The State of Open Data 2023

The longest-running longitudinal survey and analysis on open data.

With opening remarks from Springer Nature's CPO, **Harsh Jegadeesan**, and Digital Science's CEO, **Daniel Hook**. Authors **Mark Hahnel**, **Graham Smith**, **Niki Scaplehorn**, **Henning Schoenenberger** and **Laura Day**.

DIGITAL science

SPRINGER NATURE

figshare

Contents

01

Opening remarks

"Over the past eight years, open data and open research have undergone a rapid transition – from being academic concepts and the domain of the few, to becoming more widely accepted and, in some countries, mandatory for researchers and institutions.

At the same time, The State of Open Data Report has become a unique, long-term resource chronicling the establishment of open data, attitudes towards it, and researchers' experiences of data sharing.

At Digital Science, we're proud of the role we've played in capturing researchers' sentiments toward and experiences of open data during a time when both practice and community have found their footing. We hope that the act of reporting these early steps back to the wider research community has helped to contextualize and normalize good practice in why and how research data should be shared. Furthermore, these reports themselves, together with the data that underpin them, now form a valuable resource to the community to understand the longitudinal development of open data practices and sentiments in our community."

Dr Daniel Hook,
CEO
Digital Science

"Researchers are at the heart of what we do. Understanding their motivations for engaging in open research practices is critical if we, as a community, are to develop sustainable routes towards open science. But why is open science important? Because ensuring easy and open access to all parts of research supports accessibility, usability and re-usability – and this is key in helping to ensure research can be built upon and gets into the hands of those that can effect change to tackle the world's most challenging issues.

This is central to our mission at Springer Nature, and The State of Open Data report, the only industry report offering an annual snapshot of open research trends, plays a central role in this. The data it provides enables publishers, funders and institutions to gain clear insights from researchers, helping us understand the roles we need to play in driving accessible research. This year, we have expanded that further with the first publication of a partner report by the Computer Network Information Center of the Chinese Academy of Sciences, looking at open data in China.

Together with our partners Figshare and Digital Science, we are delighted to present this report, and continue to work collaboratively to better understand and drive the solutions needed to support data sharing."

Harsh Jegadeesan,
Chief Publishing Officer
Springer Nature

02

About the survey

Background

The State of Open Data survey continues to provide a detailed and sustained insight into the motivations, challenges, perceptions and behaviours of researchers towards open data. The survey is a collaboration between Figshare, Digital Science and Springer Nature.

The State of Open Data survey contains approximately 55 questions and is intended to take 20 minutes for respondents to complete. There are also 3 open questions, all of which are optional. The survey is translated into French, German, Japanese and Simplified Chinese. An incentive of five $100 gift cards was offered through a prize draw, with ten stuffed toys for winners of a prize draw in China.

There were 6,091 usable responses from the State of Open Data survey this year. The survey received a total of 7,042 responses, some of which were screened out during the data-cleaning process. Please note that some of the questions in the survey were not mandatory or relevant to all respondents.

The largest proportion of responses were completed in English, representing 77% of the sample. This was significantly higher than the languages of other respondents, with the second largest being Chinese at 10%, followed by Japanese and German which each made up 5% of responses. The smallest proportion of responses were completed in French, at 4%.

Professional experience

All respondents were questioned to ensure they were either currently or had recently participated in developing research results. Respondents who had not published within the last 5 years were screened out of the survey; in total, 332 responses were screened out at this stage.

79% had published or submitted a manuscript within the last year, while 15% did so within the last 1-2 years, and 6% within 3-5 years. This is slightly lower than previous responses – 82% of respondents in 2022 stated they had published or submitted a manuscript within the last year, and 25% within the last 5 years.

When asked in what year their first peer-reviewed research article was published, 24% of respondents reported this as being before 2000. This is also a decrease compared to 2022, where a third of respondents reported publishing their first peer-reviewed article before 2000.

When asked about their job title, 38% selected senior research roles (professors, associate professors, research directors or lab directors), while another 22% selected early career roles (PhD/master's students, postdoctoral candidates, research assistants or undergraduate students). In comparison, in 2022, 31% of respondents classed themselves as senior researchers, and 31% as early career researchers.

Field of interest
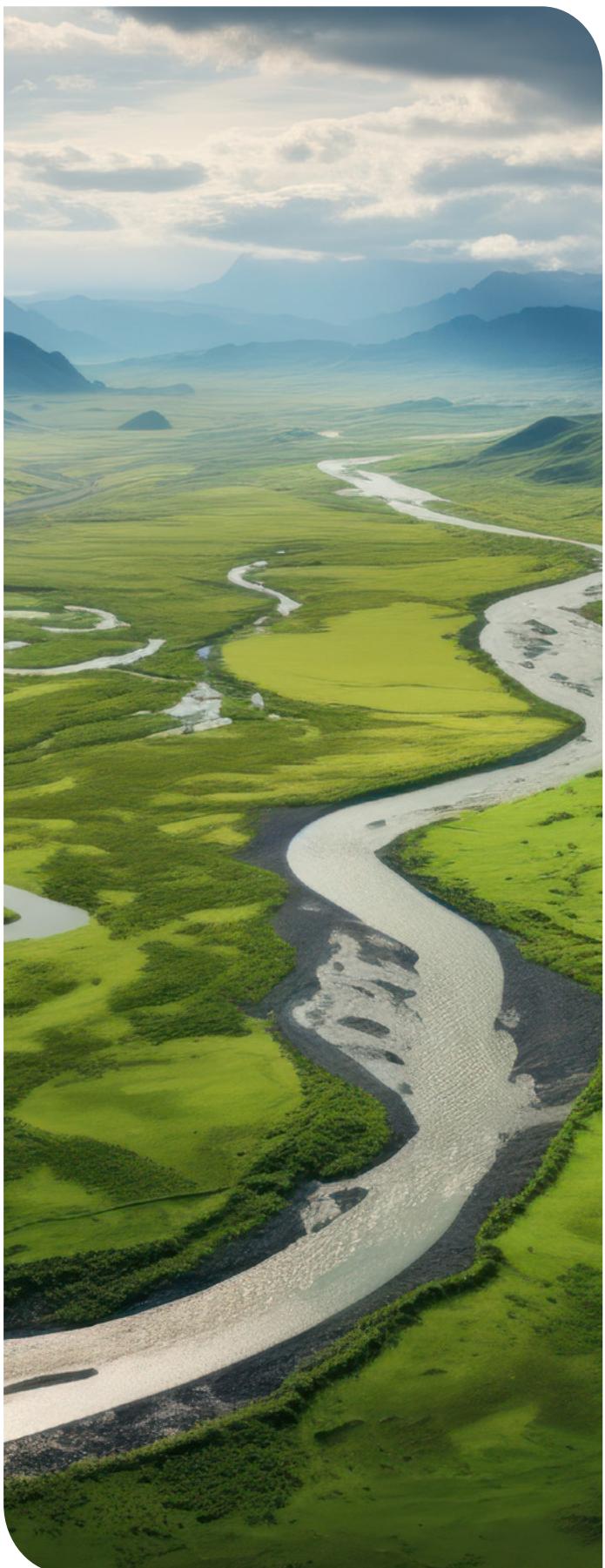
The survey received responses from a broad range of disciplines. The largest proportion of respondents were from medicine (23%) which is consistent with previous responses – in 2022, 23% were also from this field. This was followed by biology (16%) and engineering (10%). The proportion of responses from social sciences (9%) and earth and environmental sciences (8%) were also consistent with 2022, where 9% and 7% of respondents reported working in these fields, respectively.

Respondent institutional information

The majority of respondents said they worked or studied in a university (64%), up from 54% in 2022. This is followed by research institutions (13%), which was similar in 2022, at 14%. A further 7% worked in a hospital or other healthcare setting and 5% worked at a medical school.

Regional comparison

The country with the largest individual response size was India, representing 12% of responses. This was followed by China at 11% and the US at 9%. This differs slightly from 2022 – the highest response rates were from China and the US, which each accounted for 11% of responses. Other nations with response rates of over 100 were Germany (6%), Japan (5%), Italy (4%), Ethiopia and the UK (each at 3%), as well as Turkey, Brazil, Spain, Canada, Pakistan, Egypt, France and Nigeria (each making up 2% of responses).

03

Key takeaways from the State of Open Data 2023

Support is not making its way to those who need it

Almost three-quarters of respondents had never received support with making their data openly available.

One size does not fit all

Variations in responses from different subject expertise and geographies highlight a need for a more nuanced approach to research data management support globally.

Challenging stereotypes

Are later career academics really opposed to progress? The results of the 2023 survey indicate that career stage is not a significant factor in open data awareness or support levels.

Credit is an ongoing issue

For eight years running, our survey has revealed a recurring concern among researchers: the perception that they don't receive sufficient recognition for openly sharing their data.

AI awareness hasn't translated to action

For the first time, this year we asked survey respondents to indicate if they were using ChatGPT or similar AI tools for data collection, processing and metadata creation.

Support is not making its way to those who need it

Almost three-quarters of respondents have never received support with planning, managing or sharing research data.

With the global increase in policies and mandates to share data openly, who researchers are approaching for support becomes a pertinent question.

If respondents stated that they were aware of the concept of a data management plan, they were then asked if they have access to support from specialist data managers and we saw over 50% of our respondents state that they do have access to specialist research data managers in their research setting, but who else has been providing support?

Almost three quarters of respondents **had never received support** with planning, managing or sharing research data. When respondents were asked if they had ever received support with managing or making their data openly available, only 23% said they had. Of that 23%, 61% received support from informal internal sources such as colleagues or supervisors. Two other sources of support ranked highly with our respondents; their institutional libraries (31%) and their research office/in-house institutional expertise (26%).

Do you have access to support from specialist data managers?

- No — 781
- Yes, based in my department — 220
- Yes, based in my lab — 150
- Yes, based in my library — 326
- Don't know — 40
- Other, please give details — 429

Graph showing the responses to the question *Do you have access to support from specialist data managers?* This question was only asked if respondents stated that they were aware of the concept of a data management plan. This graph shows the number of respondents for each answer.
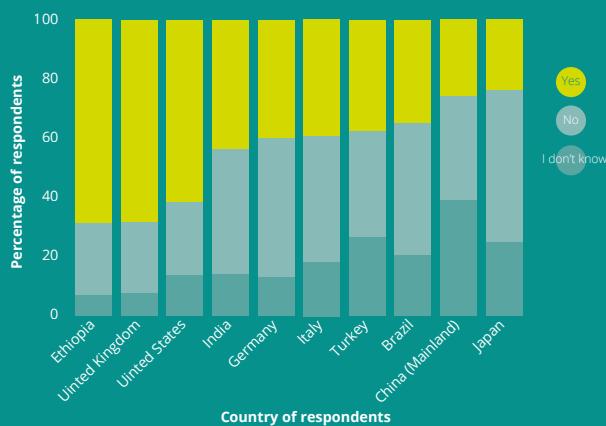
One size does not fit all

We need a nuanced approach towards research data management globally
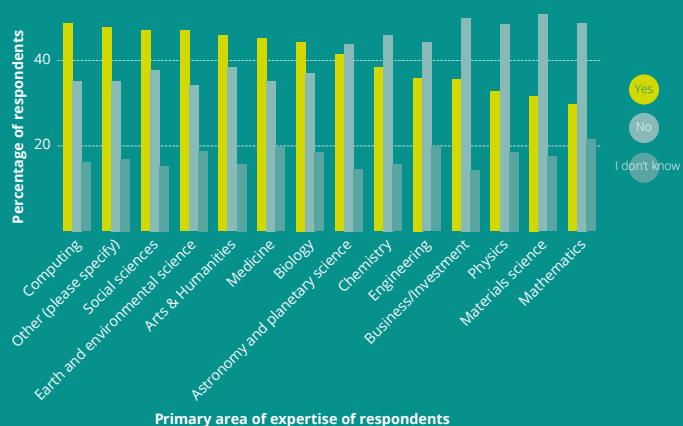
The variation in responses from different geographies in The State of Open Data 2023 clearly indicates that there are vast differences in the current approaches and attitudes towards data sharing around the world. When looking at key awareness-focused questions such as whether respondents are aware of the concept of a data management plan, we see considerable variation across different regions.

