

# Bridging policy and practice in **data sharing**

An investigation into what is driving successful data sharing in repositories

Mark Hahnel, Digital Science, Graham Smith, Springer Nature, Ann Campbell, Digital Science





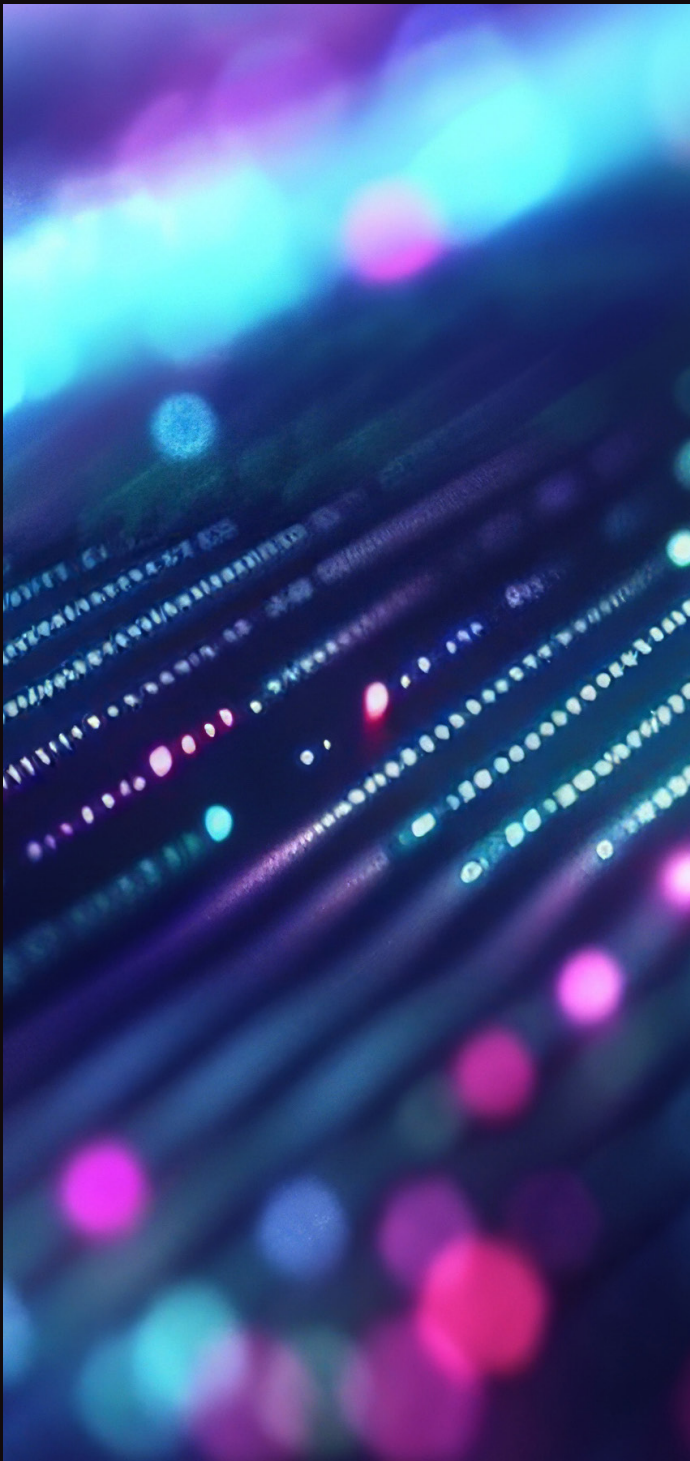
# Our aims

Research data forms the foundation of research and provides the evidence behind published articles. The more we share data openly, the more it will help create a more equitable, fairer, and less wasteful research ecosystem.

The State of Open Data survey continues to provide a detailed and sustained insight into the motivations, challenges, perceptions, and behaviors of researchers towards open data. Now in its ninth year, the survey is a collaboration between Figshare, Digital Science and Springer Nature. Uniquely this year we wanted to go beyond understanding the thoughts and attitudes of researchers and, for the first time, look into what they were actually doing.

By combining three different data sources (Dimensions, Springer Nature Data Availability Statements (DAS) and the Make Data Count and DataCite Citation Corpus, from now on referred to as MDC DCC, we reveal linkages between peer reviewed published research and datasets being made available. We believe this jump from understanding what people say they are going to do to actively showing what they are doing, is an important step in driving change and understanding how to bridge the gap between policy and practice in open data sharing. The report identifies trends and builds our understanding at a country, institution, and funder level of what is actually happening in order to work out what works and learn from different approaches.

It's only by collaborating with other actors in the research ecosystem and publishers, funders, repositories, and government agencies working together, that we can tailor our interventions and drive real change in a more equitable way. This report aims to shed light on global trends, regional differences, and the evolving landscape of data sharing practices. Understanding these patterns is crucial for developing strategies that encourage open data practices, ultimately enhancing the reproducibility, efficiency, and integrity of scientific research.



*Data publishing is having a coming of age moment. The State of Open Data has been essential in tracking how researchers globally feel about the changes in the ways they disseminate their research. The report suggests we are at the point of finding ways to reward researchers for their data, this too is essential and we are excited to be involved in playing a role as Digital Science. We're delighted to partner with Springer Nature once more and look forward to seeing the ongoing success and impact of the reports."*

Daniel Hook, CEO  
Digital Science



*We are incredibly proud of the State of Open Data report, and our partnerships with Figshare and Digital Science. Each year the report provides valuable insights for publishers, institutions, and funders about the roles we need to play in better supporting researchers with open research practice.*

*With a focus for 2024 on what actions researchers are taking, we have even more qualitative evidence as to what we, as a community, need to do to help drive forward a more reproducible research ecosystem – an ecosystem that is critical for open science and accelerating solutions to the world's most urgent challenges."*

Harsh Jegadeesan, Chief Publishing Officer  
Springer Nature

## Contents

<b>02</b>	<b>Our aims</b>	
<b>04</b>	<b>Introduction</b>	
<b>06</b>	<b>Methodology</b>	
	Data linkages in the Make Data Count Data Citation Corpus	06
	Data linkages in the Dimensions Data Citation Corpus	06
	Data linkages in Data Availability Statements	07
<b>08</b>	<b>Findings</b>	
	Background on key country trends	08
	Familiarity with FAIR principles	09
<b>11</b>	<b>New insights &amp; analysis</b>	
	Country specific focus	14
	Country analysis by dataset	16
	Trends at a funder level	22
	Trends at a university level	24
<b>27</b>	<b>Conclusions</b>	
<b>28</b>	<b>Recommendations and next steps</b>	



# Introduction

Data sharing is a fundamental pillar of open research, enhancing transparency, reproducibility, and collaborative advancement across diverse research fields

For nine years, [The State of Open Data](#) has focused on researchers’ attitudes towards and experiences of open data. Whilst it is very important to understand what is incentivizing researcher behaviors, what they say they are going to do is not always what they do. While we have still carried out the survey for 2024, we are excited to explore a new set of quantitative data as well as the qualitative survey results.



During the past decade, we have seen several initiatives that aim to standardize the way in which open data is presented in the scholarly output landscape. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse. In support of this assertion, and to encourage good practice, the [Joint Declaration of Data Citation Principles](#) was published by [FORCE11](#). More recently, the [S-Index Prize from the National Cancer Institute \(NCI\)](#) is looking to measure individual researchers based on their data sharing habits.


Several academic publishers have implemented mandates for sharing data associated with research articles. The implementation of these policies can vary by discipline, but major publishers like Springer Nature, Taylor & Francis, Wiley, and the Public Library of Science (PLOS) have been notable advocates. For example, Springer Nature’s approach not only mandates data availability for certain journals and

data types but also provides guidance and infrastructure to support researchers in sharing their data. One of the first publishers to require data sharing was PLOS, which introduced its data policy in 2014, mandating that all authors provide access to their data as a condition of publication. This set a precedent for other publishers to follow.

Taken together, this now means that there are ways to track whether researchers are making their data openly available and to what standard. In this report, we draw on several sources in order to investigate some of the patterns we see forming, with a view to better understand the real motivations for data sharing and ultimately effect change. As the scientometrics of data citation and linking is a nascent field, none of the sources used in this analysis are definitive. Instead, we find that exploring multiple sources allows us to have confidence in patterns that are emerging.

## Data Sources


In this analysis, we look at data from the following sources. While at this stage none are a single source of truth, they collectively help to identify trends and build an understanding of patterns in what is happening in practice:



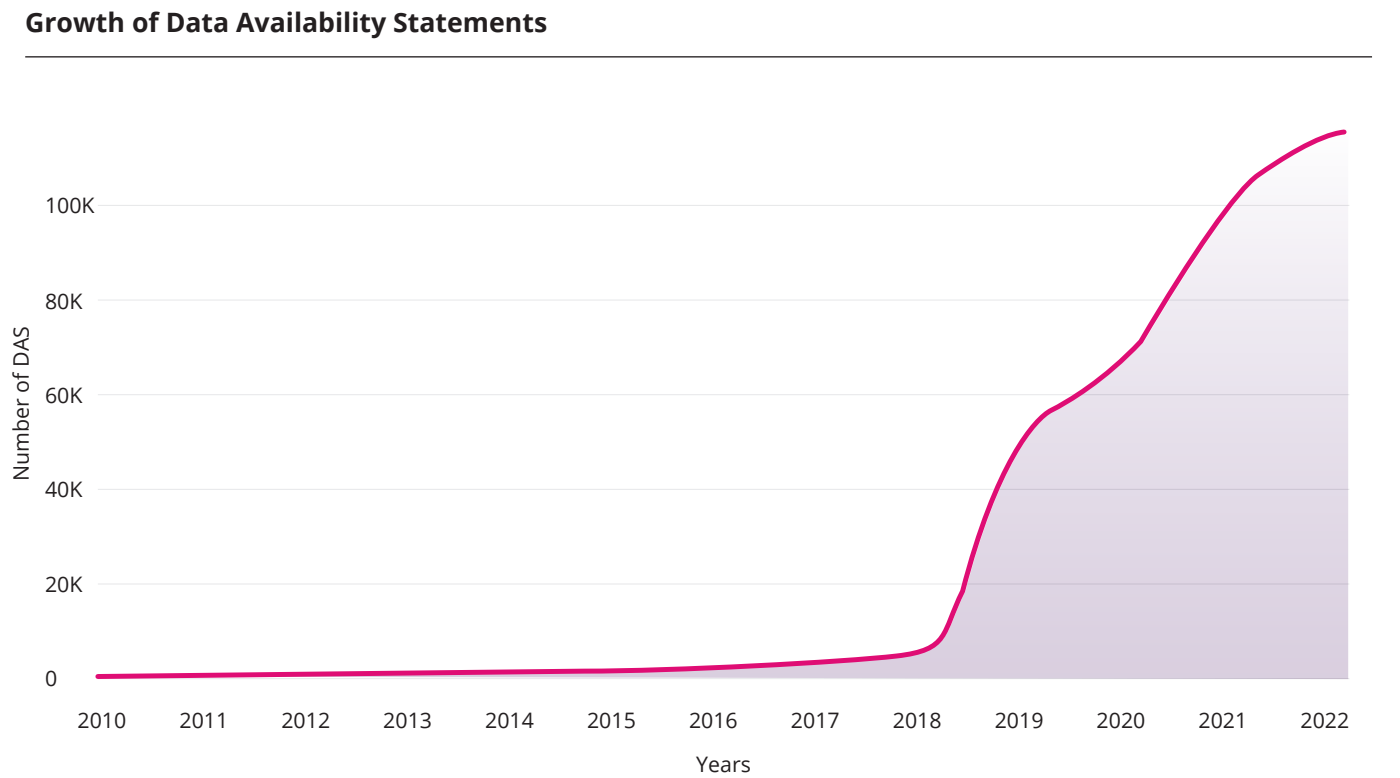
The [Make Data Count](#) and [DataCite Data Citation Corpus \(MDC DCC\)](#) is an initiative aimed at creating an open, comprehensive and centralized resource for data citations, which can dramatically improve how research data is tracked and cited. This corpus aggregates references to datasets from a wide range of sources, making it easier for researchers, funders, and institutions to monitor the impact and dissemination of research data. The primary goal is to provide open, publicly accessible data citations, addressing the long-standing challenge of evaluating the use of open data in research. The corpus includes over five million data citations from both DOI (Digital Object Identifier) and non-DOI sources, such as accession IDs, helping to ensure a broader scope of research data is represented.



