a similar lack of proper control in the production of a novel food item. As we face the complex challenge of climate change, trust in research will become even more critical, especially as climate policy is so politically charged and climate research itself is often the subject of public debate. A recent paper from the International Science Council makes the case for ensuring data behind research is available so as to strengthen the trustworthiness of research.

So, the underlying message of this State of Open Data report should be one of cautious optimism, but with some pointers for change. Researchers largely want to share their data, but the current system fails to support or adequately reward them for doing so and we are still a long way from a world where it is the norm to share fully-curated data. Until then, researchers are left to navigate a system that makes it harder than not to share and where, most alarmingly, the public may only fully understand the importance of data sharing when it's shown to have gone dramatically wrong. There's no time to lose. We need to strengthen confidence in research as we seek to address the looming global challenge of climate change.

Contributor biographies



Natasha Simons

Natasha Simons is Associate Director, Data & Services, for the Australian Research Data Commons (ARDC). She leads a team that collaborates nationally and internationally to solve key challenges that will improve infrastructure, policies, and practices to enable FAIR data. She is responsible for delivery of ARDC's National Data Assets Initiative that leverages strategic partnerships to develop a portfolio of national-scale data assets supporting leading-edge research. Natasha has a co-chair role in a range of international groups including the Research Data Alliance and the Global Open Science Cloud and she is a member of the FORCE11 Board of Directors.

 <https://orcid.org/0000-0003-0635-1998>



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>



Megan Hardeman

Megan is the Product Marketing Manager at Figshare. Prior to this, she spent five years as the Engagement Manager at Figshare, working closely with institutions and researchers to provide guidance and best practices for storing and sharing research data.

 <https://orcid.org/0000-0002-1911-7503>



Dr Connie Clare

Connie is the Community Manager at 4TU.ResearchData, an international repository for science engineering and design research data. Her focus is to engage researchers and data support professionals about data management, and to bring discipline-specific communities together to stimulate the creation of FAIR data through use of the 4TU.ResearchData repository.

 <https://orcid.org/0000-0002-4369-196X>



Sara Gonzales

As the Data Librarian at the Galter Health Sciences Library & Learning Center at Northwestern University, Ms. Gonzales develops and delivers training in data management and data cleaning to clinical researchers, support staff and students at the Feinberg School of Medicine. She leads the CTS Personas project, developing 1-page persona profiles representing employees in clinical and translational science. She chairs the Galter Library's Digital Initiatives Working Group (DIWG), charged with executing and communicating digital system improvements being implemented across the library and related university systems. She also co-chairs the DIWG's Metadata Subcommittee, a group committed to metadata standards and improvement related to library cataloging and maintenance of the institutional repository. She also serves as the Community Manager of the international InvenioRDM repository software development project and Assistant Director of the National Evaluation Center of the National Network of Libraries of Medicine.

 <https://orcid.org/0000-0002-1193-2298>



Damon Strange

Damon Strange is a freelance management consultant, currently working as a Project Manager for the Sustainable Digital Scholarship (SDS) service, at the University of Oxford. Playing a pivotal role, and managing the project which led to the implementation and launch of the SDS service, he continues to support improvements to the service, as well as working with researchers at Oxford to find the right technical solutions for their research data. Damon has trained in Urban Studies and Planning (BA, MA, University of Sheffield), as well as holding a number of professional Project Management qualifications (Agile, PRINCE2, APM, Scrum), and prior to Oxford University has supported the successful management of numerous projects within other HEI & Local Government organisations.

 <https://orcid.org/0000-0002-5851-718X>



Graham Smith

Graham Smith is Research Data Manager at Springer Nature, where he develops, promotes and embeds research data initiatives across the organisation. His role reflects the growing recognition of research data as first class research outputs, working to offer researchers the best options for the data behind their publications. Throughout his career he has worked with a wide range of researchers, editors, policy-makers and technical solutions to develop a data specialist viewpoint, implementing data curation and metadata services in public sector and commercial settings, including at the Natural History Museum in London.

 <https://orcid.org/0000-0001-9520-0109>



Daniel G. Kipnis, MSI

Dan is the Life Sciences Librarian in the Campbell Library at Rowan University in Glassboro New Jersey, United States. He has been an academic librarian for twenty years and his research interests include information and digital literacy in STEM and institutional repositories. He is active with the New Jersey ACRL chapter as Co-Chair of the research committee and the Philadelphia Area Science, Technology and Engineering Librarians (PASTEL) community.

 <https://orcid.org/0000-0002-4589-5106>

Keisuke Iida

Keisuke Iida is a JST officer and works at the Department for Information Infrastructure. He oversees J-STAGE Data's day-to-day operations to ensure adherence to policies and metadata quality.

 <https://orcid.org/0000-0001-5924-0728>



Nobuko Miyairi

Nobuko Miyairi is Scholarly Communications Consultant, based in Tokyo, Japan. Her service caters to a wide range of business needs from academic societies, research institutions, publishers and solution vendors. Her work focuses on the changing landscape of scholarly communications in light of the open science movement. Nobuko is a librarian by training and earned an MLIS from the University of Hawaii at Manoa.

 <https://orcid.org/0000-0002-3229-5662>



Veliswa Tshetsha

Veliswa Tshetsha is a Senior Coordinator for Open Scholarship at the University of Pretoria (UP) in South Africa. Her role is to deliver coordination of activities to ensure continued development and delivery of the UP's research outputs curation and dissemination service, particularly in the area of open access. Her key focus areas include engagement with researchers to advocate and promote best practice in research dissemination, ensuring policy compliance and research impact. She provides guidance, training, and support by delivering a range of services, which enable research outputs to be effectively exposed, managed, curated and measured.

 <https://orcid.org/0000-0002-9981-6960>



Rosina Ramokgola

Rosina Ramokgola is a Data Curation Officer, responsible for the UP Research Data Repository known as Figshare and Research Data Management, under Scholarly Communications (Research Data Management Unit) at the Department of Library Services, University of Pretoria. She is a Tuks Alumni and holds a master's degree from University of Pretoria. Rosina has been a member of the University of Library Services for a decade where her career has been focused primarily on client services. Currently, her focus is data literacy, data science, and research data management. She is an active member of NeDICC (Network of Data and Information Curation Communities) and Library Carpentries. She is passionate about new trends in the Library, particularly those that enhance our service.

 <https://orcid.org/0000-0001-6645-4628>



Pfano Makhera

Pfano Makhera is a Metadata Specialist at the University of Pretoria's Department of Library Services. Her main focus is on descriptive metadata creation on the Open Scholarship and Research Data Repositories. Her background and passion for cataloguing influences her adherence to metadata standards in order for her work to meet international metadata standards. Since taking her role as a metadata specialist, her recent interest now includes RDM and Open Science.



Prof Ginny Barbour

Ginny Barbour is Director of Open Access Australasia and is Co-Lead, Office for Scholarly Communication, Queensland University of Technology (QUT). She was one of the three founding editors of PLOS Medicine. She was Chair of COPE (2012-2017), and is currently Vice-Chair of DORA Steering Committee, a Plan S Ambassador and member of the NHMRC's Research Quality Steering Committee. She is an editorial advisor to medRxiv.

 <https://orcid.org/0000-0002-2358-2440>

Part of **DIGITAL**science



digital-science.com



10 years of figshare

Available online also:

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