Awareness of DMP by country (top 10 countries)

Awareness levels of the concept of a data management plan broken down by countries. This graphic highlights the awareness levels from the ten countries from which we saw most respondents.
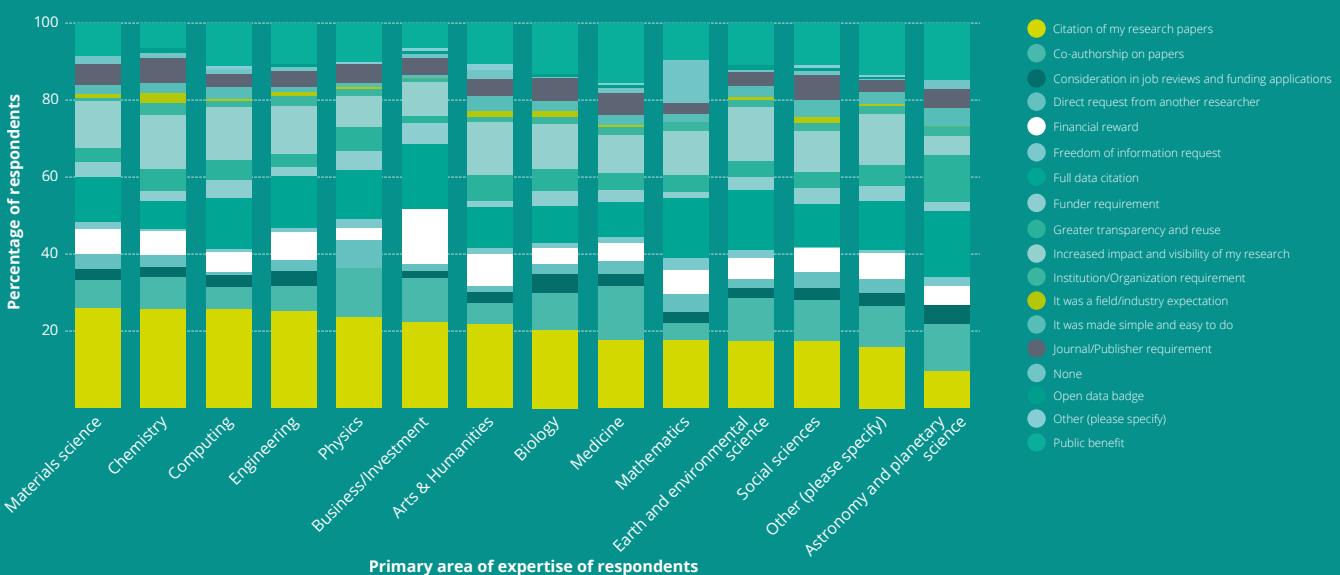
Awareness of the concept of a data management plan by primary area of expertise

Awareness of the concept of a data management plan broken down by primary area of expertise and ordered by percentage of those that said 'yes'.

There is also significant variation when we break the respondents down by their primary discipline. Differing levels of data management plan awareness are evident from different expertise areas, as well as different motivations for data sharing.

Motivations for sharing data openly by primary area of expertise

Percentage distribution of responses to what the primary circumstances were that would motivate them to share their data, broken down by primary area of expertise.

These varying responses highlight that a nuanced approach to research data management is needed to proficiently encourage and support data sharing worldwide.
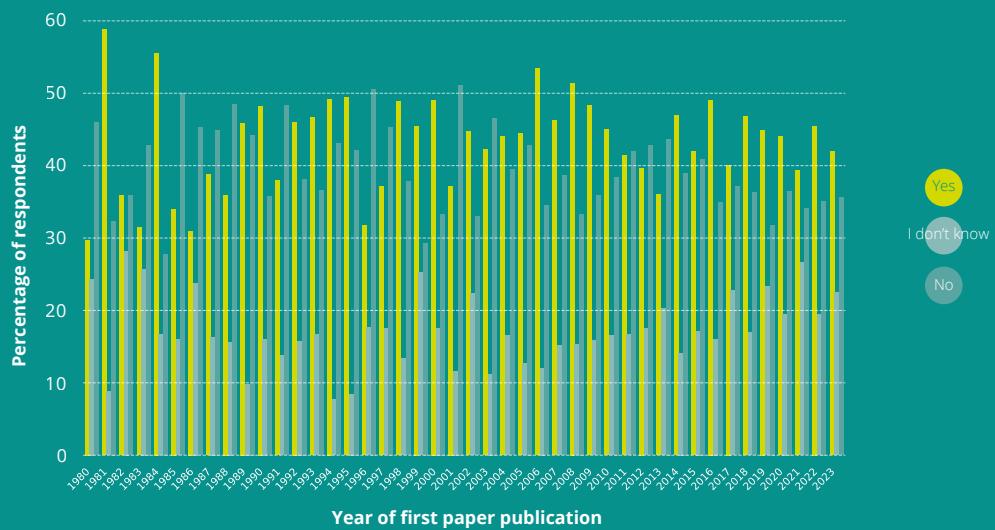
Challenging stereotypes: are later career academics really opposed to progress?

Career length is not a significant factor in open data awareness or support levels
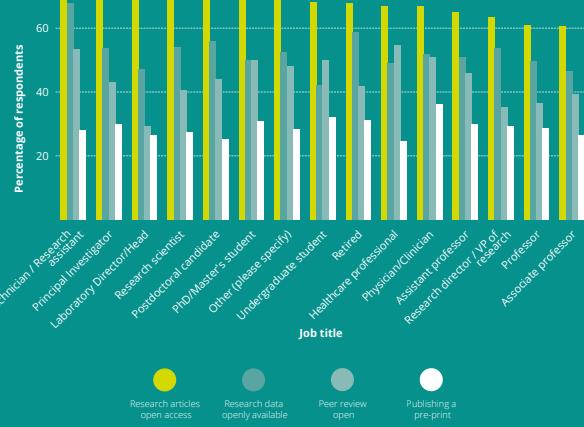
From the responses to the 2023 survey, it appears that the length of academics' careers is not a significant factor in terms of awareness levels of the concept of a data management plan or support levels of data being made openly available common practice.

Awareness of data management plans by first year of publication

Percentage of respondents

Year of first paper publication

Legend: Yes / I don't know / No

Awareness of the concept of a data management plan broken down by the first year that the respondent published a peer-reviewed article.
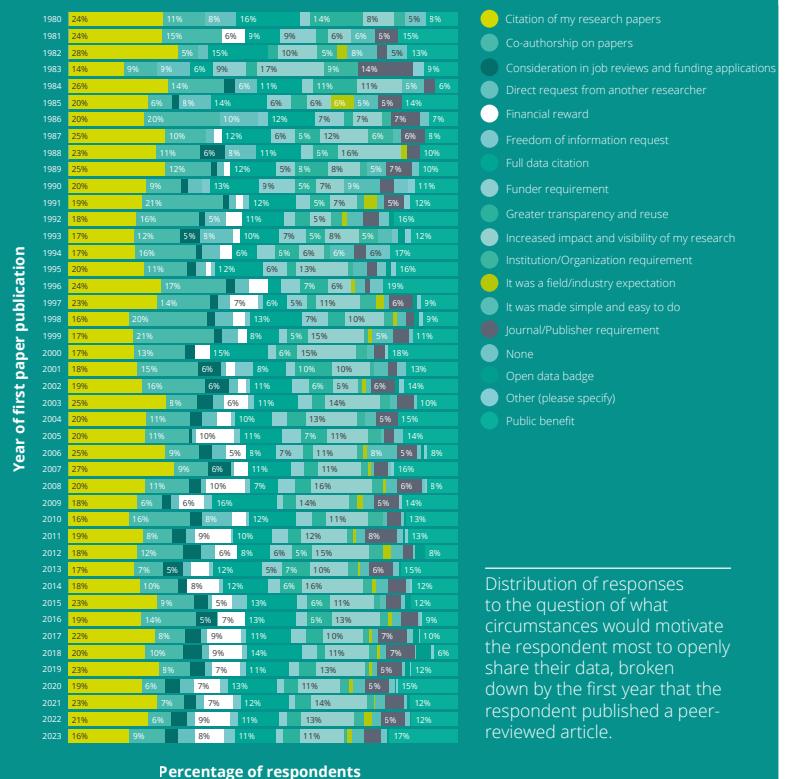
This indicates that the idea that early career researchers are driving data sharing forward and that more established, later career academics are opposed to progress in the space is a misconception. Researchers spanning all career lengths are meeting the same challenges and share the same motivations for data sharing.

Percentage of 'strongly agree' responses for selected questions by job title

Percentage of respondents

Job title: Technician / Research assistant, Principal Investigator, Laboratory Director/Head, Research scientist, Postdoctoral candidate, PhD/Master's student, Other (please specify), Undergraduate student, Retired, Healthcare professional, Physician/Clinician, Assistant professor, Research director / VP of research, Professor, Associate professor

Legend: Research articles open access / Research data openly available / Peer review open / Publishing a pre-print

Percentage of 'strongly agree' responses to the question of whether four selected core open practices should be 'common scholarly practice', broken down by job title.

Motivations for data sharing by publication year

Year of first paper publication

Percentage of respondents

Legend:
- Citation of my research papers
- Co-authorship on papers
- Consideration in job reviews and funding applications
- Direct request from another researcher
- Financial reward
- Freedom of information request
- Full data citation
- Funder requirement
- Greater transparency and reuse
- Increased impact and visibility of my research
- Institution/Organization requirement
- It was a field/industry expectation
- It was made simple and easy to do
- Journal/Publisher requirement
- None
- Open data badge
- Other (please specify)
- Public benefit

Distribution of responses to the question of what circumstances would motivate the respondent most to openly share their data, broken down by the first year that the respondent published a peer-reviewed article.

Credit is an ongoing issue

## Researchers still do not feel they receive appropriate credit for openly sharing their data

For eight years running, The State of Open Data survey has revealed a recurring concern among researchers: the perception that they don't receive sufficient recognition for openly sharing their data. The 2023 survey responses continue this trend, highlighting an ongoing issue within the research community that needs to be addressed in the future.

Do you think researchers currently get sufficient credit for sharing data?

Longitudinal survey data from 2019-2023 for the question 'Do you think researchers currently get sufficient credit for sharing data?'

In terms of what would motivate researchers to share their data, the responses remain very similar to previous years, with full citation of research papers or a data citation ranking highly. However, '***public benefit***' was the second most popular selection for respondents when asked 'which circumstances would motivate you most to share your data?'

Which of these circumstances would motivate you most to share your data?

Circumstance	Percentage
Citation of my research papers	20%
Public benefit	12%
Increased impact and visibility of my research	12%
Full data citation	11%
Co-authorship on papers	10%
Financial reward	6%
Journal/Publisher requirement	5%
Greater transparency and reuse	5%
Funder requirement	4%
Consideration in job reviews and fundingapplications	3%
Direct request from another researcher	3%
It was made simple and easy to do	3%
Institution/Organisation requirement	2%
None	1%
Freedom of information request	1%
It was a field/industry expectation	1%
Open data badge	1%
Other (please specify)	0%

Percentage of respondents

Graph showing the percentage of respondents that would be motivated by certain circumstances to share their data openly.

AI awareness hasn't translated to action

The space is nascent but opportunities abound
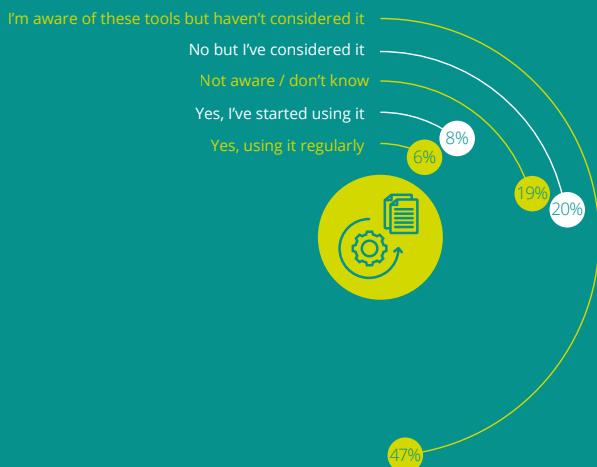
For the first time, this year we asked survey respondents to indicate if they were using ChatGPT or similar AI tools for data collection, processing and metadata collection.