[Dimensions](#) is a research and innovation insights platform that aggregates and connects a vast range of scholarly information sources to provide comprehensive data for research discovery, analysis, and impact tracking. Launched by [Digital Science](#), it includes content like publications, grants, patents, clinical trials, datasets, and policy documents. With full text access to over 100 million research articles, we can easily find linkages between academic research papers and DataCite DOIs. We can also interrogate Data Availability Statements in said research papers.



Springer Nature Data Availability Statements (DAS) - a DAS is a section of a research paper where authors specify where and how the underlying data supporting the findings can be accessed, or provide reasons for any data restrictions. It ensures transparency and supports reproducibility and compliance with open data policies. DAS became more common in the mid-2010s, driven by journal mandates, funder requirements, and open science initiatives like Plan S and the FAIR principles, promoting standardized data sharing across disciplines. This growth can be seen in the chart below.



# Methodology

This year we have combined data sources to analyze linkages between papers and datasets

## Data linkages in the MDC DCC

The Data Citation Corpus is a project by DataCite and Make Data Count funded by [Wellcome](#), which has as focus the development of a comprehensive, centralized, and publicly available resource of data citations from a variety of sources. The first release of the MDC DCC was delivered on January 30, 2024. The second release was shared on August 23, 2024. More details can be found [here](#).

### MDC DCC

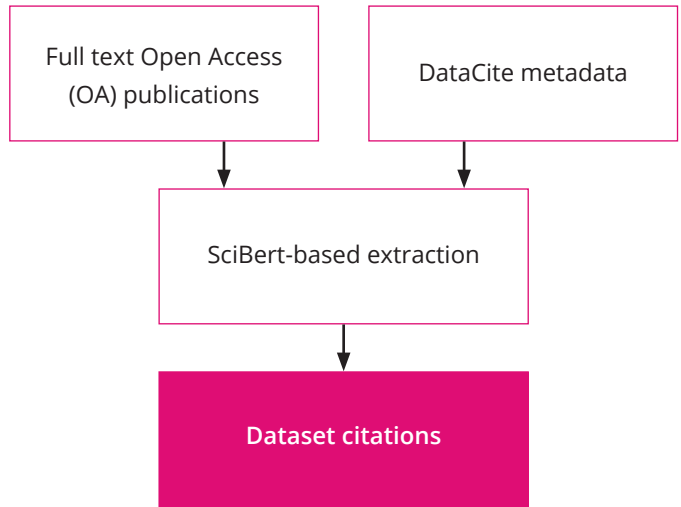


Figure 1. MDC DCC link (citation) extraction protocol.

Please note: whilst it provides citations to accession numbers, identifiers can be inconsistent and 6% self-citations of data or publications. Based on 1.9m publications cite data, 3.4m datasets are cited of which 1.4m have a DOI and 1m are in Dimensions. 413k dataset IDs are multi-cited

## Data linkages in the Dimensions Data Citation Corpus

Dimensions is a powerful, AI-enhanced research platform by Digital Science, designed to provide an interconnected view of the global research landscape. It aggregates vast data from diverse sources, including over 140 million publications, 7 million grants, and additional files and metadata such as datasets, clinical trials, patents, and policy documents. This database allows us to search for DataCite DOIs in the full text of the articles. The resultant database does not aggregate links to accession numbers like the MDC DCC.

### Dimensions

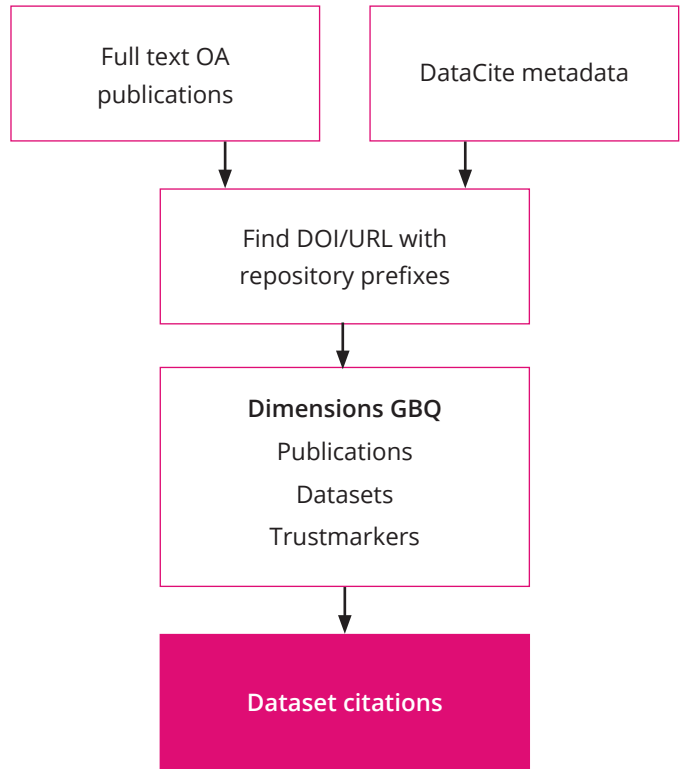


Figure 2. Dimensions Data Citation Corpus Data link (citation) extraction protocol.

This provides good, clean metadata and identifiers and extracts from full text also for non-open access. However extraction of DOIs is difficult in PDFs (improvements expected in the future) and there are no citations of accession numbers. 1.4m of 91.2m (1.5%) publications cite data and 3m out of 32m (9%) datasets are cited. 90k datasets are multi-cited by 0.25m publications.

## Data linkages in Data Availability Statements (Springer Nature Journals, 2019–2022)

Traditional bibliometric measures, such as data citations, often underrepresent the true extent of data sharing practices. This creates challenges in accurately assessing how data is disseminated and reused within the scholarly community. To bridge this gap, analyzing Data Availability Statements (DAS) provides a more nuanced understanding of data sharing behaviors. DAS offer direct insights into researchers’ intentions and practices regarding data accessibility, moving beyond what citation metrics can capture. By examining DAS, we can uncover patterns and trends that inform policies and support mechanisms aimed at promoting open science.

By conducting an extensive analysis on DAS, this study investigates data sharing patterns in Springer Nature journals from 2019 to 2022. Utilizing the Dimensions database, we identified articles containing key DAS identifiers such as “Data Availability Statement” or “Availability of Data and Materials” within their full text. Digital Object Identifiers (DOIs) of these articles were collected and matched against Springer Nature’s XML database to extract the DAS for each article.

The extracted DAS were categorized into specific sharing types using text and data matching terms. For statements indicating that data are publicly available in a repository, we matched against a predefined list of repository identifiers, names, and URLs. The DAS were classified into the following categories:

1. Data are available from the author on request.
2. Data are included in the manuscript or its supplementary material.
3. Some or all of the data are publicly available, for example in a repository.
4. Figure source data are included with the manuscript.
5. Data availability is not applicable.
6. Data are declared as not available by the author.
7. Data available online but not in a repository.

These categories are non-exclusive: more than one can apply to any one article. Publications outside the 2019–2023 range and non-article publication types (e.g., book chapters) that were initially included in the Dimensions search results were excluded from the final dataset. Articles were included in the final analysis after applying the exclusion criteria. Upon processing, it was found that only 370 results were returned for Botswana across the five-year period; due to this low number, Botswana was not included in the DAS focused country-level analysis.

This analysis is naturally limited to Springer Nature publications. It does not assess the accuracy of the DAS in the context of each individual article. There was no manual verification of the categories applied; as a result, terms used out of context could have led to misclassification. Approximately 5% of articles remained unclassified following text and data matching due to these limitations.



“DAS offer direct insights into researchers’ intentions and practices regarding data accessibility.”

Graham Smith, Open Data Programme Manager  
Springer Nature



# Findings



“Everyone’s on board with FAIR principles but different regions and funders are executing it in different ways and some are having more success than others. So if we can find out what is successful, then we can advance that for everybody.”

Mark Hahnel,  
VP Open Research, Digital Science



## Trends at a country level

### Background

Earlier in 2024, in an effort to emphasize the depth of the data that is made openly available by The State of Open Data survey, we released a re-analysis undertaken by a team of undergraduates from King’s College London: [The Global Lens: Highlighting national nuances in researchers attitudes to open data](#). As the title alludes to, this report took a deep dive into the differences in responses from researchers based in different countries, to demonstrate that global trends don’t always align with national trends.

The following figures, taken from that report, highlight country-based differences in enthusiasm for open data, knowledge of the [FAIR Data Principles](#), and motivating factors for making data open.

### Should funders withhold funding from (or penalize in other ways) researchers who do not share their data if the funder has mandated that they do so at the grant application stage?

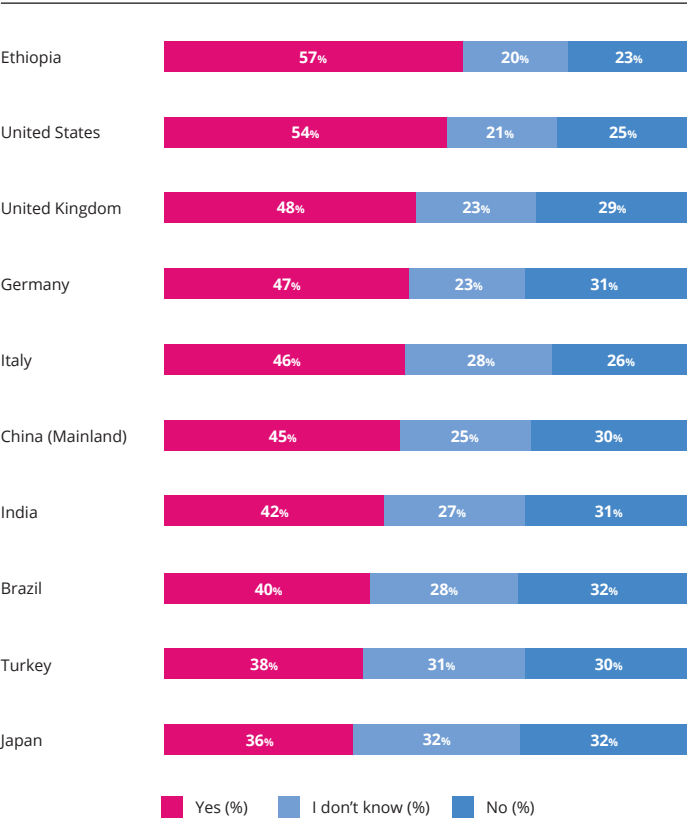


Figure 3. Percentage of responses to survey question “Should funders withhold funding from (or penalize in other ways) researchers who do not share their data if the funder has mandated that they do so at the grant application stage?” (State of Open Data 2023)