The most common response to all three questions was
*'I'm aware of these tools but haven't considered it.'*
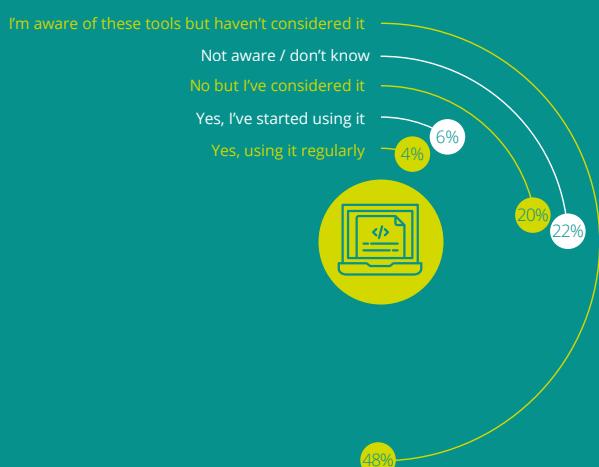
Are you using ChatGPT or similar AI tools for data collection?

I'm aware of these tools but haven't considered it
No but I've considered it
Not aware / don't know
Yes, I've started using it
Yes, using it regularly

7%
5%
18%
19%
51%

Are you using ChatGPT or similar AI tools for data processing?

I'm aware of these tools but haven't considered it
No but I've considered it
Not aware / don't know
Yes, I've started using it
Yes, using it regularly

8%
6%
19%
20%
47%

Are you using ChatGPT or similar AI tools for metadata creation?

I'm aware of these tools but haven't considered it
Not aware / don't know
No but I've considered it
Yes, I've started using it
Yes, using it regularly

6%
4%
20%
22%
48%

We have now benchmarked researchers' use of ChatGPT and similar AI tools in regards to research data and its management and we're looking forward to seeing how these responses develop in years to come, in light of the fast moving space of AI tools and their applications.

DIGITAL science

SPRINGER NATURE

figshare

Mark Hahnel.
Founder, CEO - Figshare,
Digital Science

04

Key insights: Analysis to action. The need for a nuanced approach to the momentum of open research data

The journey through the eight years of The State of Open Data survey has witnessed a remarkable evolution in the open data sphere, where the confluence of technology, policy, and community engagement has sculpted a dynamic and resilient environment. From fostering transparency to catalysing innovation, the impact of open data continues to be felt. Kicking off the commentary around the data (all of which is open for your own perusal here: https://doi.org/10.6084/m9.figshare.24517123), we must not forget the end goal of open academic data. In an era where data is often heralded as the 'new oil,' we should recognize the value and potential in driving innovation, policy-making, and societal advancements that open academic data holds.

The 2023 report is intentionally more analytical than in previous years. As the report has gained international traction, we want to provide as much detail as possible in this original 2023 report. This year also sees the launch of a partner report from the Computer Network Information Center of the Chinese Academy of Sciences - using The State of Open Data survey results to guide their understanding of how Chinese researchers are responding to data publishing requirements. The variance observed across different geographical locales, research disciplines, and career stages, however subtle, has also prompted us to create some recommendations for stakeholders in order to recognize disparities and address them through targeted interventions, policies, and support mechanisms. We anticipate and encourage further feedback on the report and national, funder or even organizational analysis of the data published alongside the report.

As a longitudinal survey, we want to track how both emotions and actions are driving an exponential increase in the number of datasets published year on year. In the survey data, we can see a shift as researchers respond to new funder policies coming into place, a focus on trust in a post-COVID world and the rapid emergence of new tools and technologies, such as artificial intelligence (AI). We see
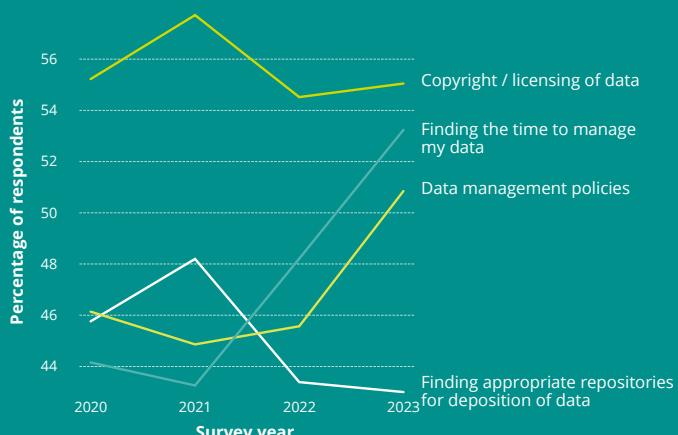
progress in researchers adopting data management plans and dig into the importance of this later in the report. This is an early signal with regards to researchers recognising the importance/or expectations their funders and institutions are placing on data.

The big challenges in promoting the publication of open academic data remain the same: Credit and concerns. Graham Smith talks more about these concerns in his key insights piece: 'Beyond policy, meeting researchers' practical needs'. We may also be seeing some fatigue in enthusiasm for open research in general as open data policies come into place and researchers find themselves with even less time to comply with the directives of their funders. 54% of respondents felt that funders should make sharing research data part of their requirements for awarding grants. In 2022, 52% agreed – slightly less. However, this proportion has dropped from 2019, where 69% agreed.

- Respondents from universities were more likely to disagree (26%).
- Similar to the national mandates, professors were less likely to agree to make the sharing of research data part of their requirements.
- The proportions were significantly higher for open science advocates.
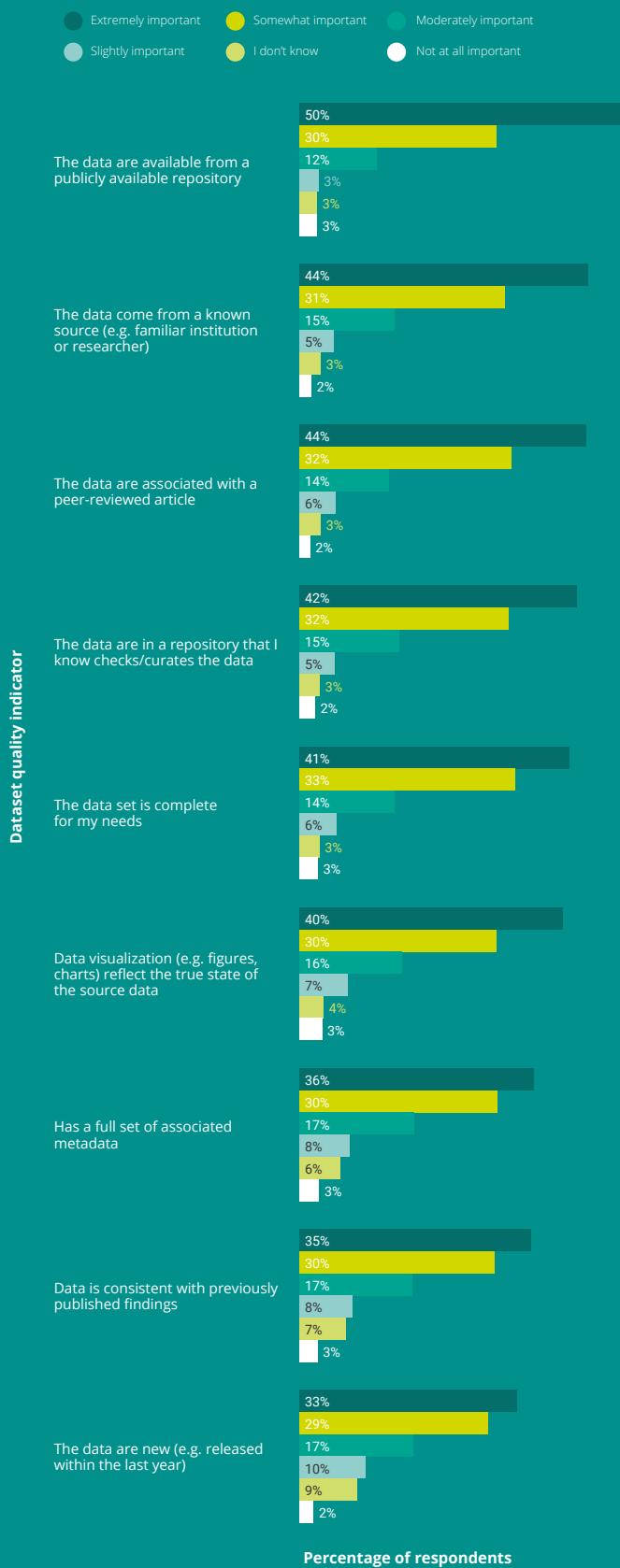
46% of respondents agreed that they felt funders should withhold funding from, or penalize, researchers who do not share their data if it was previously mandated by the funder at the grant application stage. This was two percentage points higher than in 2022, but far lower than in 2019, when the agreement level was 69%. There were three areas which were common for respondents to feel they needed help with in regard to making their data openly available. 'Copyright/licensing of data' (55%), 'Finding the time to manage their data' (53%) and 'Understanding the data management policies that apply to them' (51%) were all selected by over half of the respondents, significantly high proportions.

What areas, if any, do you feel you need help with in regard to making your research data openly available?

Graph showing longitudinal data of what areas respondents feel they need help with in regard to making their research data openly available.

Finding appropriate funds was also a highly chosen area, with 45% of respondents selecting it. Interestingly, last year, the same concern about data rights and licensing topped the list with 55% of researchers wanting more guidance on it.

Citation credit
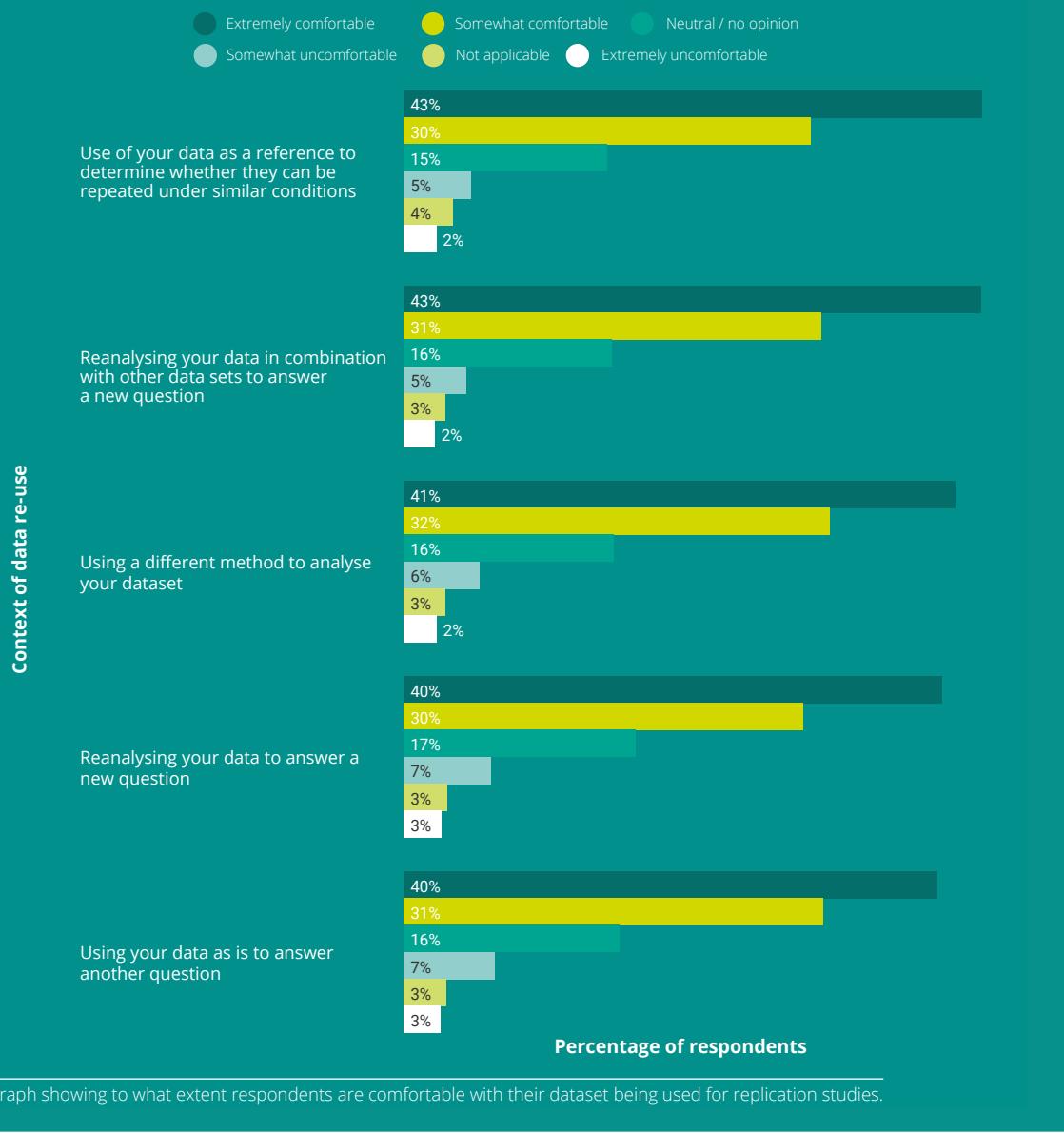
Researchers consistently say they will be motivated by citations of their data and their research papers. While paper citation is rewarded in the academic career progress scoring system, credit for data citations remains low. Tracking of data citations themselves is a nascent space. Make Data Count is soon to release a Data Citation Corpus, which will enable repositories to consistently track the impact of data citations. This provides funders with a stepping stone to start giving career-advancing credit to researchers. There is of course the question of whether researchers are actually re-using open data. One interesting area of the survey is investigating the perceived quality of open datasets, as a proxy to 'trustworthy data' and how comfortable those researchers are at having their own data re-used. The graph to the right suggests there are several essential criteria that need to be met by a dataset in order for it to be considered trustworthy. Novelty and previous findings are less relevant than the data being findable, accessible, interoperable and re-usable (FAIR).