## How familiar are you with the FAIR data principles in relation to open data?

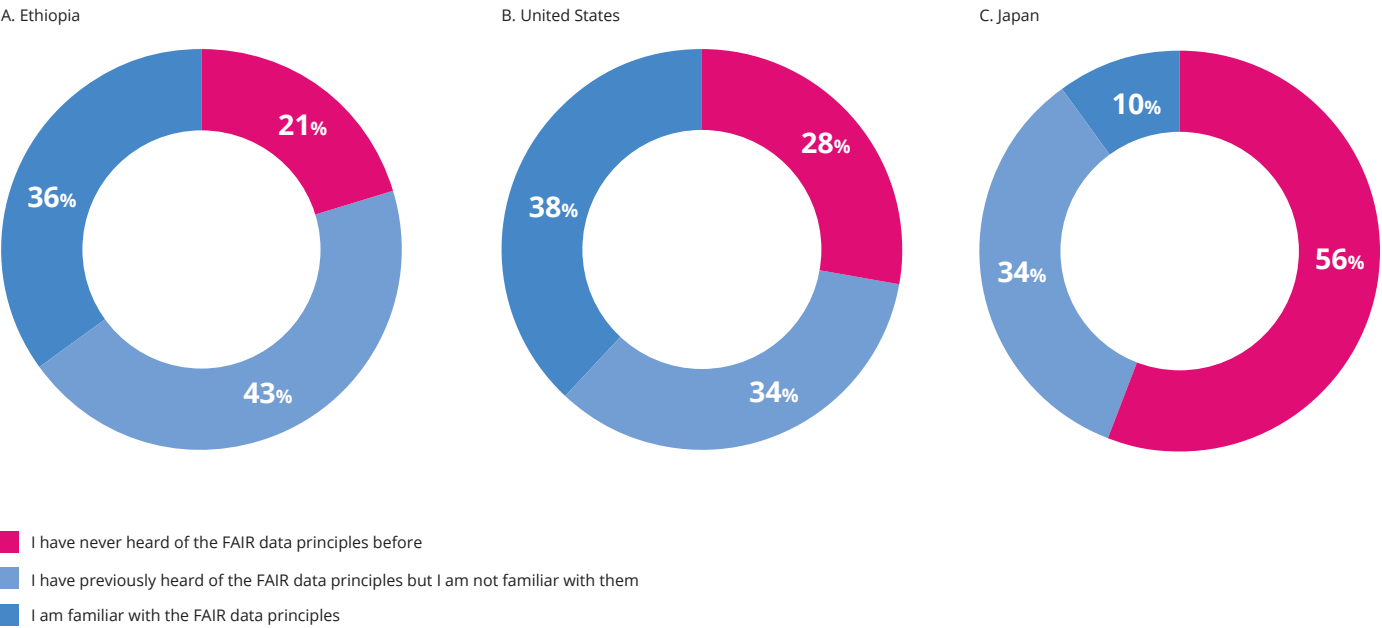
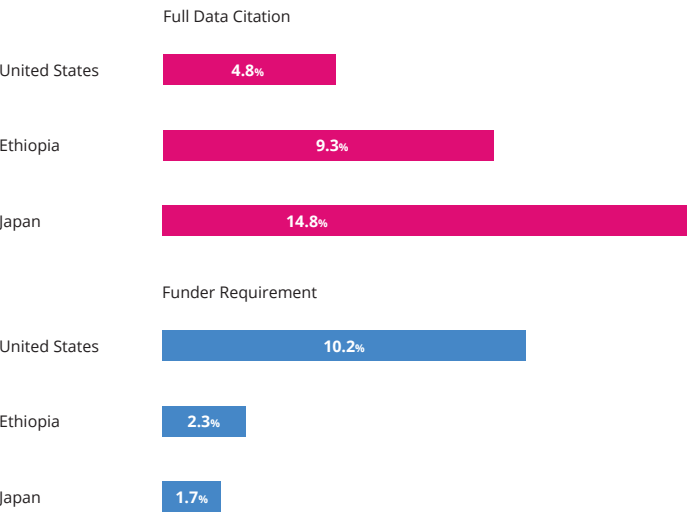


Figure 4. Percentage of responses to survey question ‘How familiar are you with the FAIR data principles in relation to Open Data?’ for A. Ethiopia, B. United States and C. Japan (State of Open Data, 2023)

The pie charts above show researchers’ familiarity with the FAIR data principles. These principles are often included and form the basis of funder and government policies regarding open data. The FAIR principles are a set of guidelines for making data easier to share and reuse, by ensuring that the data is: Findable, Accessible, Interoperable and Reusable (FAIR).

The 2023 State of Open Data survey responses shown in Figure 4 highlight that Ethiopia and the United States have a similar proportion of researchers that are familiar with the principles (36% and 38% respectively), while the majority of respondents from Japan had never heard of the FAIR data principles (56%).



The United States has the lowest percentage of researchers that are motivated by citation of their data (4.8%) while having the highest percentage motivated by funder requirement (10.2%). Conversely, Ethiopia and Japan show similarity with a higher importance of motivation through citation of their data (9.3% and 14.8% respectively) and a similar low importance of motivation from funder requirement at (2.3% and 1.7% respectively).

Figure 5. Percentage of responses to survey question “Which one of these circumstances would motivate you the most to share your data?” for Ethiopia, United States, Japan (State of Open Data 2023).



Data related retractions by country in 2023

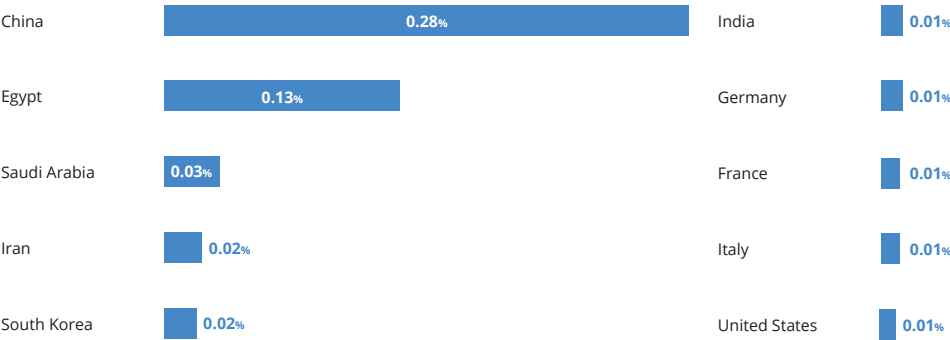


Figure 6. Retractions for data related reasons (concerns/issues about data, duplication of data, error in data, plagiarism of data, unreliable data) in the Retraction Watch Database in 2023 as a percentage of papers published, by country

Unlocking that there are differences in the motivations at a country level has helped us identify where training may be more important to advance open data principles. However, this is just indicative of how researchers are behaving. This report aims to pair a new quantitative analysis to the qualitative data of the State of Open Data survey over the last nine years. The analysis in the following pages highlights how similar patterns are emerging for countries between how enthusiastic and knowledgeable researchers are about open research, and how good their open data practices are.

While the hypothesis that funders or institutions with stronger open access and open data policies will have better open data practices is evident, there seem to be fluctuations based on the type of research they fund and their geolocation. The country-based analysis emphasizes these differences with strong continental disparities. This could be due to cultural norms, such as a strong focus on IP in Asia. There are also continental differences in publication rates that highlight a difference in scale. Data-related retractions (above) and other research integrity measurements also shine a light on the difficulties in rolling out new policies at scale whilst maintaining trustworthy research practices.



# What's new?

## New insights and analysis

Different segments of the academic landscape - across countries, funders, and institutions - demonstrate varied success in open data practices, influenced by policy environments, funding conditions, and institutional resources. Recognizing and addressing these disparities through targeted support and harmonized policies can further accelerate the global adoption of open data practices. In doing so, we can foster a more equitable and accessible research ecosystem, where data sharing is valued, rewarded, and ultimately becomes a cornerstone of academic success. The success of open data initiatives often reflects the broader policy environment within a country. We hypothesize that countries with proactive open science policies generally lead in open data compliance and engagement. Conversely, countries where data policies are less defined, or where infrastructural limitations exist, often experience slower progress in adopting open data practices.

Funders play a pivotal role in determining the level of open data compliance, as they set the terms for data sharing and are often the primary source of mandates. Public funders, especially in the United States and Europe, frequently require open data as part of their grant conditions. In contrast, private funders may vary in their approaches, depending on organizational goals and focus areas. Some private foundations have invested directly in open data infrastructure, while others may not prioritize open data as explicitly. This variability creates a mixed landscape in which researchers funded by public grants may be more actively involved in data sharing than those relying on private sources.

Institutions play a crucial role in supporting or hindering open data practices, as they provide the resources, infrastructure, and professional recognition for data sharing. Institutions with robust data management services, such as dedicated repositories or support teams, tend to have higher compliance rates. These institutions are often located in regions or countries where open data is a policy priority, highlighting the interplay between national policies and institutional practices. Conversely, institutions with limited resources or insufficient digital infrastructure may struggle to support open data practices, even if their researchers are motivated to share data.



Identifying patterns in high-adoption regions allows us to pinpoint effective policies, resource allocation, and institutional practices. For example, countries with comprehensive open science policies may experience higher adoption rates, suggesting that clear national mandates and infrastructure support are critical incentives. Similarly, funders with strong data-sharing requirements, particularly public ones, often see greater compliance among their grantees, indicating that mandates backed by funding are effective motivators. This approach of “reverse engineering” successful strategies allows us to develop best practices and replicate them across various contexts.

By examining different rates of open data adoption, we gain insights into the incentives that drive success at the funder, country, or institutional level.



“*Targeted incentives that consider these diverse influences will help bridge gaps in adoption.*”

Mark Hahnel, VP Open Research  
Digital Science

Country	Position (MDC DCC)	Position (Dimensions)	Position (Dimensions DAS)	Average Position
Ethiopia	5	2	1	2.67
Botswana	1	5	3	3
United Kingdom	2	3	6	3.67
China	3	7	2	4
Germany	6	1	5	4
France	7	4	4	5
United States	4	6	7	5.67
Japan	8	8	8	8
India	9	9	9	9

Table 1. Average position by country

We utilized three distinct databases to evaluate the success of countries in publishing open data. While each database produced different rankings based on its unique criteria, we observed notable correlations between the rankings. This consistency suggests that, despite methodological differences, the relative success of countries in open data initiatives shows similar trends across multiple sources of evaluation. The datasets used for this ranking were the MDC DCC, Dimensions links to DataCite DOIs from papers, and Dimensions DAS availability.

In our analysis of data sharing practices, we have chosen to focus on nine countries that represent a diverse spectrum of research publishing and data sharing dynamics. The selection criteria for these countries are twofold:

1. The top five publishers of research globally
2. The remaining four countries were selected based on their unique characteristics and relevance to the topic of data sharing. While there is no formal scientific basis for their selection, we believe that their inclusion adds valuable insights into different cultural, economic, and regulatory environments surrounding data sharing.

By focusing on these nine countries, we aim to provide a comprehensive overview of data sharing practices that reflects both the leaders in research publishing and a broader array of perspectives from diverse contexts. This approach allows us to draw meaningful conclusions about trends and challenges in data sharing across different regions.

Country	MDC DCC	Dimensions
India	--	--
Japan	-	--
China	+	-
Ethiopia	+	++
Botswana	+	+
France	+	+
United States	+	-
United Kingdom	+	+
Germany	+	+

Comparison to mean of group (- below, -- far below, + above, ++ far above).

Table 2. Comparison of the percentage of publications with an author based in an institution in the analyzed countries, that have been linked to a dataset in the MDC DCC or the Dimensions Corpus

When plotting each country against the mean percentages for the nine countries analyzed, we can see which countries are having more success in encouraging researchers to make their data openly available. It also highlights differences in the locations of the data, comparing MDC DCC counts which include accession numbers and the Dimensions analysis which does not. Accession numbers are primarily used for subject-specific repositories, especially in fields that produce large volumes of data like genomics, proteomics, and structural biology.

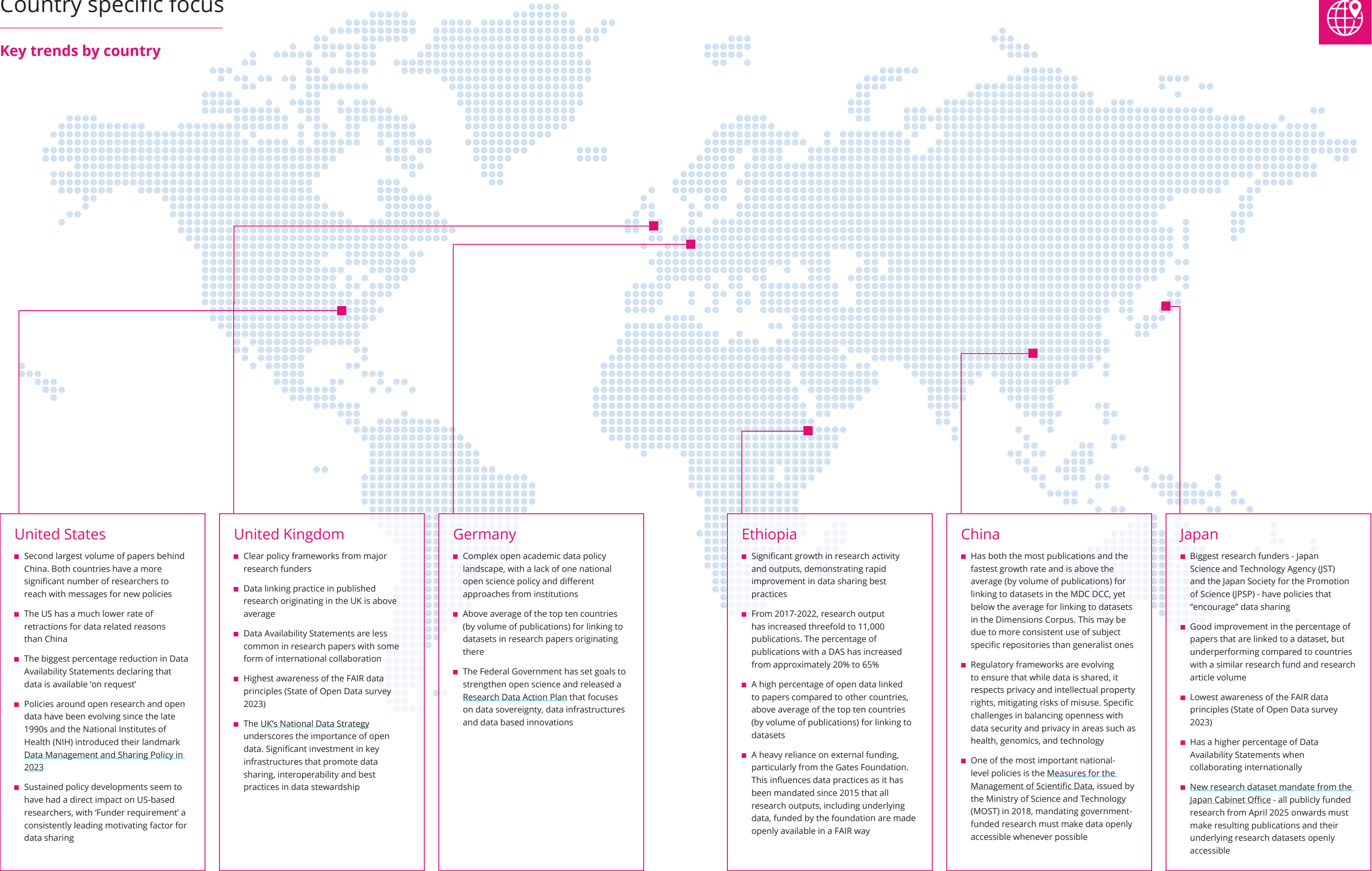
Relying solely on a single database, such as the MDC DCC, to analyze links from datasets to publications can present challenges in ensuring data accuracy and completeness. As a relatively new resource, the MDC DCC is still in the process of refinement and may contain gaps or inconsistencies in how data citations are tracked and recorded. This incompleteness can limit its reliability, especially as the database may not yet encompass the full spectrum of data sharing practices across disciplines, institutions, or regions. Furthermore, using only one source of dataset-publication links provides a narrow perspective and risks bias in understanding overall trends in open data practices.

To improve accuracy and validate the trends observed in the MDC DCC, it is essential to cross-reference its findings with other sources, such as the Dimensions database and Springer Nature’s Data Availability Statements. These additional resources offer different methodologies for cataloging data-publication links, and comparing results across them can reinforce the MDC DCC’s insights by showing correlation and consistency. Dimensions, for instance, tracks citation relationships comprehensively across scholarly content, while Springer Nature’s Data Availability Statements document data-sharing practices within publications, making both useful complements to the MDC DCC. When correlations among these databases demonstrate alignment, they collectively provide a more trustworthy and comprehensive picture of open data practices, strengthening the conclusions drawn from any one source alone.



# Country specific focus

## Key trends by country





# Country analysis by dataset



Percentage links to datasets in MDC DCC from countries with more than 50,000 publications per year

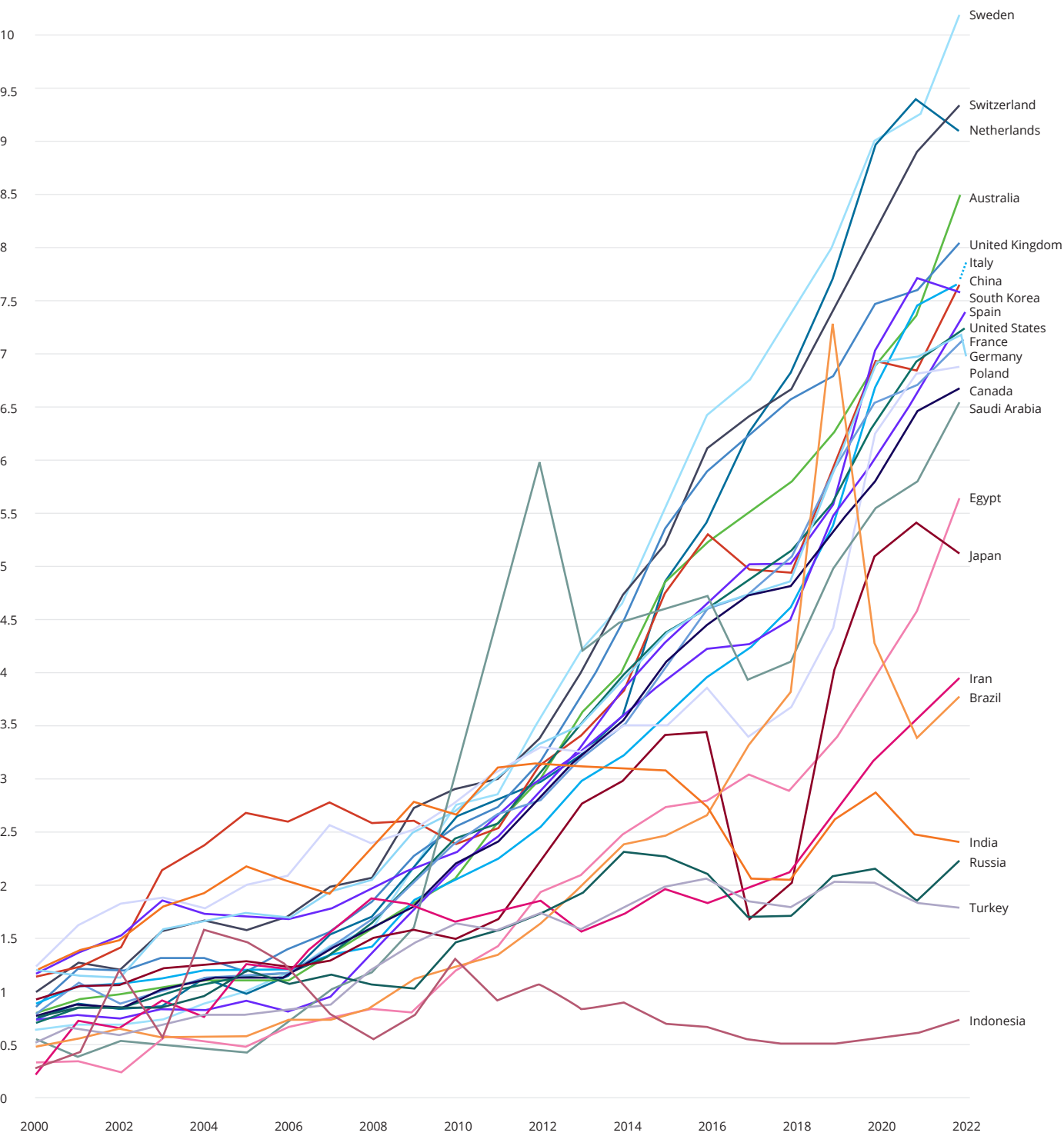


Figure 7. Percentage of papers linking to datasets when comparing countries with more than 50,000 publications per year

Highest percentage links to datasets MDC DCC from countries with more than 1,000 publications per year

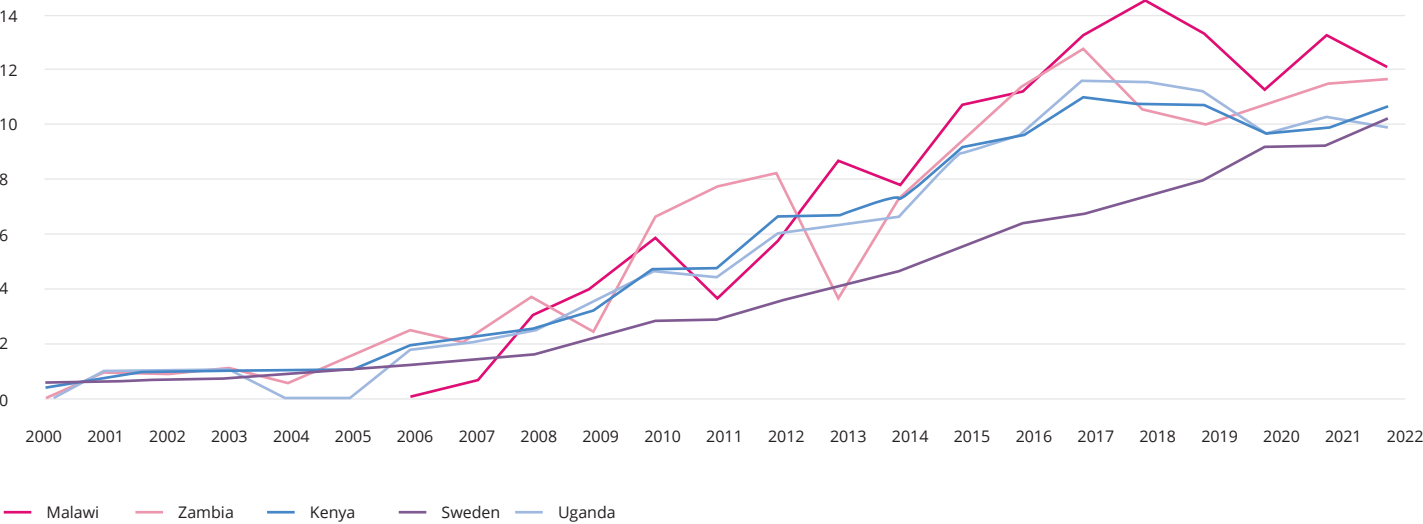


Figure 8. The top five results in terms of 'percentage of papers linking to datasets' when comparing countries who publish more than 1,000 papers per year

By initially comparing the data linking percentage from countries publishing over 50,000 papers annually, we can see which countries are achieving societal change at scale.

## Europe

Europe continues to lead, with an increased diversity of countries participating in open data publishing. Smaller nations, such as Estonia and Slovenia, show strong engagement, indicating that open data policies are effective even in countries with lower publication volumes. The EU's supportive stance on open science principles appears to have a widespread impact across both large and smaller research-producing nations.