We hear the success stories of large-scale academic projects making use of FAIR data and artificial intelligence (AI) to drive systemic change in a research field (such as DeepChem, ClimateNet and DeepTrio), but need to dig deeper into academic motivations to re-use datasets. Interestingly, data being publicly available in an open repository came out as the most important factor when determining the quality of

How would you rate the importance of each of the following with reference to determining the quality of an openly accessible dataset?

Graph showing how respondents rate the importance of certain criteria for determining the quality of an openly accessible dataset.

a dataset. We also see a little wariness in researchers about others 're-interpreting' their data when compared to eg. re-use for replication. Researchers are more comfortable than uncomfortable when it comes to their own data being re-used, as highlighted in the graph below. However, the fact that nearly 10% of those surveyed are still uncomfortable about their data being re-used in any format highlights potential roadblocks in the path to fully open research at scale.

This year's report takes a more nuanced approach to the survey responses. We investigate whether the results are consistent across countries, research subjects and the career stage a researcher is at. It feels like the right time to do this. While a global funder push towards FAIR data has researchers globally moving in the same direction, it is important to recognize the subtleties in researchers' behaviours based on variables in who they are and where they are.

To what extent are you comfortable with others using your data for replication studies

Legend:
- Extremely comfortable
- Somewhat comfortable
- Neutral / no opinion
- Somewhat uncomfortable
- Not applicable
- Extremely uncomfortable

Context of data re-use

Use of your data as a reference to determine whether they can be repeated under similar conditions
- 43%
- 30%
- 15%
- 5%
- 4%
- 2%

Reanalysing your data in combination with other data sets to answer a new question
- 43%
- 31%
- 16%
- 5%
- 3%
- 2%

Using a different method to analyse your dataset
- 41%
- 32%
- 16%
- 6%
- 3%
- 2%

Reanalysing your data to answer a new question
- 40%
- 30%
- 17%
- 7%
- 3%
- 3%

Using your data as is to answer another question
- 40%
- 31%
- 16%
- 7%
- 3%
- 3%

Percentage of respondents

Graph showing to what extent respondents are comfortable with their dataset being used for replication studies.

The State of Open Data survey aspires to continue being a compass; guiding efforts, informing strategies, and illuminating the pathway towards an open data future that is not just robust but also inclusive and equitable. As we dig deeper into the findings of this report, it is important to not only celebrate the milestones achieved but also to cognitively engage with the challenges and opportunities that lay embedded in the data. As such, we are also providing some recommendations to different audiences in order to help move the space further, faster. May this call to action serve as a catalyst for informed discussions, strategic alliances, and innovative solutions in our collective journey towards a more open, inclusive, and data-driven future.

■
Graham Smith, Open Data
Programme Manager at
Springer Nature

05

Key insights: Beyond policy, meeting researchers' practical needs

Since its inception, The State of Open Data has charted the shifting landscape of data-sharing policies, with researchers increasingly facing requirements to make data available for evaluation and re-use. In 2023 a number of these have come into practical effect, for example, the NIH Data Management and Sharing Policy, and deadlines for implementing the WHOSTP Nelson memo into US federal agencies' policies.

Already we see that researchers publishing in the last year are significantly more likely to share data due to a funder requirement than those publishing earlier. However there are also clear challenges in the level of support for researchers to comply with data policies; almost three-quarters of respondents had never received support with planning, managing or sharing research data. Where support is provided, it is more often from informal channels, such as colleagues and supervisors. This is noteworthy, particularly when half of the respondents who were aware of data management plans also indicated that they had access to data managers or curators at their institution.
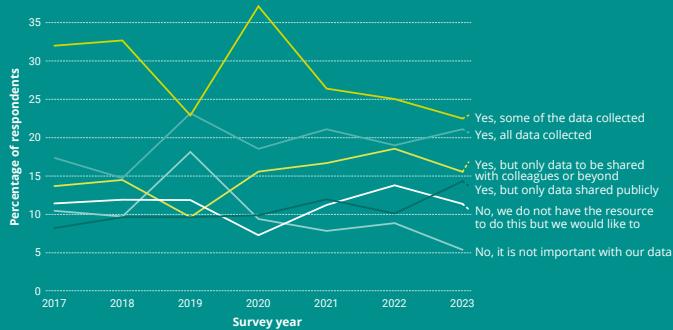
We can infer that effective support is not making it to researchers, while they are being squeezed by growing requirements. The data also suggests areas of improvement for those providing support; those who had published or submitted a manuscript within the last year were significantly less likely to describe their support levels as 'good' or 'excellent'. Additionally, these respondents were less likely to rely on funders and professional third parties for support.

Reducing the burden on researchers

Looking at the areas researchers tell us they need help, an ever-present issue has been that of time to curate data. While any data curator will tell you good data management takes time, certain main challenges in this area seem solvable, for example: 'finding appropriate repositories for deposition of data' (identified by 41% of respondents). Interestingly this is highlighted more (45%) in biological

sciences where, broadly speaking, the establishment of data repositories is more mature than in other research areas. This suggests that while solutions exist, they are not easily accessible in the research lifecycle.

During or after data collection, do you curate/prepare your data for sharing (whether privately or publicly)?

Graph showing the percentage of respondents who curate/prepare their data and to what extent.

Publishers can play a key role in reducing this burden. At Springer Nature, our vision for open science encompasses a set of open outputs such as data, code and protocols, all linked via the research article. We aim to make open science easily accessible, more prominent in our journals and embedded across disciplines, with authors empowered to share their data, opening them up for further re-use and interrogation.

One initiative in this area is the standardization of our research data policy which will embed the requirement for Data Availability Statements across over three thousand journals. This move is designed to increase transparency around underlying data and enhance the integrity of the scientific record. As part of the change, we are making author and editor guidance more straightforward, supporting authors and editors with compliance.

Alongside policy requirements, the **rollout of the Figshare integration across the Nature Portfolio** is providing a practical means to make data sharing in repositories easier for authors. This streamlined tool has seen great uptake from authors, with 7,500 data submissions since its launch in April 2022, equivalent to 15% of manuscript submissions. The first year's worth of publications since integration shows more authors are using repositories overall (a 12% increase), which supports our hypothesis that more prominent and accessible data solutions can have a demonstrable impact on data sharing at a journal. The initiative has since expanded to 37 Nature Portfolio titles, including *Nature* and *Nature Communications*.

Complementing our data initiatives is a similar code-focused integration with Code Ocean and the acquisition of protocols. io, both strengthening our vision for linked open science outputs available alongside the article.

But in order to take into account future developments and to meet the constantly evolving needs and requirements of the scientific community, our work must not stop here. A range of commentators in The State of Open Data reports over the years have highlighted the need for different entities in this space to work together to solve the challenges of data sharing and accessibility. Bodies such as the Research Data Alliance, CODATA and FORCE11 have provided invaluable forums for cooperation between institutes, researchers, publishers, funders and numerous other actors. The ongoing collaboration between the Computer Network Information Center of the Chinese Academy of Sciences, Springer Nature, Digital Science and Figshare on The State of Open Data is another such positive sign in global data collaborations. One thing The State of Open Data results tell us is there are solvable challenges that need practical solutions to back up this cooperation.

One area we will undoubtedly see growing in response is AI. While the results from this year's survey don't yet show a clear picture, data on which issues to address will be key as we think about the tasks best suited for automation. At a time when much generative AI is focused on language, data-focused automation is a more nascent area but one that will likely play a large role in navigating current obstacles in research. Niki Scaplehorn and Henning Schoenenberger explore this further in the piece 'AI and open science - the start of a beautiful relationship?'

06

From generalized trends to a demographic-tailored approach

In all 8 editions of The State of Open Data, we have explored longitudinal trends in researchers' attitudes towards data publishing. As this nascent field develops, we can delve deeper into the nuanced thinking among various researchers, considering their location, primary research field, and even research seniority. What follows begins to unravel how differences in the types of research being conducted can significantly affect researcher attitudes. Factors like cultural norms, funding, data privacy laws, perceived data value, and lack of resources can influence attitudes towards data sharing.
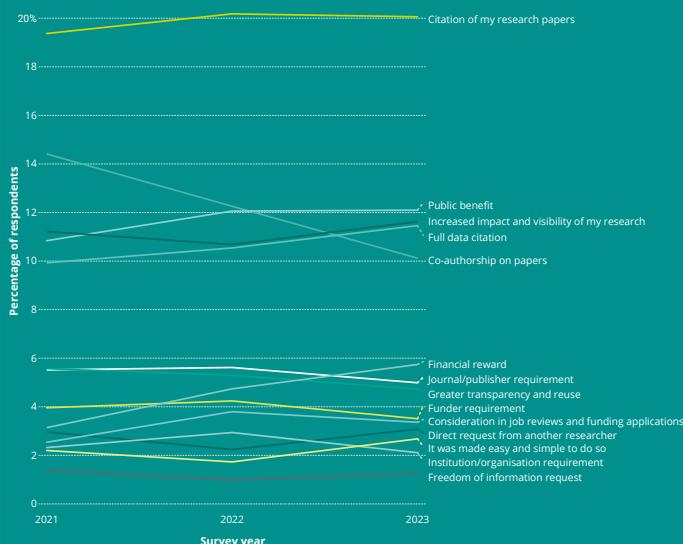
When it came to potential problems with sharing datasets, the most commonly reported was around 'The inclusion of sensitive information or requiring study participant permissions before sharing' (39%). Those within hospitals or other healthcare settings were significantly more likely than other organization or institution types to select this (51%), as were those in private companies (50%). This concern was greater for those based in China (57%), Canada (54%) and the United Kingdom (53%) than other territories, suggesting that local legislation may play a role.

'Concerns about misuse of data' and 'Unsure about copyright and data licensing' were the next most reported concerns, with just under a third (32%) selecting each. China and the US were more likely to select 'Concerns about misuse of data'. Those who were concerned about the misuse of data were significantly more likely to work in social sciences (38%) or medicine (36%). A lack of clarity about copyright and data licensing was significantly more likely to be a concern for technicians/research assistants (45%).

Motivation to share data

The primary circumstance that would motivate respondents to share data was 'Citation of their research papers' (65%), which was also the top factor in 2022. Credit for dataset citation counts has not yet been quantified in a standard way. The Generalist Repository Ecosystem Initiative (GREI)

What circumstances would motivate you to share your data?