## North America

The United States and Canada maintain high levels of open data publishing, as seen in larger countries. Even smaller research contributors, such as Mexico, are beginning to adopt open data practices, though the intensity is less pronounced compared to the larger, more research-intensive nations.

## Regions with increased open data but varied progress:

### Asia

China is doing well at achieving societal change at scale. Some nations with lower publication volumes - such as Malaysia and the Philippines - are making efforts to contribute to open data. However, large parts of the region, particularly in South Asia, still exhibit lower levels of open data publishing, indicating persistent barriers related to policy or infrastructure.

### Latin America

This region shows promising increases in open data participation, with countries like Chile and Colombia joining Brazil in adopting open data practices. Nonetheless, significant variability exists across Latin America, and some countries remain slower to adopt these practices, likely due to resource constraints and limited national mandates for data sharing.

If we expand the analysis to include countries that are publishing more than 1,000 papers per year we see some promising trends within Africa, where countries like South Africa, Ghana, and Kenya are leading regional open data initiatives. It is notable and encouraging that in this part of the analysis, four out of the top five countries leading in open data publishing are African nations. Historically, African nations have faced barriers in research visibility and access, often due to limited resources, digital infrastructure, and capacity-building opportunities.

Links between papers in the MDC DCC and datasets, filtered by country

Percentage of papers with a link to a dataset in the MDC DCC, by country

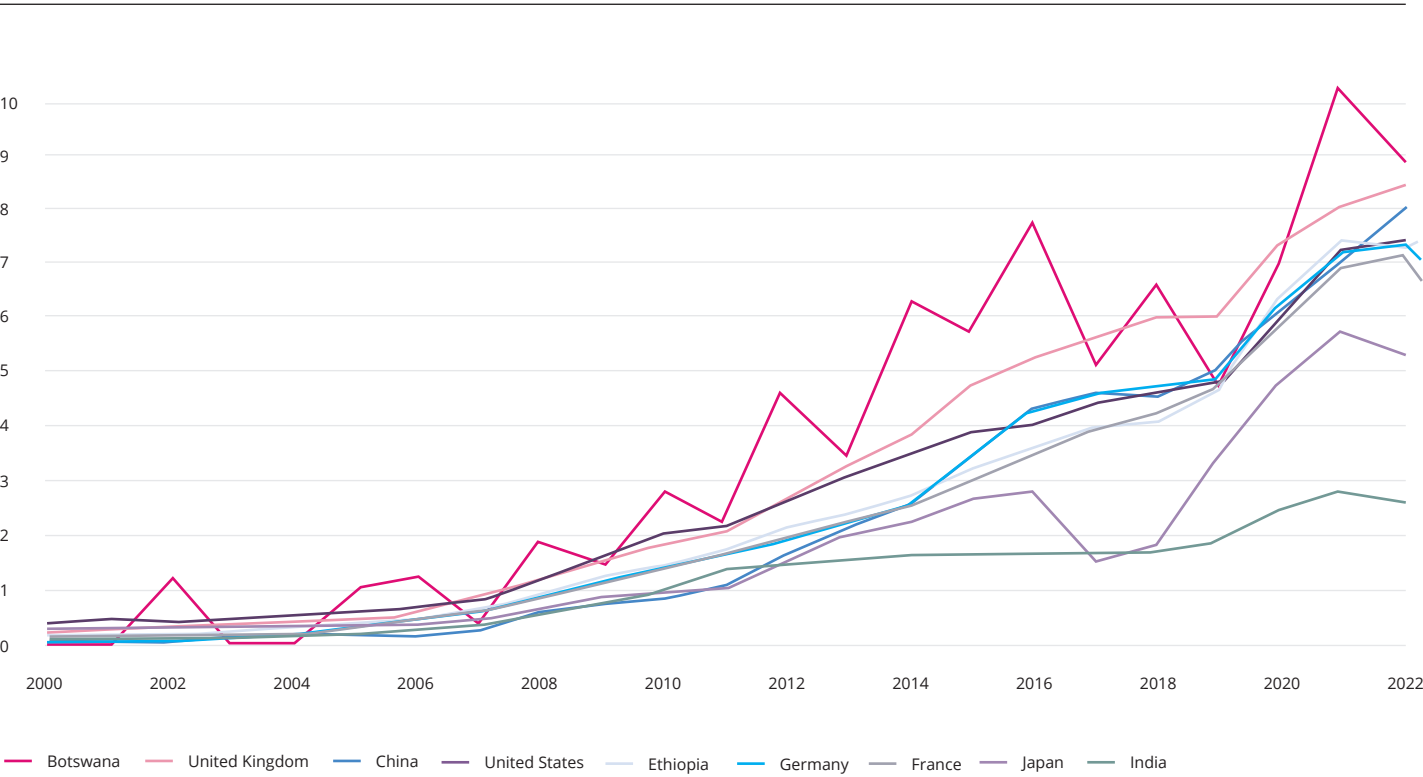


Figure 9. Percentage of papers linking to datasets in geographically diverse countries in the MDC DCC, to compare citation trends with those we see in the Dimensions corpus

Links between papers in the Dimensions corpus and datasets, filtered by country

We can see similar trends when looking at the percentage of links between papers in the Dimensions corpus and datasets, filtered by country.

Percentage of papers with a link to a dataset in Dimensions, by country

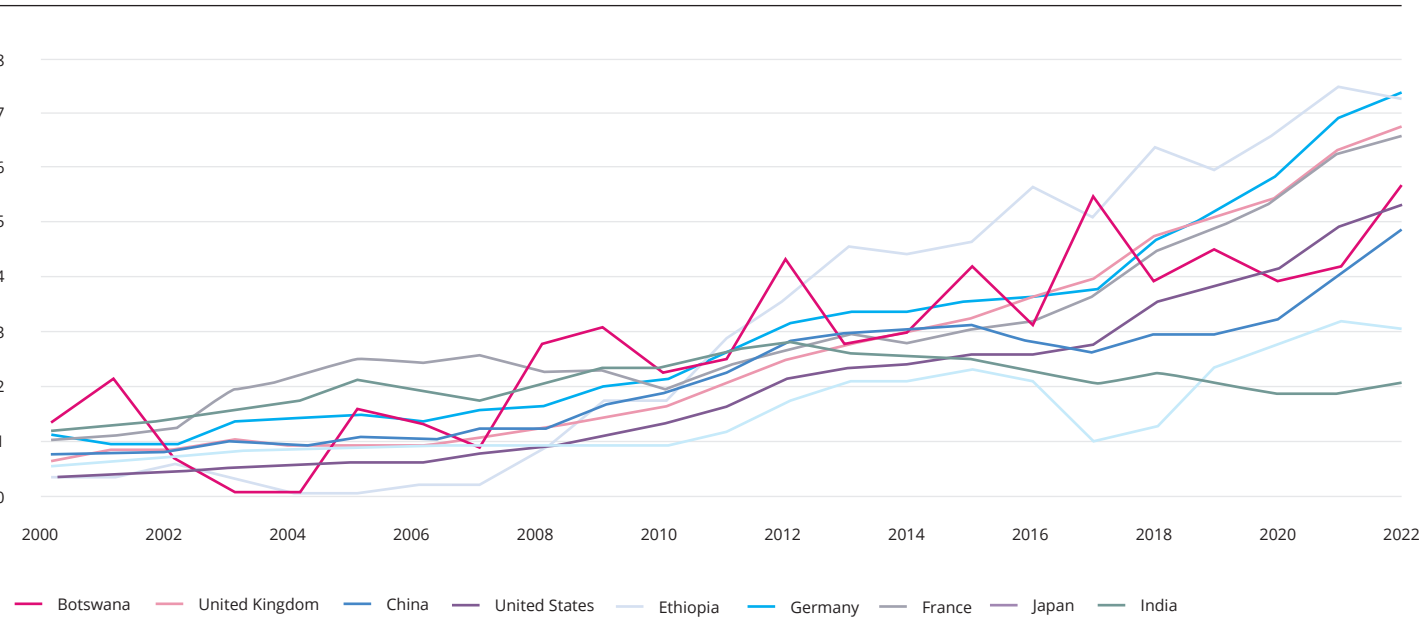


Figure 10. Percentage of papers linking to datasets in geographically diverse countries in the Dimensions Corpus, to compare trends with those we see in the MDC DCC

Uptake of Data Availability Statements, by country

Using Dimensions Trust and Integrity markers, we can analyze what percentage of papers have a DAS. This does not necessarily mean that there is data available, as we discuss later in this report. In these two plots, both the MDC DCC and Dimensions Corpus reveal consistent country-level trends in the prevalence of dataset links in academic publications. In both visualizations, countries that lead in open data sharing, such as the United States, the United Kingdom, and Germany, show a higher percentage of publications with dataset links. This alignment suggests a correlation in how open data practices are distributed globally, with specific regions demonstrating stronger data-sharing cultures. These shared trends between the MDC DCC and the Dimensions Corpus indicate that open data policies and mandates are likely having a similar impact across datasets, supporting the reliability of these findings.

Additionally, both plots show regions where dataset linking is less common, particularly in parts of Asia and Latin America. This consistency underscores potential challenges in open data adoption, such as policy limitations or infrastructural barriers, that are affecting multiple datasets. By identifying similar trends across both sources, these comparisons reinforce the validity of using multiple databases to track data-sharing practices and demonstrate that the MDC DCC’s patterns are not isolated but part of a broader and reproducible observation in academic data sharing practices.

Percentage of papers with a Data Availability Statement - Dimensions Corpus

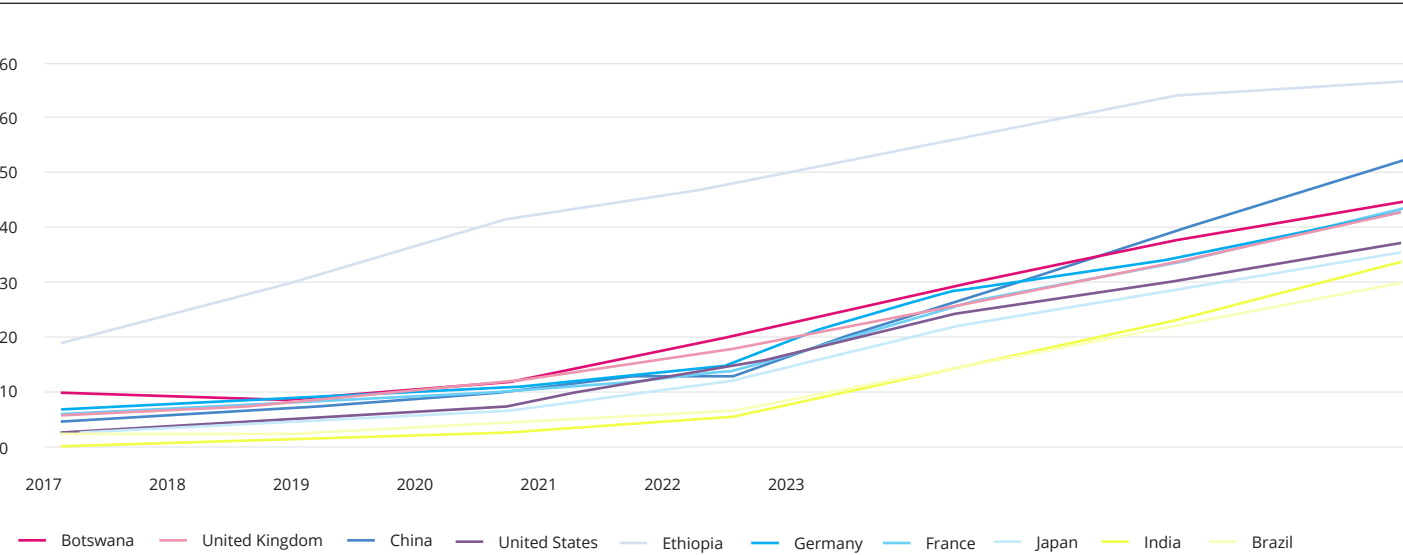


Figure 11. Percentage of papers with a Data Availability Statement - Dimensions Corpus (2017-2023)



Data sharing methods in Springer Nature Data Availability Statements, by country

Data Availability Statements - Data available “on request”

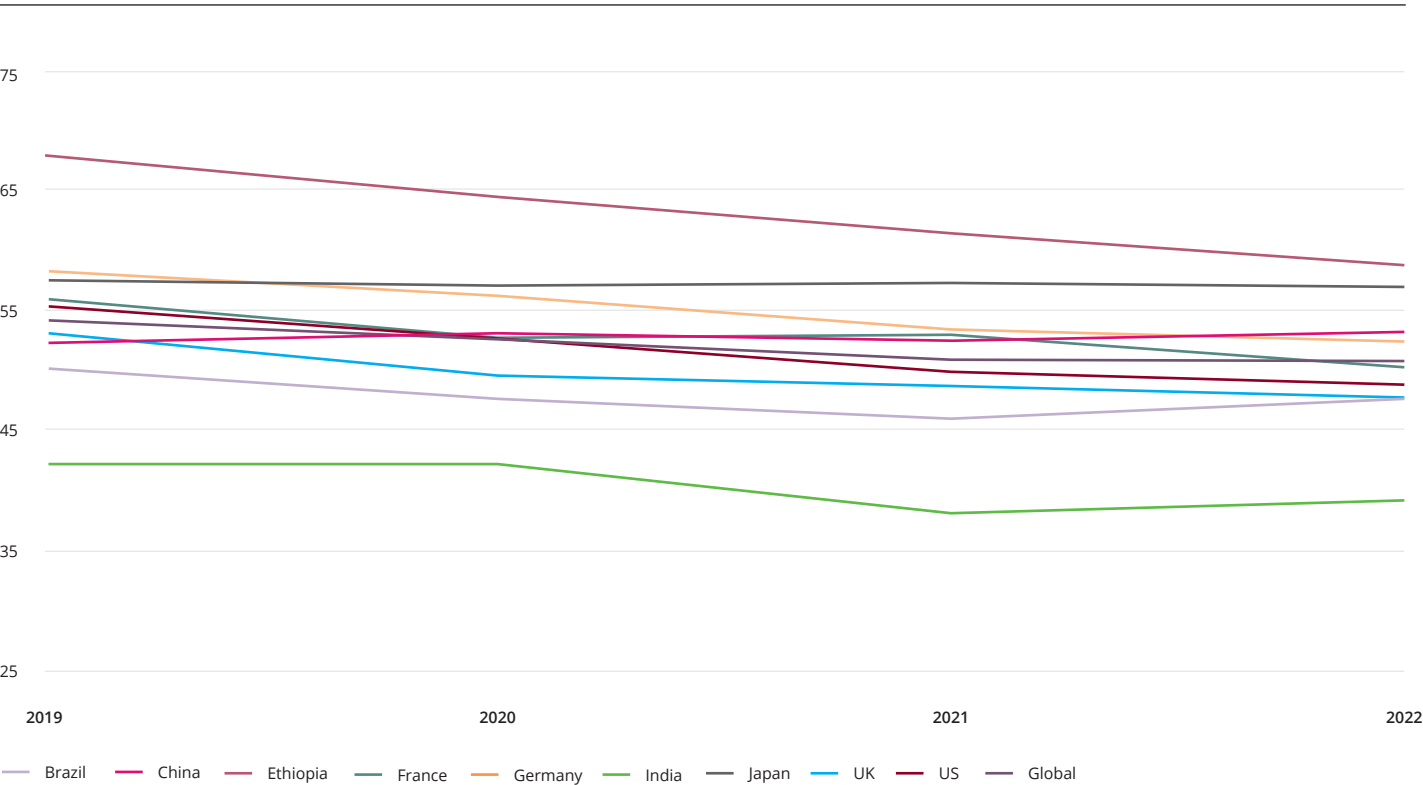


Figure 12. Percentage of Data Availability Statements that listed “Data Available on Request” from Springer Nature Journals (2019-2022)

“On request” data sharing is still dominant but in decline

Our analysis revealed that sharing data upon request remains the most common method across all countries, but it is in decline. This method, where authors indicate that data can be obtained by contacting them directly, suggests a preference for controlled data sharing, possibly due to concerns over data misuse, privacy, or intellectual property.

Between 2019 and 2022, there was a noticeable decrease in “on request” sharing in nine out of ten regions analyzed, as well as globally, with reductions ranging from approximately 1% to 9%. For instance, Ethiopia experienced the greatest decline of 8.79%, signaling a potential shift towards more open sharing practices.

In countries like China, Germany, and Japan, over 50% of articles still use “on request” sharing, indicating a sustained reliance on traditional data sharing practices. The UK, US, and Brazil have seen rates drop below 50%, hinting at a gradual move towards more open and accessible methods of data sharing.

India is the only country represented with consistently under 50% use of “on request”. However, this is mainly complemented by an increase in DAS declared as “not applicable” over the time period (rather than increasing data being shared in repositories for example). Ethiopia had

the highest rates of “on request sharing” but also saw the greatest decline (8.79%). China is the only country not to see a decline in “on request” with rates remaining broadly stable and a slight increase of 1.07%. Japan was similarly consistent, seeing only a 0.76% drop.

Modest regional increases in repository sharing

Repository sharing, where data are deposited in publicly accessible repositories, is considered a cornerstone of open science due to its facilitation of data accessibility and reuse. Our analysis showed that repository sharing has remained relatively consistent overall, with small gains observed in several countries.

Germany demonstrated a notable increase in repository sharing, rising from 23.13% in 2019 to 26.59% in 2022. This growth suggests a positive trend towards embracing open science practices within the German research community.

Similarly, the US and Ethiopia showed slight increases in repository use. The US, UK, Germany, and France exhibit similar patterns in repository sharing, clustering around a 25% sharing rate. This similarity may reflect shared policies, funding mandates, or cultural attitudes towards data sharing within these countries.

Conversely, India and Brazil experienced declines in repository sharing, highlighting regional disparities and

Data Availability Statements - “Data in manuscript”

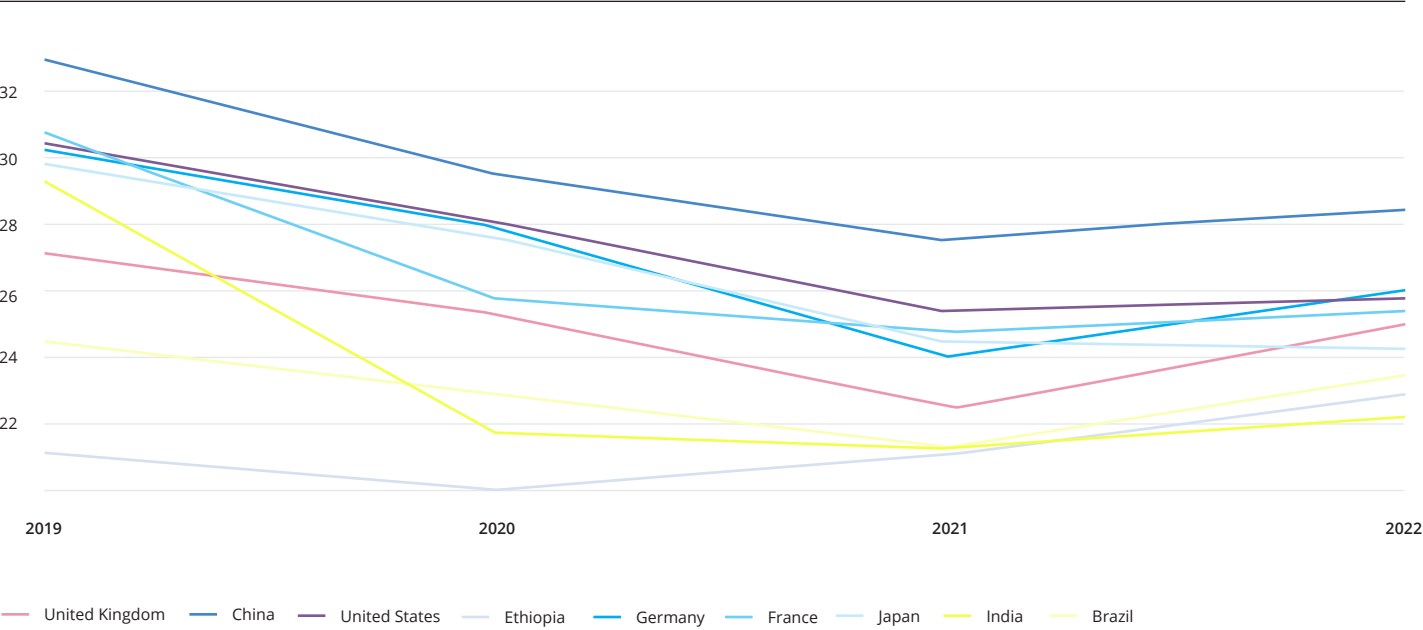


Figure 13. Percentage of Data Availability Statements that listed “Data in Manuscript” from Springer Nature Journals (2019-2022)

potential barriers to adopting open science practices. In 2022, only 8.50% of Indian authors and 5.52% of Ethiopian authors deposited data in repositories, indicating challenges such as limited access to repositories, lack of awareness, or insufficient institutional support.

Comparison with previous data

Previous analysis of BMC journals (part of Springer Nature) indicated that data available “on request” was previously more dominant, being present in 60% of BMC articles in 2017–2018. While there is a difference in the precise methodology and journals included, this earlier study followed a broadly similar approach of DAS categorization to the analysis presented here, and together they track an overall decline in data available “on request” statements in Springer Nature journals over time.

Springer Nature has expanded the requirements for DAS to disciplines and journals with less-established data sharing norms over the analysis period. While there is a clear need for discipline-specific approaches and support to improve repository sharing, “on request” DAS are declining.

Addressing the challenges and disparities

In summary, our analysis of data sharing practices in Springer Nature journals from 2019 to 2022 reveals a complex and evolving landscape. While “on request” data sharing remains dominant, it is in decline, and modest gains in repository sharing are observed in some regions. However, these positive trends are uneven, with significant regional disparities that may reflect differences in resources, infrastructure, and community norms.

The findings underscore the importance of providing practical support to researchers, beyond implementing policies, to encourage the adoption of open science practices. This includes investing in infrastructure, offering training and resources, and fostering a culture that values and rewards data sharing.

Addressing the challenges and disparities identified in this study is crucial for advancing open science globally. By leveraging insights from DAS analysis, stakeholders can better understand researchers’ behaviors and needs, tailoring interventions to promote reproducibility, efficiency, and integrity in research worldwide.

# Trends at a funder level



We can also compare the percentage of papers citing a dataset in the MDC DCC or the Dimensions corpus, filtered by the funding agency associated with the paper.