Graph showing the percentage of respondents motivated by different circumstances to share their data.

is working on this through its Data Citation working group. However, until new reward systems are embedded into career progress, paper citations will remain the key driver.
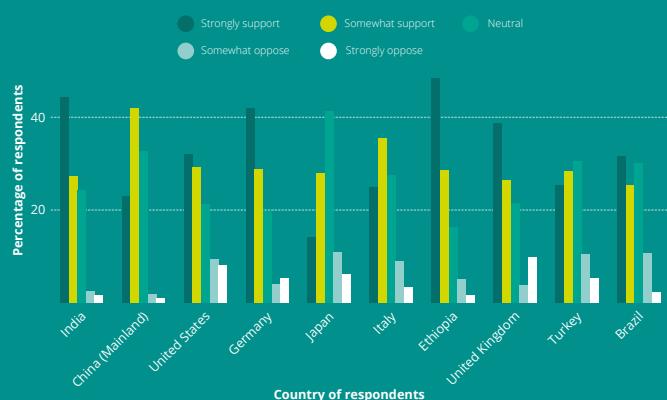
- Respondents working within computing were significantly more likely to select paper citations as their primary motivation for sharing data (75%) than those in other disciplines. Additionally, those based in Japan were significantly more likely to select this (75%) than those in other locations.

- Respondents working in astronomy and planetary science (73%), arts and humanities (67%), computing (67%) and earth and environmental sciences (62%) were all significantly more likely to see 'Full data citation' as a motivating factor in sharing their data. Respondents affiliated with research institutions (60%) shared this perspective. Moreover, respondents from China (69%), Germany (62%), and the United States (58%) were also

significantly more likely to select 'Full data citation' as their motivating factor.

National mandates

There was overall support for a national mandate for making research data openly available – 64% of respondents supported the idea. A little over a third (34%) strongly supported the idea. Only 11% opposed the idea, a significantly low proportion. Indian and German respondents were more likely to support this idea (both 71%). In fact, Japanese and Chinese respondents were more likely to be neutral around the idea of a national mandate (41% and 33%, respectively), than other countries. A similar pattern is seen when researchers were asked about their support for a mandate in their own country (below).
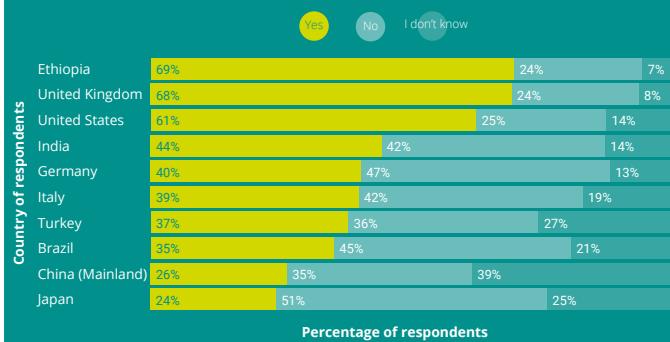
Graph showing the percentage of respondents that support national data mandates in their country, showing data for the 10 countries with the highest number of respondents.

When examining the ten countries with the most survey respondents, Ethiopia tops the chart in terms of the highest percentage of respondents who 'Strongly support' open data, followed by India, Germany, and the United Kingdom. Japan has the lowest percentage of respondents who 'Strongly support' open data. Interestingly, the biggest funder of Ethiopian-based publications is the Bill and Melinda Gates Foundation, which has a strong open data publishing policy: 'Each accepted article must be accompanied by a Data Availability Statement that describes where any primary data, associated metadata, original software, and any additional relevant materials can be found.'

The biggest funder of Japanese-based publications is the Japan Society for the Promotion of Science (JSPS). The second biggest Funder is Japan Science and Technology Agency,

which does 'require open data archiving' and since 2017 has required funded researchers to develop a data management plan defining how to manage research data, and to manage data accordingly. We observe similar results when assessing awareness of data management plans. Ethiopia has the highest awareness as a percentage of those surveyed, while Japan appears to have the least awareness.

Graph showing the levels of awareness of the concept of a data management plan, broken down by country, showing data for the 10 countries with the highest number of respondents.
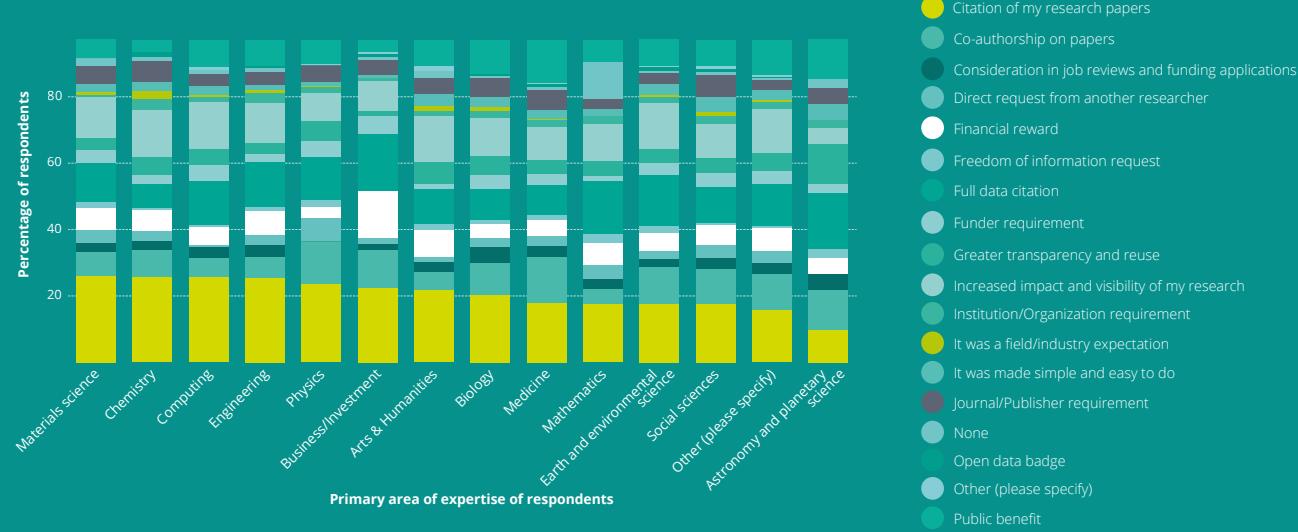
Subject-based trends

Nation-states may have different strengths in their policies and promotion of open academic data. The same could be true of research fields. Research always strives to be collaborative across multiple areas, but there is often a sentiment that single-subject-focused research domains can exist in a silo. Thus, it's unsurprising that researchers in different fields have varied reasons for choosing to share or not share their data. Coupled with this is a real heterogeneity in the outputs of research in different fields. Data is a broad term but can be simplified to mean the outputs that researchers create to back up the findings they publish in papers or monographs. As such, the outputs generated in the social sciences may be quite different from those in the life sciences.

The previous State of Open Data surveys have shown that motivations for sharing data include advancing science, increasing transparency, gaining recognition, fulfilling funding requirements, and promoting interdisciplinary collaboration. For the first time this year, we decided to examine differences in research fields. Across all subjects, a consistent message emerges that research impact and credit are key drivers when asked, 'What motivates you to share your data?'. In all fields, 'Citation of research papers,' 'Full data citation,' and 'Increased impact and visibility of my research'

score highly, but there does seem to be a subject-based preference for citation of papers, or citation of data. For example, as highlighted below, researchers from 'Astronomy and planetary Science' are more motivated by a 'Full Data Citation' than more citations to their papers.

Motivations for sharing data openly by primary area of expertise.

Graph showing the percentage of respondents that would be motivated by certain circumstances to share their data, broken down by subject area of expertise of the respondents.

Making data publicly available

89% of respondents make their data available publicly, with 11% having responded, 'I do not share my data beyond my immediate collaborators' (n=6091).

Most respondents indicated that they made their data publicly available 'On the publication of the research' (37%) or 'When the research is complete' (18%), suggesting this usually happens towards the end of the research cycle. Similarly, 13% selected that sharing happens 'In the process of submitting a research article', again indicating that sharing of data takes place in the latter stages.

• Respondents in biology were significantly more likely to make their data publicly available 'On publication of the research' (47%), whereas engineers were significantly more likely to do so 'When research is complete' (23%).
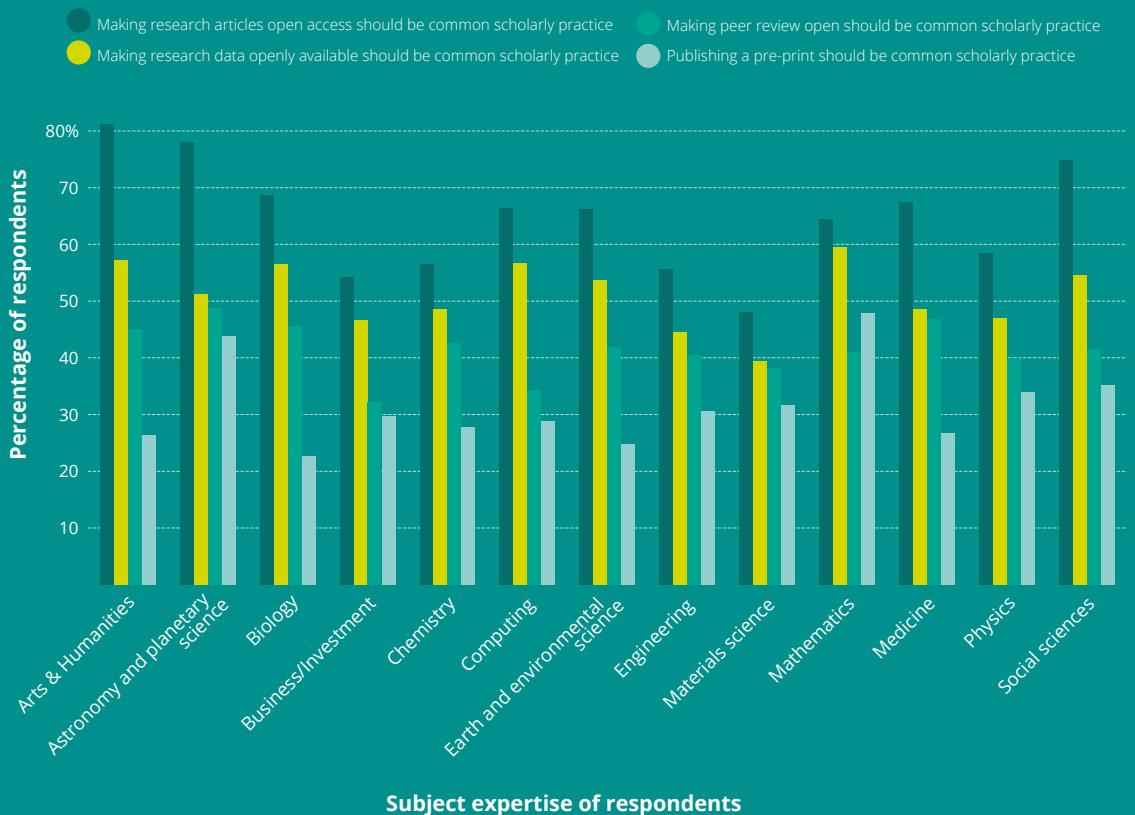
This being said, 17% indicated that they make their data publicly available only 'Upon request from others'; indicating that they do not routinely share their data as part of the

process. Similarly, 11% selected 'I do not share my data beyond my immediate collaborators', again suggesting sharing data is not the norm for all researchers.

• Those within medicine (20%) and social sciences (23%) were significantly more likely to share their data 'Upon request from others'.
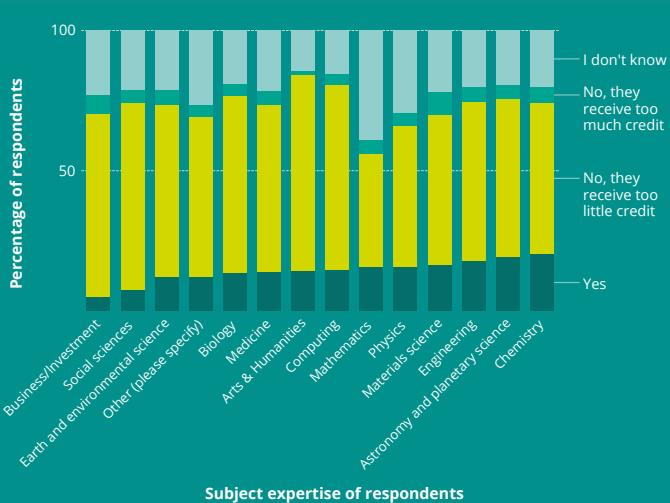
When we explore a general level of 'Open to Openness', we observe subject-based variations. The chart below shows what percentage of researchers 'Strongly agreed' with open practices around open research articles, datasets, peer review and preprints. The general trend is that researchers prioritize the importance of these outputs being open in that order, with open research articles evoking the strongest positive sentiment. Strong support for open data varies from 39% in Materials Science to 59% in Mathematics. Materials Science interestingly has less than 50% of researchers who 'Strongly Agree' with any of these open practices. This again highlights the need for a more subject-based approach to outreach and education around open practices.