Percentage of papers funded by large funders with a link to a dataset in the MDC DCC

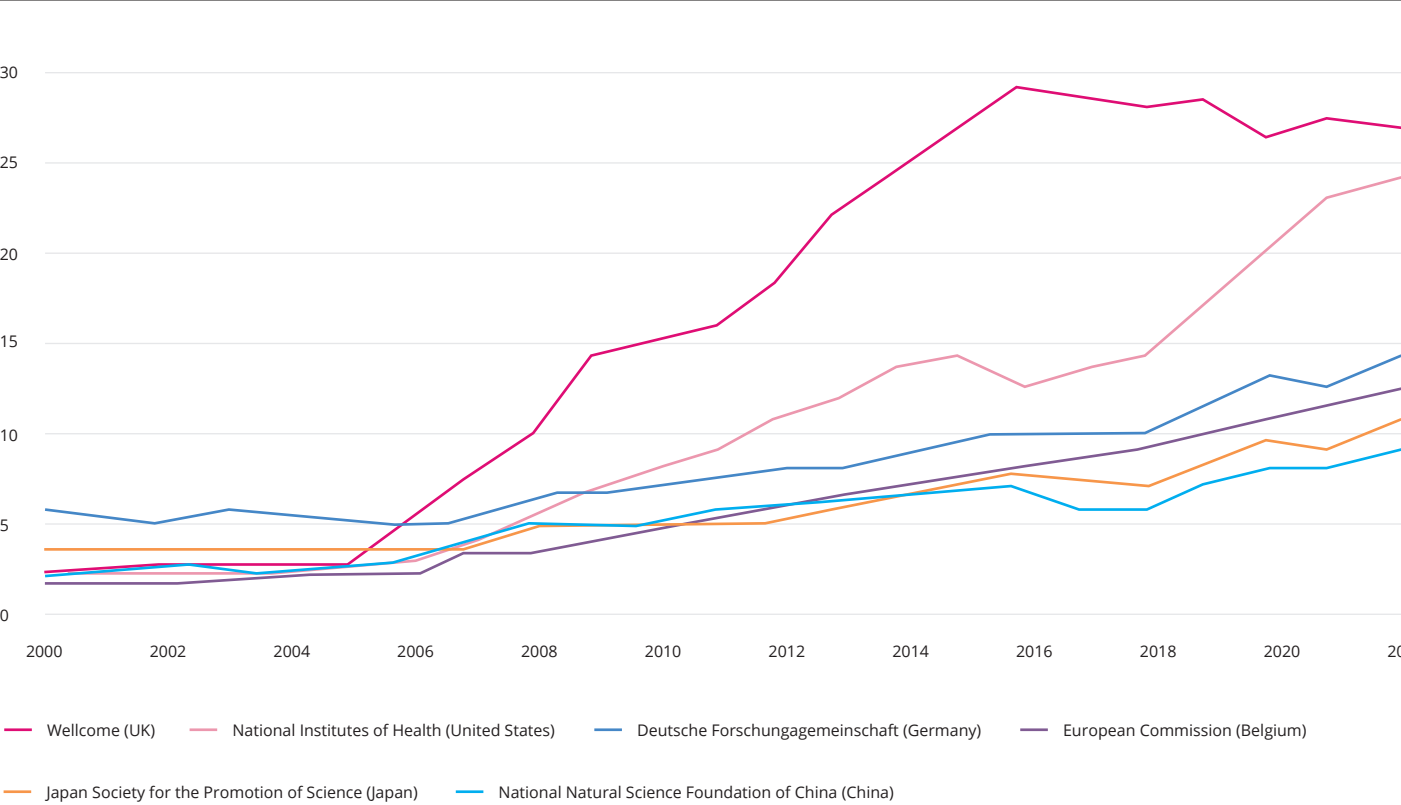


Figure 14. Percentage of papers citing a dataset in the MDC DCC when filtered by the funder of the paper

Funders were chosen to give a broad geographic spread, whilst accounting for those which fund the largest amount of publications.

Wellcome has had an [open access policy](#) since 2005 which requires that all research papers that have been accepted for publication in a peer-reviewed journal, and are supported in whole or in part by Wellcome funding, be made freely available through the PubMed Central (PMC) and Europe PubMed Central (Europe PMC) repositories as soon as possible and in any event within six months of the journal publisher’s official date of final publication.

Wellcome mandates that all research it funds must share data openly, particularly research with implications for public health. Wellcome enforces its policies by making compliance with data-sharing mandates a condition for funding. This proactive approach, combined with the data here suggests researchers are likely to follow through on data-sharing practices, resulting in higher dataset-linking rates in Wellcome-funded publications. While earlier and well-defined

policies generally support higher dataset-linking rates, other factors - such as the strictness of mandates, enforcement mechanisms, and support resources - also play a significant role.

The [National Institutes of Health \(NIH\) Public Access Policy](#) is an open access mandate, drafted in 2004 and mandated in 2008, requiring that research papers describing research funded by the National Institutes of Health must be available to the public free through PubMed Central within 12 months of publication. In 2003, NIH launched the NIH Data Sharing Policy, which requires grant applicants seeking \$500,000 or more in direct costs to include a data-sharing plan in their application.

In the case of the European Research Council (ERC), its policies on open data and the FAIR principles were implemented relatively early (around 2017) and aligned with the broader European Union emphasis on open science. However, as the middle performing among the five funders investigated here, this placement could indicate that ERC’s

policies, while strong, may be somewhat less enforceable or have less institutional support in certain areas compared to NIH, for example. NIH, as a national agency with a long-standing mandate for open data, directly ties compliance to funding, which may contribute to its higher position. ERC’s mandate, although comprehensive, spans multiple countries with varying compliance structures, potentially influencing its dataset-linking rate.

Thus, while policy timing is crucial, other elements - like mandate enforcement and consistent support structures

- also significantly impact the overall success of data-sharing practices. ERC’s mid-range position suggests that a combination of factors, including but not limited to policy timing, contributes to dataset-linking success across funders. Interestingly, funders appear to have much better rates overall in percentages of papers linking to datasets, over institutions and even countries. This may be due to inconsistencies in available open metadata to do this analysis, or it could be that researchers are much more inclined to listen to the people that fund them.

Number of NIH funded papers with a link to a dataset - based on the MDC DCC

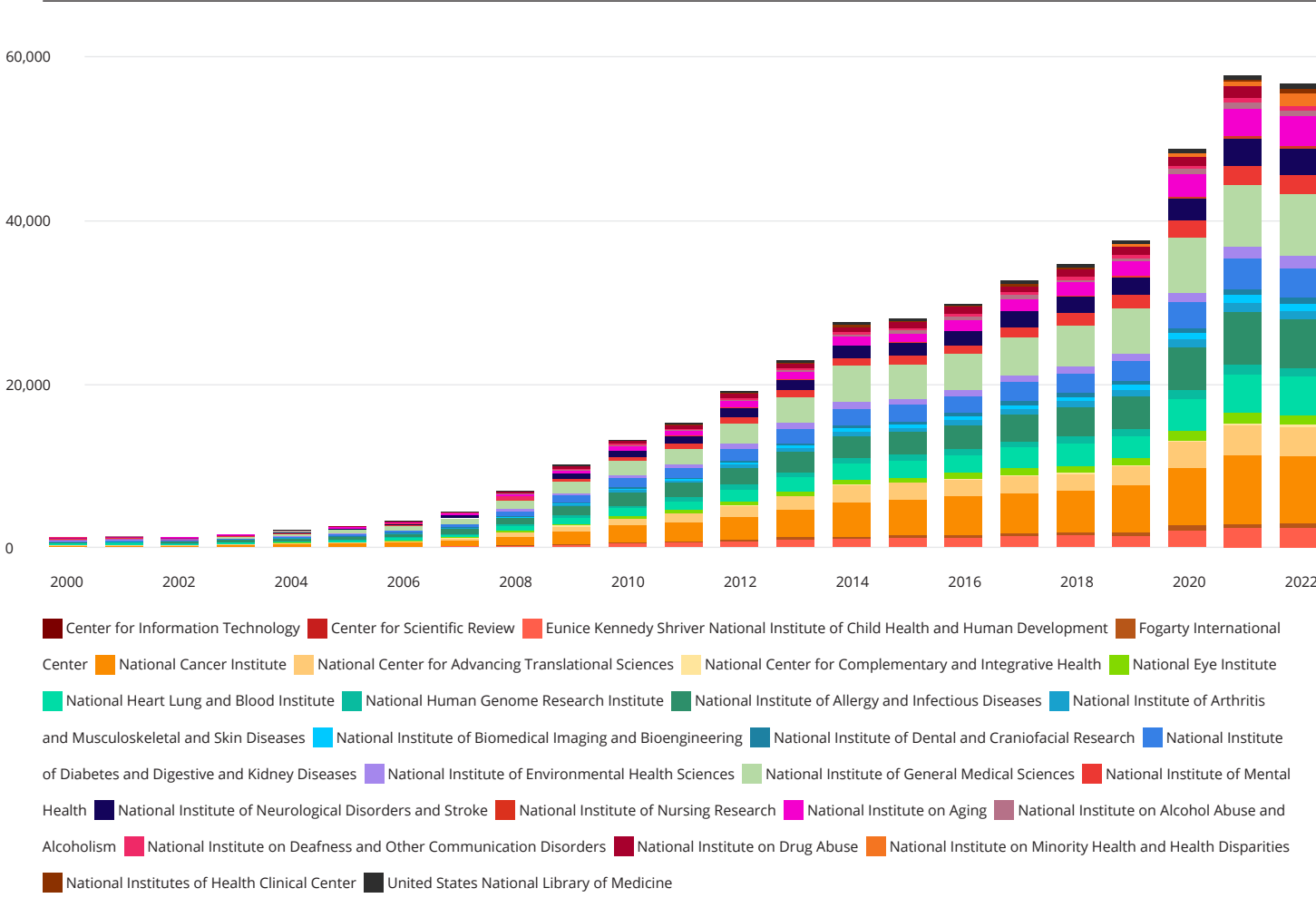


Figure 15. Percentage of papers citing a dataset in the MDC DCC for NIH funded papers over time

## Measuring impact of funder policies

In this example, by tracking the growth of papers linked to datasets funded by the NIH, we can quantify and track the impact of any changes to funder policy. The NIH policy has evolved, and in 2023, NIH implemented the [NIH Data Management and Sharing \(DMS\) Policy](#), which requires all NIH-funded researchers to submit a data management and sharing plan. We will be able to track the impact of this mandate, and any potential rollback of [Office of Science and Technology Policy \(OSTP\) memo](#) due to lack of funding, or changes of governmental focus, going forward.





# Trends at a university level



Many universities now have data-sharing policies as part of their efforts to promote open science and research transparency. These policies are increasingly being developed in response to funder requirements (such as mandates from the NIH or the EU's Horizon 2020), as well as a growing recognition of the benefits of data sharing for

advancing research and fostering collaboration. The following graphs look at the percentage of papers that link to datasets, with one or more authors from each academic institution. The analysis focussed on the universities with the highest volume of publications since 2010.