Open to Openness by Subject (Percentage that 'Strongly agreed')

- Making research articles open access should be common scholarly practice
- Making research data openly available should be common scholarly practice
- Making peer review open should be common scholarly practice
- Publishing a pre-print should be common scholarly practice

Percentage of respondents

80%
70
60
50
40
30
20
10

Subject expertise of respondents

Arts & Humanities, Astronomy and planetary science, Biology, Business/Investment, Chemistry, Computing, Earth and environmental science, Engineering, Materials science, Mathematics, Medicine, Physics, Social sciences

Graph showing the levels of 'openness' to making certain open practices 'common scholarly practices' broken down by subject area of expertise of the respondents.

Do researchers receive enough credit for sharing their data?

Percentage of respondents

100

50

Subject expertise of respondents

Business/Investment, Social Sciences, Earth and environmental science, Other (please specify), Biology, Medicine, Arts & Humanities, Computing, Mathematics, Physics, Materials science, Engineering, Astronomy and planetary science, Chemistry

- I don't know
- No, they receive too much credit
- No, they receive too little credit
- Yes

Graph showing whether respondents feel researchers receive enough credit for data sharing, broken down by the subject area of expertise of the respondents.

Researchers do not receive enough credit for sharing their data, according to researchers in all major fields except Mathematics, where the majority of researchers responded with 'I don't know'. Arts and Humanities and Social Science researchers have the largest majority who think that they do not get sufficient credit for sharing data.

In summary, the variations in responses from different countries and subject areas suggest that a nuanced approach to Research Data Management (RDM) is necessary to effectively encourage and support data sharing. While some subjects may not align as seamlessly with open data practices as others, we have previously observed that when communities support open data, it more reliably becomes the norm. This is evidently less of a cultural shift for digital-born research practices, as observed in genomics. However, we have also witnessed the effects in older research fields such as ecology, where a concerted effort has driven open data into the perceived requirements of research publishing. A lack of subject-specific or thematic generalist repositories may lead to a one-size-fits-all approach, and this is something that could be addressed by societies and subject-based research communities.

Even with a global push towards mandating open data publishing, the follow-through from different countries and funders will largely define the speed of success and return on investment for the respective country. Now is the time to compare and contrast how your country is performing compared to those trying to reach similar goals.

Now is the time to compare and contrast how your country is performing compared to those trying to reach similar goals.

07

Researchers at all stages of their careers share the same optimism and concerns around open data

Planck's principle is the view that scientific change does not occur because individual scientists change their minds, but rather that successive generations of scientists have different views.

'A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die and a new generation grows up that is familiar with it'.

Intriguingly, this year's The State of Open Data report suggests that the differences observed between countries and primary subjects do not extend to researchers who have been publishing for a long time or to specific job descriptions. This challenges stereotypes and misconceptions about later career academics being opposed to progress and contradicts the message that early career researchers are the primary drivers of change in this area. Researchers, at all stages of their careers, face the same challenges, seek the same rewards, and demonstrate a similar level of openness to embracing open data.

Researchers, at all stages of their careers, face the same challenges, seek the same rewards, and demonstrate a similar level of openness to embracing open data.

It seems that the length of an academic career does not influence how aware researchers are of aspects like data management plans (DMPs), nor does it affect their motivations for sharing their data. There are indeed differences, but it is encouraging that there are not huge differences in the ways in which we have to educate academics on open data. We examined both academic job titles and the year in which a researcher first published a paper to gain deeper insights into their awareness and motivations regarding open data.

Incentives and recognition for sharing data are happening

Just over half (54%) of those surveyed already have received some form of recognition for sharing their data; the most common form of recognition received was 'Full citation in another article' (39%) followed by 'Co-authorship on a paper that used my data', which was selected by 23%. Other kinds of recognition included 'Consideration of data sharing in a job review' (7%), 'Consideration of data sharing in a grant application' (8%) and 'Open data badge' (4%).

- PhD/master's students were significantly more likely to have received 'Consideration of data sharing in a job review' (10%), as had research assistants/technicians (15%) and undergraduate students (27%).

However, over a third (37%) had never received any credit/recognition for sharing their data. 33% reported they had been involved in collaboration as a result of data they had previously shared. 60% reported that they hadn't.
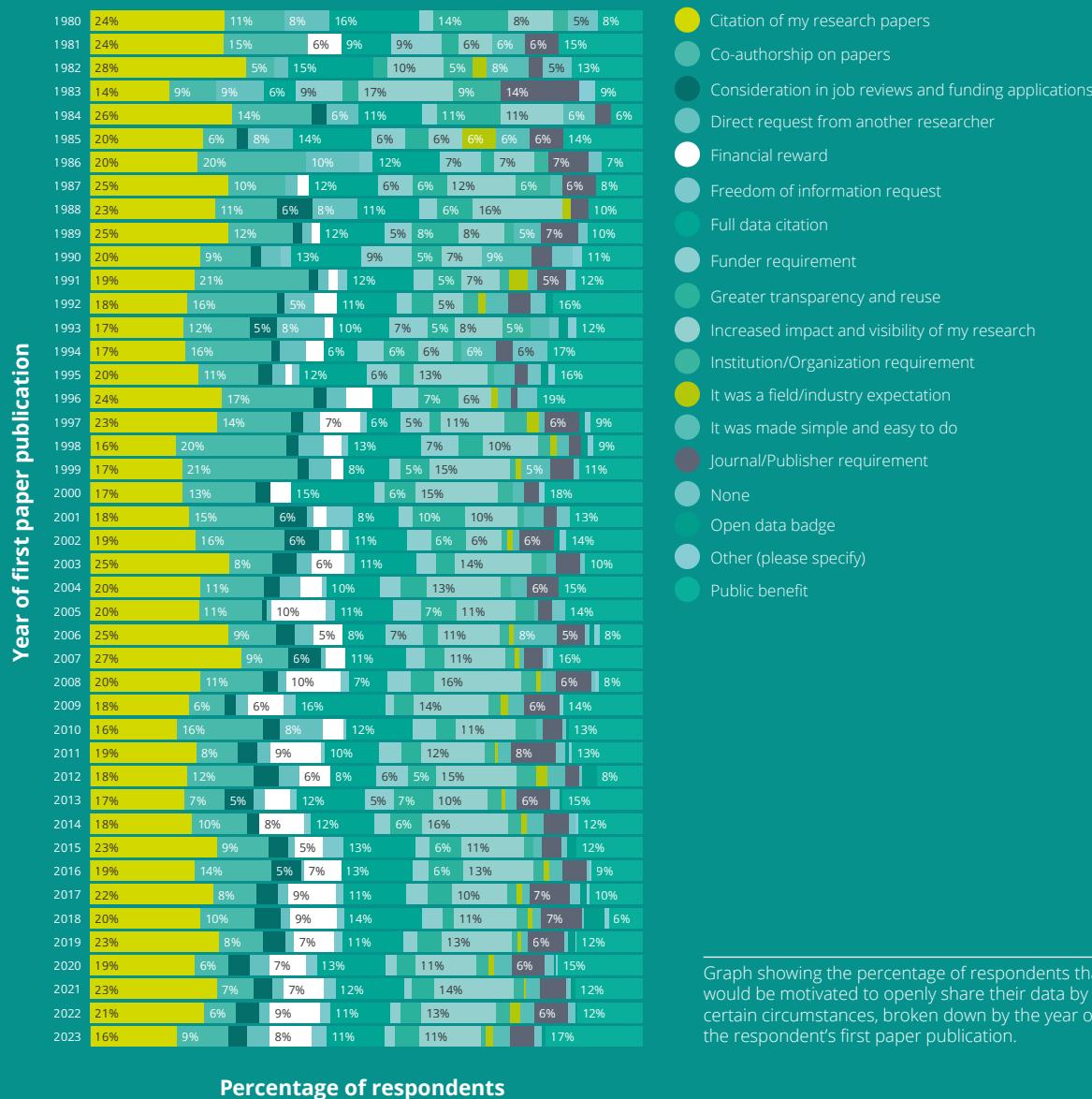
- Professors were significantly more likely to indicate that they had been involved in collaboration (39%), as were research directors or VPs of research (48%) – perhaps suggesting those in senior roles are more likely to get these opportunities. Those working within earth and environmental sciences also indicated this (40%).

- PhD/master's students were significantly more likely to indicate that they had not (69%), again suggesting collaboration opportunities may be more available to senior roles.

Trends based on year of first publication

The motivations for sharing research data are largely consistent among academics who published their first paper between 1980 and 2023. Each line in this chart represents one of those groups. While there is some variation year by year, no general trend is apparent. A consistent message prevails across all stages of academic career progression: 'Researchers receive insufficient credit for sharing their data openly.' Coupled with the motivations for sharing research, it appears that we can communicate with all researchers on the same level regarding how to garner more credit for their research and how sharing their research data openly can help improve the trust around their research papers and result in more citations to the paper.

Motivations for data sharing by publication year

Year of first paper publication

Year									
1980	24%	11%	8%	16%	14%	8%	5%	8%	
1981	24%	15%	6%	9%	6%	6%	6%	15%	
1982	28%	5%	15%	10%	5%	8%	5%	13%	
1983	14%	9%	9%	6%	9%	17%	9%	14%	9%
1984	26%	14%	6%	11%	11%	11%	6%	6%	
1985	20%	6%	8%	14%	6%	6%	6%	6%	6%	14%
1986	20%	20%	10%	12%	7%	7%	7%	7%	
1987	25%	10%	12%	6%	6%	12%	6%	6%	8%
1988	23%	11%	6%	8%	11%	6%	16%	10%	
1989	25%	12%	12%	5%	8%	8%	5%	7%	10%
1990	20%	9%	13%	9%	5%	7%	9%	11%	
1991	19%	21%	12%	5%	7%	5%	12%		
1992	18%	16%	5%	11%	5%	16%			
1993	17%	12%	5%	8%	10%	7%	5%	8%	5%	12%
1994	17%	16%	6%	6%	6%	6%	6%	17%	
1995	20%	11%	12%	6%	13%	16%			
1996	24%	17%	7%	6%	19%				
1997	23%	14%	7%	6%	5%	11%	6%	9%	
1998	16%	20%	13%	7%	10%	9%			
1999	17%	21%	8%	5%	15%	5%	11%		
2000	17%	13%	15%	6%	15%	18%			
2001	18%	15%	6%	8%	10%	10%	13%		
2002	19%	16%	6%	11%	6%	6%	6%	14%	
2003	25%	8%	6%	11%	14%	10%			
2004	20%	11%	10%	13%	6%	15%			
2005	20%	11%	10%	11%	7%	11%	14%		
2006	25%	9%	5%	8%	7%	11%	8%	5%	8%
2007	27%	9%	6%	11%	11%	16%			
2008	20%	11%	10%	7%	16%	6%	8%		
2009	18%	6%	6%	16%	14%	6%	14%		
2010	16%	16%	8%	12%	11%	13%			
2011	19%	8%	9%	10%	12%	8%	13%		
2012	18%	12%	6%	8%	6%	5%	15%	8%	
2013	17%	7%	5%	12%	5%	7%	10%	6%	15%
2014	18%	10%	8%	12%	6%	16%	12%		
2015	23%	9%	5%	13%	6%	11%	12%		
2016	19%	14%	5%	7%	13%	6%	13%	9%	
2017	22%	8%	9%	11%	10%	7%	10%		
2018	20%	10%	9%	14%	11%	7%	6%		
2019	23%	8%	7%	11%	13%	6%	12%		
2020	19%	6%	7%	13%	11%	6%	15%		
2021	23%	7%	7%	12%	14%	12%			
2022	21%	6%	9%	11%	13%	6%	12%		
2023	16%	9%	8%	11%	11%	17%			

Percentage of respondents

Legend:
- Citation of my research papers
- Co-authorship on papers
- Consideration in job reviews and funding applications
- Direct request from another researcher
- Financial reward
- Freedom of information request
- Full data citation
- Funder requirement
- Greater transparency and reuse
- Increased impact and visibility of my research
- Institution/Organization requirement
- It was a field/industry expectation
- It was made simple and easy to do
- Journal/Publisher requirement
- None
- Open data badge
- Other (please specify)
- Public benefit

Graph showing the percentage of respondents that would be motivated to openly share their data by certain circumstances, broken down by the year of the respondent's first paper publication.

When discussing the education of researchers regarding the benefits of sharing research data, it's noteworthy that, although the desire for impact and career progression is consistent and support for open practices is strong, potential differences may exist in the concerns preventing researchers from making data available.

Percentage of researchers who list the cost of data sharing as something that puts them off data sharing

Job title	Percentage of respondents
Undergraduate student	34%
Principal Investigator	33%
Technician / Research assistant	31%
Laboratory Director/Head	28%
Research director / VP of research	26%
Professor	25%
Other (please specify)	24%
Associate professor	24%
PhD/Master's student	23%
Research scientist	23%
Physician/Clinician	23%
Healthcare professional	23%
Assistant professor	22%
Postdoctoral candidate	19%
Retired	15%

Graph showing the percentage of respondents who would be deterred from sharing their data openly by the associated costs, broken down by job title.

A prime example involves researchers hindered from sharing their data openly due to cost-related concerns. There's a slight skew towards undergraduate students and Principal Investigators (PIs). As budget holders, PIs likely harbour concerns about how this will impact the day-to-day management of their projects. Conversely, undergraduate students are more likely to worry about how it will affect their personal finances. Numerous avenues allow researchers to publish data at no direct cost to them, or to apply for support through their funding bodies. Just over a third of respondents (34%) indicated that their institution or organization would meet the costs of making research data openly accessible. This was followed by respondents' own funds (30%) and then 'Funds identified in your grant for this purpose' (20%). Associate professors and professors were significantly more likely to meet this cost using their own funds (34% each). Research scientists (42%) and research assistants/technicians (52%) were significantly more likely to have this cost met by their institution/organization.

These results are encouraging, showcasing the academic community's willingness to embrace open academic data. Unlike country- and subject-based approaches, we can ensure researchers at all stages comprehend the benefits of making data openly available while tailoring support to alleviate their concerns. In terms of actions, these data highlight a need for more inclusive outreach when organizing discussions, forums and panels in the open research space.

08
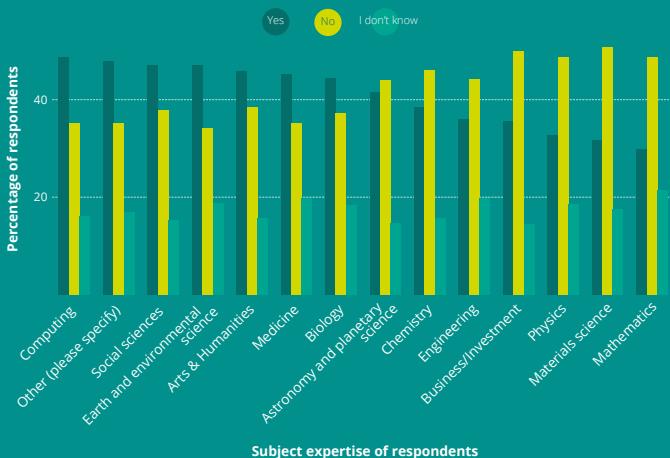
The importance of data management plans (DMPs)

Only 43% of respondents were aware of what a data management plan (DMP) is. Surprisingly, this represents a dip from the previous year, where over half (52%) claimed awareness. The awareness metrics further vary within specific demographics: those who recently published a manuscript were more familiar (44%) with DMPs, and those within the government or local government sectors showcased even higher awareness at 57%. However, certain disciplines like engineering, mathematics, material science, and physics lagged behind in awareness compared to other fields. The graph to the right represents respondents' awareness of the concept of a data management plan. The data is segmented based on respondents' primary area of expertise, highlighting differences in awareness across various fields.

While awareness is the first step, competence in drafting a functional DMP is another challenge. A mere 22% felt fully equipped to develop a DMP, while a significant 44% believed they would need moderate to extensive training on the subject. Interestingly, geographical location also played a role. Respondents from the US exuded greater confidence, with 31% feeling fully competent, whereas the figures were lower for Japan and Italy, with 8% and 9% respectively. Of those who know about DMPs, an encouraging 74% had created one. However, a notable subset of this demographic (3%) admitted to needing further training, underscoring the gap between awareness, actual implementation, and comprehensive understanding.
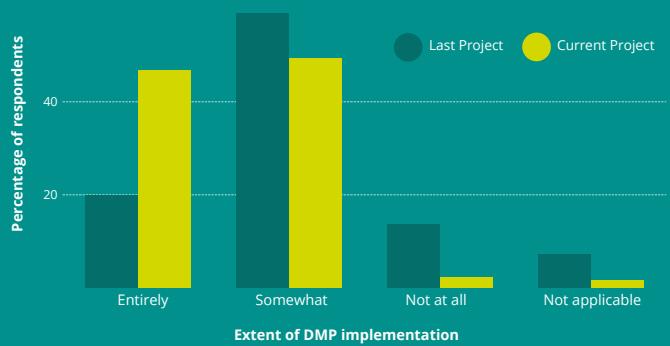
Encouragingly, there is a significant increase in the number of respondents who are fully implementing a data management plan in their current project compared to their last project. An impressive 79% created a DMP for their latest projects, with 96% doing so for ongoing projects. The percentage of those feeling they've fully implemented a DMP in ongoing projects (47%) is much greater than those from completed projects (20%).

Are you familiar with the concept of a data management plan?
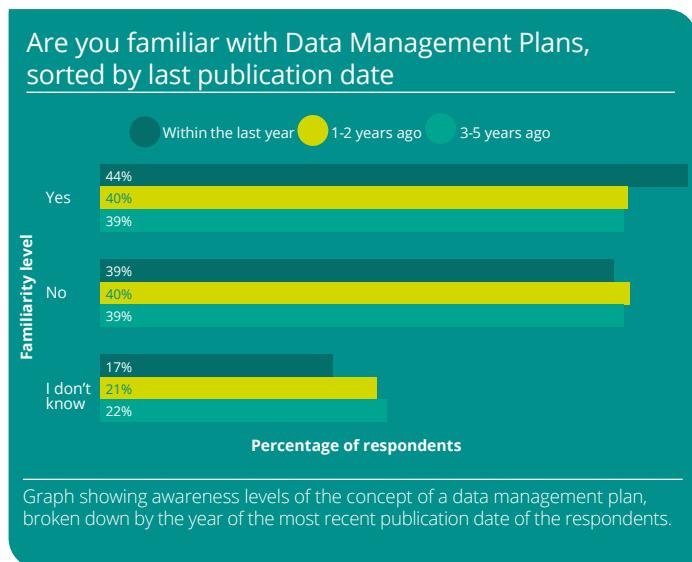
Graph showing the percentage of respondents that are aware of the concept of a data management plan, broken down by subject area of expertise of the respondents.

To what extent have you implemented a data management plan

Graph showing the extent to which respondents implemented a data management plan in their last project compared with their current project.
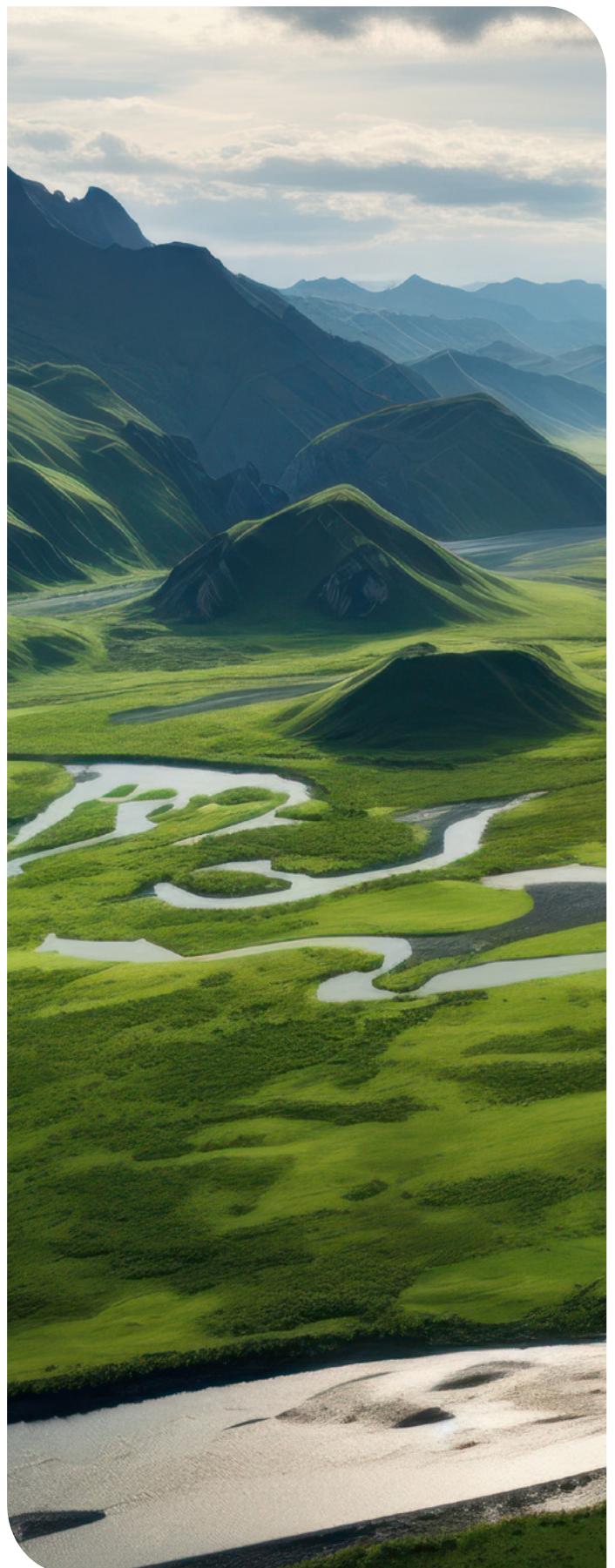
Diving deeper into the motivations, it emerged that a considerable number were influenced by external factors rather than personal convictions. Requirements from funders (42%) and institutions (36%) ranked high, while only 29% created a DMP out of personal choice. Other driving factors included expectations within research fields, recommendations from supervisors, and requirements from specific journals. For those making their debut in research publication in 2022 and 2023, the journal requirement seemed more pronounced, hinting at a possible trend towards mandatory DMP submissions in the future.

Are you familiar with Data Management Plans, sorted by last publication date

Within the last year   1-2 years ago   3-5 years ago

Familiarity level

Yes
- 44%
- 40%
- 39%

No
- 39%
- 40%
- 39%

I don't know
- 17%
- 21%
- 22%

Percentage of respondents

Graph showing awareness levels of the concept of a data management plan, broken down by the year of the most recent publication date of the respondents.

Challenges

While the progress in the DMP space seems promising, issues remain. Technical challenges topped the list with 34% struggling with data storage and organization. Time constraints and the lack of trained staff followed closely. Financial challenges, changing research dynamics, and memory lapses in following the DMP also emerged as barriers. Medicine and material sciences faced unique challenges, emphasising the need for a tailored approach to data management across different disciplines.

In conclusion, while the awareness and implementation of data management plans are gaining traction, there exists a gap in confidence and competence. With the myriad challenges faced in its application, there's an evident need for structured training, tailored guidance, and perhaps a rethinking of how DMPs are designed and executed across various disciplines.

■ Niki Scaplehorn, Director,
Content Innovation,
Springer Nature

■ Henning Schoenenberger,
Vice President Content
Innovation, Springer Nature

09

AI and open science - the start of a beautiful relationship?

With the incredible pace of recent developments in AI, researchers are rapidly confronting the difficult question of how to safely use these tools to accelerate their everyday work. However, according to this year's State of Open Data survey, only a small portion of researchers have so far considered using these tools to collect, analyse and annotate their data. In this section, we discuss the potential for AI to overcome barriers to open science, and highlight how open science will become ever more important in an AI-augmented world.

Unleashing the power of small data

AI techniques have been improving our article searches for longer than many would realize, but the ability of the latest generation of large language models (LLMs) to extract meaning from large volumes of text has already created a raft of new tools for scientists to help navigate and make sense of the scientific literature. These tools are among the first to gain the widespread attention of researchers exploring the possibilities of using generative AI. Rather than relying on keywords to identify papers of interest, LLM-based products can identify papers based on how closely they match a particular research question, and quickly extract and summarize details across a large number of papers to present a balanced answer, backed up by references. This approach has the potential to save hours or even weeks of research time, although it comes at the expense of a level of transparency and control over which papers are highlighted, and ultimately a risk of systematic biases that, for now, can only be mitigated by careful comparison of augmented and more manual approaches.

While these tools primarily search and process text, it is already possible for large 'multimodal' models to make sense of images and other data types. For open science, this raises the exciting prospect of being able to unlock the vast array of data that are hidden in previously machine-unreadable formats - within figures and tables or in supplementary files,

for example. While open science often focuses its efforts on big, highly re-usable data, AI is poised to unleash and make sense out of a tidal wave of small data that is otherwise rarely re-used.

Making meta better

While sharing data can be easy, thanks in no small part to generalist repositories such as Figshare, sharing data in full accordance with the principles of findability, accessibility, interoperability and re-usability (FAIR) can be considerably more challenging. Researchers often share data without including the vital metadata required to understand how it was created and how it should be re-used. Very often, that information is buried within an associated research article, but inconsistencies in how data and articles are linked, and the complexity of the article itself, can make it prohibitively difficult for others to confidently re-use or reproduce the data.

With most science funders now encouraging and even mandating the sharing of all underlying data at the point of publication, journals are playing an increasingly proactive role in promoting good open science practices, and LLMs have the potential to support both authors and journals through that process. Although human oversight will always be necessary to detect potential errors, AI-based drafting of metadata, data availability statements and even associated data descriptor papers will be valuable tools to improve standards of data sharing.

Your personal data scientist

LLMs are not limited to searching and summarizing - they are already set to transform how we analyse and visualize data. For many of us, data analysis begins with a spreadsheet and ends with copying and pasting a graph into a figure, leaving the data behind to languish on our hard drives. Analysing and visualizing data using code

is usually considered to be a more challenging option, reserved for more complex analyses and those with sufficient technical expertise.

The ability of LLMs to write code is, however, dramatically changing this. It's now possible to use an LLM as a personal data scientist, which can explore a complex dataset, propose interesting research questions, design a step-by-step statistical analysis, and draw up graphical representations of the results in a matter of minutes. Importantly, the data and the resulting figure remain linked together by code, making the figure a dynamic artefact that can be reanalysed and redrawn, and making the raw data an integral part of the figure. These tools are still far from perfect: it's still essential to make sure that an analysis driven by AI makes scientific and statistical sense. But for open science, AI raises the tantalizing prospect of data finally taking its place at the heart of our scientific articles, with integrated transparency and reproducibility as to how the data were processed and analysed.

A virtual collaborator

For most of us, until recently, the idea of an artificial scientist designing and even carrying out their own research projects felt very much like a prospect for the distant future. But just as LLMs can support human data analysis, the prospect of autonomous agents that can sift through the world's scientific data and identify unanticipated patterns and connections is no longer far away. For example, AI could be used as a co-pilot within a research team - not only scanning the literature for new papers of interest but running analyses that integrate the team's most recent results with publicly available data to propose new experiments and new collaborations.

The impact of these new players on the data ecosystem may be complex. On the one hand, the use of AI will greatly increase the chances that shared data will be re-used, creating new demand for open science. At the same time, fear of data 'misuse' is still a significant barrier to data sharing, and if mechanisms to ensure proper credit for data creators and good practices around authorship and collaboration are not in place, the threat of data being snapped up by a competitor bot could act as a disincentive to early sharing of data. These are not new dynamics, but their impact could be magnified and accelerated by the widespread use of AI.

Open science to the rescue

A clear risk of generative AI for research publishing more generally is the potential for paper mills to create fake articles much more quickly and easily than was previously possible. There are many beneficial uses of these technologies in scientific writing, especially as an opportunity to level the playing field and broaden access to publication, and so even if the detection of AI-generated text was completely reliable, banning AI-generated text from scientific publications would not represent a viable solution to this problem.

While addressing the misaligned incentive structures that lead researchers to use paper mills in the first place should be our ultimate goal, open science has an important part to play in providing proof of authenticity in research. Insisting on the sharing of all raw data makes it much more difficult to construct an entirely falsified article, especially with the development of increasingly sophisticated AI-based tools to detect data manipulation. The prospect of an AI 'arms race' between data falsifiers and manipulation detectors is a grim vision of the future, but without open science, the struggle to protect the integrity of the scientific record will be lost long before that. Instead, we should work towards ensuring that our vision for the positive impact of AI on open science becomes a reality.

The prospect of an AI 'arms race' between data falsifiers and manipulation detectors is a grim vision of the future, but without open science, the struggle to protect the integrity of the scientific record will be lost long before that. Instead, we should work towards ensuring that our vision for the positive impact of AI on open science becomes a reality.

10

Recommendations for the academic community

Understanding the 'state of open data' in our specific research setting

Respondents from different geographies and different disciplines responded to the survey with great variation. A purely global and cross-disciplinary approach to research data management and its promotion is clearly not sufficient or sustainable. Understanding the 'state of open data' in our specific research setting, whether that be at a national level, a university level or a departmental level could be the key to truly engaging with the communities we're trying to reach. There is a clear need to tailor support that's dependent on awareness and support levels specific to the particular context that we're operating within.

Action Request:
- **For policymakers**: Engage with researchers and institutions to understand specific needs and challenges within their context and tailor open data policies accordingly. Establish advisory committees, consult with experts and collaborate with scientific organizations.

- **For researchers**: Actively participate in surveys, forums, and discussions to voice the specific challenges and needs related to open data within your specific setting.

- **For academic institutions**: Conduct internal surveys and workshops to understand the unique needs of different departments and establish data management support that is contextually relevant.

- **For publishers**: Facilitate collaboration between researchers and academic institutions by organizing workshops or conferences that bring stakeholders together to share experiences and knowledge.

Credit where credit's due

The issue of credit, and researchers feeling they don't receive enough of it for sharing their data openly, has consistently been evident in The State of Open Data since its inception. Credit for data sharing could take many forms, from a citation to career-related recognition or progression. Community initiatives like CoARA (Coalition for Advancing Research Assessment) are tackling this head-on. This initiative launched by the EU Commission, the European University Association (EUA), Science Europe, and with over 350 signatories, has principles guiding this space. We need to give more rewards and incentives for data sharing when assessing research. One way in which receiving credit for sharing data openly could be more likely is if it's always linked to the final publication as clearly as possible. Some publishers are piloting innovative solutions such as the Public Library of Science (PLOS) and the pilot of their 'accessible data button'. This was the seemingly simple addition of a button linking to an open dataset associated with a PLOS article in selected repositories; the pilot saw a 20% relative increase in views of the datasets included. More eyes on a dataset could mean more potential for re-use, a citation or simply further recognition. Both publishers and repositories could begin thinking about the way they link datasets to articles and work to increase visibility, therefore creating opportunities for more recognition.

Action Request:
- **For all participants in the space**: Define an inclusive and all-stakeholder approach. Societal change requires looking outside of not-for-profits and into industry and beyond in order to effect change.

- **For funders and academic institutions**: Establish clear policies that recognize and reward researchers for sharing data openly and integrate these contributions into career progression evaluations.

- **For publishers**: Adopt and implement mechanisms, like the 'accessible data button', that enhance the visibility of open datasets linked to publications and explore additional methodologies to credit researchers for sharing their data.

- **For researchers**: Ensure to cite datasets appropriately in your research and advocate for policies within your institutions that recognize and reward open data sharing.

Help and guidance for the greater good

We need inter- and intra-community coordination when it comes to doing a better job at education. The NIH-funded GREI initiative is a prime example of a funder forcing change by preventing the siloed approaches of individual generalist data repositories. This prevents organizations from focusing on the functionality of their own platforms, with training or materials specific to their infrastructure or repository, but instead empowers them to offer more general guidance documentation and advice on data sharing that goes beyond a user choosing to use their specific product or platform. Being more collaborative and contributing to the bigger picture could be of greater benefit to researchers. While the awareness and implementation of data management plans are gaining traction, there exists a gap in confidence and competence. With the myriad challenges faced in its application, there's an evident need for structured training, tailored guidance, and perhaps a rethinking of how DMPs are designed and executed across various disciplines.

Action Request:
- **For publishers, librarians and software providers**: Develop and disseminate open data guides that are not product-specific and conduct workshops and webinars that cater to a broad audience regarding the essentials and best practices of data sharing.

- **For researchers**: Advocate for, and participate in, forums and workshops on data sharing and bring forward your challenges and insights to help shape better platforms and policies.

- **For funders**: Ensure that funded projects allocate resources and time for data management, and provide clearer guidelines on data sharing requirements.

Making outreach inclusive

As we've discussed within our report, early career researchers are not the only ones who support and also struggle with data sharing. Researchers and academics at all stages of their careers share the same motivations, have the same concerns and exhibit similar levels of support when it comes to open data. In an organizational setting, it may be tempting to focus on instilling and promoting core open science values among early career researchers and those who are just starting their academic journeys. One takeaway from this year's results is that those looking to engage research communities should be inclusive and deliberate with their outreach, engaging those who have not yet published their first paper as well as those who first published over 30 years ago.

Action Request:
- **For conference organizers**: Ensure representation from various career stages in discussions, panels, and talks related to open data and science.

- **Implementation of mentorship programmes with a specific focus on data sharing**, aimed at fostering collaborative learning from one another.

- **Outreach in all areas**: Have a standard set of support tools, accessible via a central repository - similar to the OA Books Toolkit for Researchers.

We would also like to make a final call to ALL actors to actively and above all regularly participate. This can take various forms, including participation in forums where experiences, needs, and requirements are openly discussed. The goal of these forums is to establish clear objectives that aim to standardize the entire data-sharing process.

Moving forward, it's essential to maintain a dynamic approach. This involves regularly monitoring, evaluating, adapting, and sharing feedback on the standardized processes. We don't see standardization as a fixed achievement but as an evolving process that needs to keep up with technological advancements and changing methodologies.

In the upcoming forums, the focus should be on discussing progress and sharing constructive feedback. This ongoing feedback loop ensures that our data-sharing standards remain relevant and aligned with the latest developments in the field, helping us stay current and effective.

11

Authors and acknowledgments

Graham Smith

Graham is the Open Data Programme Manager at Springer Nature. He works to develop and promote data-sharing tools, partnerships and initiatives across the organization's publishing activities. He has a background in geophysics and has coordinated data curation activities across the Nature, BMC and Springer portfolios, and at the Natural History Museum in London.
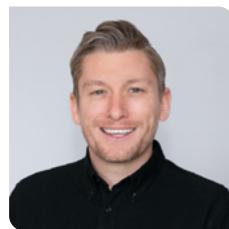
Henning Schoenenberger

Henning, Vice President Content Innovation at Springer Nature, is an accomplished leader in digital innovation and a pioneer in the early adoption of artificial intelligence in scholarly publishing. He ideated and product managed the first machine-generated research book published at Springer Nature. Henning graduated in Social Science and is based in Heidelberg, Germany.

Laura Day

Laura is the Marketing Director at Figshare, part of Digital Science. Before joining Digital Science, Laura worked in scholarly publishing, focusing on open access journal marketing and transformative agreements. In her current role, Laura focuses on marketing campaigns and outreach for Figshare. She is passionate about open science and is excited by the potential it has to advance knowledge sharing by enabling academic research communities to reach new and diverse audiences.

Dr Mark Hahnel

Mark is the CEO and founder of Figshare, which he created while 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.

Niki Scaplehorn

Niki is Editorial Director, Content Innovation, at Springer Nature, where he is developing solutions that make it easier for authors to share and interact with data, code and protocols. A cell biologist and neuroscientist by training, he was drawn to an editorial career at Cell before moving to Springer Nature, where he became Chief Life Science Editor of Nature Communications and Editorial Director for the Nature life science journals. He is based in Hamburg, Germany.

Contributors

Thanks to colleagues from Springer Nature, Figshare and Digital Science who also helped to shape this whitepaper.

Acknowledgments

Figshare, Digital Science and Springer Nature extend their thanks to Shift Insight, who collected the 2023 survey data and undertook the initial analysis, as well as our all-important respondents who provided us with unparalleled insights and enabled us to bring the survey findings to the wider research community.

A collaboration between Figshare, Digital Science, and Springer Nature