Percentage of papers linking to a dataset in the MDC DCC - highest publishing universities

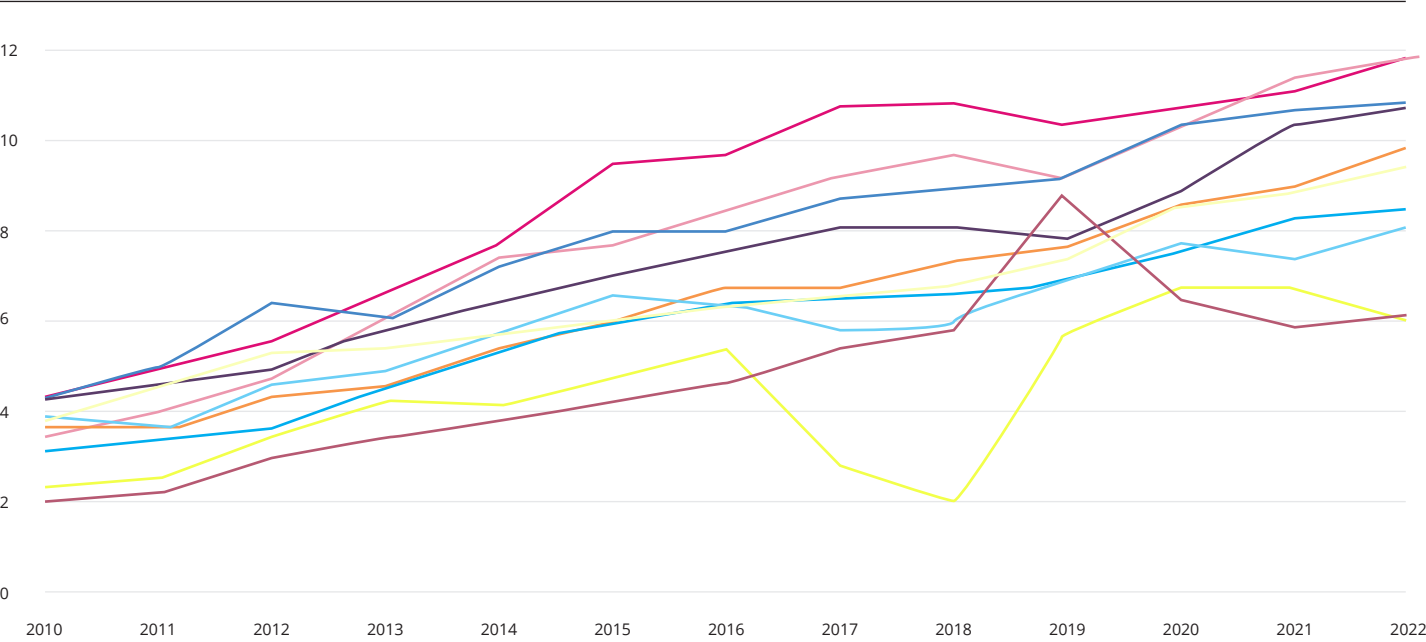


Figure 16. Percentage of papers citing a dataset in the MDC DCC when filtered by university. Universities selected as the highest publishing universities by volume of papers

Harvard University (United States) University of Oxford (United Kingdom) University College London (United Kingdom) Johns Hopkins University (United States)  
University of California, Los Angeles (United States) Stanford University (United States) University of Michigan, Ann Arbor (United States)  
University of Chinese Academy of Sciences (China) Universidade de São Paulo (Brazil) The University of Tokyo (Japan)

Percentage of papers linking to a dataset in Dimensions in the highest publishing universities

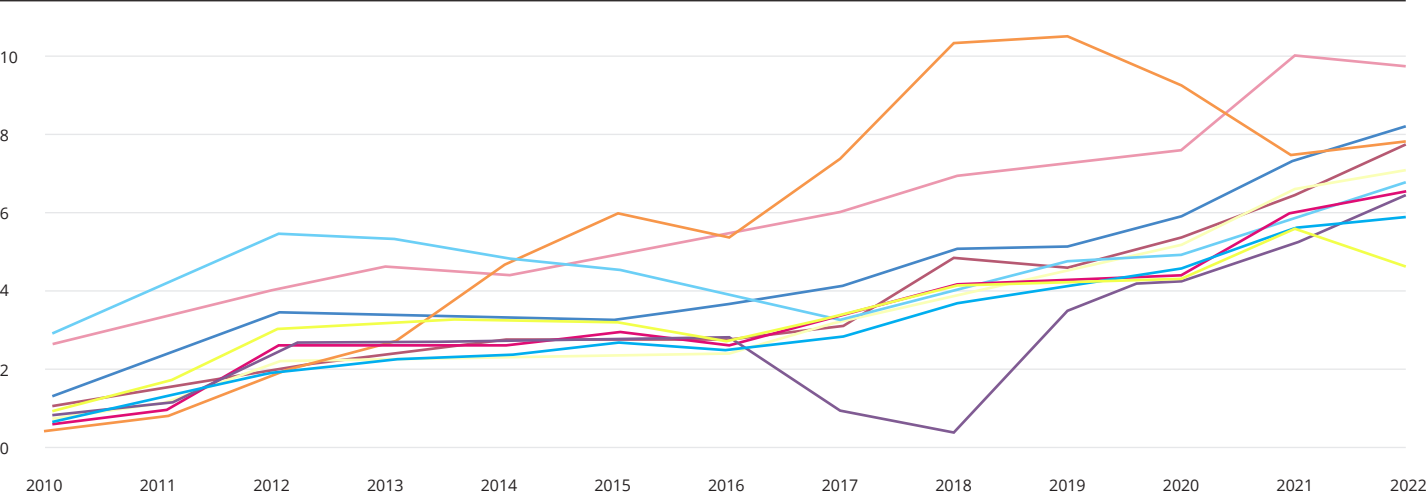


Figure 17. Percentage of papers citing a dataset in the Dimensions Corpus when filtered by university. Universities selected as the highest publishing universities by volume of papers

As with the country-based analysis, we see a strong growth in papers linking to data from universities globally. Although there are differences regionally, we see a spread of just 5-10%, which is less significant than the >85% of articles that are not linking to data by the year 2022.

N.B. The fact that these are lower percentages than those seen in other sections of the report is likely due to a lack of consistent metadata highlighting research institution affiliation with publications. While we can see trends and suggest what is happening, it may be that the data needed to analyze this consistently needs improvement.

Institution	Position (MDC DCC)	Position (Dimensions)	Average Position
University of Oxford	2	1	1.5
University College London	3	2	2.5
Harvard University	1	7	4
Stanford University	6	4	5
University of California	5	5	5
Universidade de São Paulo	9	3	6
John Hopkins University	4	8	6
University of Chinese Academy of Sciences	8	6	7
University of Michigan	7	9	8
The University of Tokyo	10	10	10

Table 3. Average position comparison for universities based in the 9 countries investigated

Table 3 (above) ranks universities by the proportion of their publications that include dataset links, highlighting significant variation across institutions. These are leading universities globally. While all universities have more papers that link to datasets in 2022 than in 2010, the rate of growth is not consistent. Harvard University has more than double the percentage of papers that link to datasets than Universidade de São Paulo or the University of Tokyo. Dimensions data reveals similar trends, with the same high-performing universities also displaying strong dataset-linking practices. This consistency across both the MDC DCC and Dimensions data supports the reliability of the observed patterns.

Subject-specific disparities may also be a factor in how researchers are sharing their data. However, the universities analyzed here have prominent programs in the biomedical and life sciences, engineering, computer science, and physical sciences. This suggest that the differences we see may be due to other factors, such as regional differences in policies and education on open data.



“Policy changes alone are insufficient to drive the desired shift towards open data practices.”

Graham Smith, Open Data Programme Manager  
Springer Nature

## Background on the universities analyzed

A broad range of universities were investigated. The universities in this list were selected as being highly respected organizations with open policies and regional variety.

**University of Oxford:** Oxford has an open access and [research data management policy](#) that mandates the sharing of research data wherever possible and appropriate, in line with funder requirements.

**Harvard University:** Harvard’s [Open Access Policy](#) applies to research outputs, including datasets. It encourages open sharing of data through repositories like Harvard Dataverse.

**University of California System:** The University of California [has open access policies](#) across its campuses, encouraging researchers to deposit their data in open repositories and comply with funder guidelines.

**Johns Hopkins University:** Johns Hopkins encourages open data practices, especially in compliance with funder mandates. It has implemented a [data management policy](#) requiring researchers to deposit datasets in public repositories whenever possible, ensuring the data is accessible, reusable, and preserved long-term. This is in line with its broader commitment to open science.

**University of Michigan:** The University of Michigan has a comprehensive [Research Data Management and Sharing Policy](#). Researchers are required to create data management plans and, where appropriate, make data openly available. Michigan also supports a range of data repositories and services to facilitate the sharing and preservation of research data.

**University College London (UCL):** UCL’s [Research Data Policy](#) requires researchers to manage and share their data responsibly and ensures that data generated from publicly funded research is open and accessible.

**Stanford University:** Stanford promotes open data as part of its broader commitment to open science and [requires researchers to make their data available](#) in line with funder mandates.

**Universidade de São Paulo (USP):** USP has a [clear commitment to open access](#) through its Open Access Repository (BDPI), established in 2012. This repository stores and provides access to the digital outputs of research, including articles, theses, and other academic work, in line with international open access standards.

**University of Chinese Academy of Sciences (UCAS):** UCAS follows the Chinese Academy of Sciences’ (CAS) Open Data Policy, which promotes open access to research data and mandates that publicly funded research outputs be deposited in institutional repositories. CAS’ data-sharing initiatives focus on making scientific data available to foster innovation and international collaboration.

**The University of Tokyo:** The University of Tokyo has developed initiatives to support [open access to research data](#), encouraging researchers to deposit data in publicly accessible repositories. While there is no single university-wide policy, Tokyo supports national efforts like the [Japan Science and Technology Agency’s \(JST\) Open Science Framework](#), which emphasizes open data sharing in research.

# Conclusions



“*This report serves several purposes. Firstly, it confirms that the qualitative survey data of the State of Open Data over the last nine years has been accurate when compared to quantitative actions of researchers worldwide. Secondly, it provides a pathway which can be taken to use data to create the carrots and rewards metrics for researchers who publish their open data.*”

Mark Hannel, VP Open Research  
Digital Science

## Key takeaways

Open data is on the brink of becoming a recognized scholarly output globally. This report serves several purposes. Firstly, it confirms that the qualitative survey data of the State of Open Data over the last nine years has been accurate when compared to quantitative actions of researchers worldwide. Secondly, it provides a pathway which can be taken to use data to create the carrots and rewards metrics for researchers who publish their open data. This is not only a great return on investment for funders on the research outputs they pay for, it also can feed a new wave of machine-generated findings and move the needle on the speed of research. Finally, it highlights that we cannot think of open data as a uniform entity. There are societal, cultural, regional, and subject-specific differences that need to be catered to whilst we work with researchers on the path towards an open data future.

- 1 **More carrots, more change**
- A need to move through a four step process of change: policy, mandate, compliance and measurement
  - A lack of credit is stopping open data sharing
  - Initiatives like an S-Index will measure the extent and effectiveness of data sharing

- 2 **Resource disparities**
- Progress in some countries is hindered by limitations in internet connectivity, institutional support, and lack of awareness
  - There is a need for enhanced collaboration, innovation, and more equitable access to knowledge

- 3 **Practical support**
- Policy changes alone are insufficient to drive change
  - Training, user-friendly repositories and clearer data policies are needed

- 4 **Acknowledging nuances by discipline**
- Tailored support and resources to address discipline specific challenges
  - A sustained effort rather than a “quick fix”



# Recommendations and next steps

This report has for the first time provided quantification of what researchers are actually doing in practice. It provides an analysis that goes beyond thoughts and attitudes into real world behaviors, responding to the question: How do we bridge policy and practice gaps? There are four key recommendations for the future:

## 1. We can now offer carrots to drive change

By also looking at data from the State of Open Data Survey, which explores researcher motivations and attitudes, we are both able to assess researcher behavior in a quantitative way, as well as a qualitative way. [Responses from the State of Open Data 2023](#) show citations are by far the main motivation for researchers to share their data (20%), although many see the public benefit (12%) and also the benefits of increased impact and visibility, through citations to papers and data.

In order to drive societal change in academia, we need to use both carrots and sticks. To succeed, we need to move the process through the following steps:

- Policy
- Mandate
- Compliance
- Measurement

In the nine years that the State of Open Data has been running, we have seen policies forming and becoming mandates. This has driven a lot of the compliance we see today. Researchers are sharing because they are told they have to as a requirement of funding. The survey continues to tell us that the main reason that is stopping them engaging with open data publishing in a more serious manner is a lack of credit for their open data. Researchers cannot get credit if there is not a way to consistently measure data metrics across platforms and repositories. The MDC DCC allows us to do this. [The NIH Data Sharing Index \(S-Index\) Challenge](#) seeks innovative approaches to quantify and evaluate data-sharing practices by biomedical researchers. The challenge is aimed at developing an “S-Index” to measure the extent and effectiveness of data-sharing, encouraging transparency and accessibility in research. Submissions are judged based on originality, feasibility, impact, and scalability of the proposed metrics.

### What circumstances would motivate you to share your data?

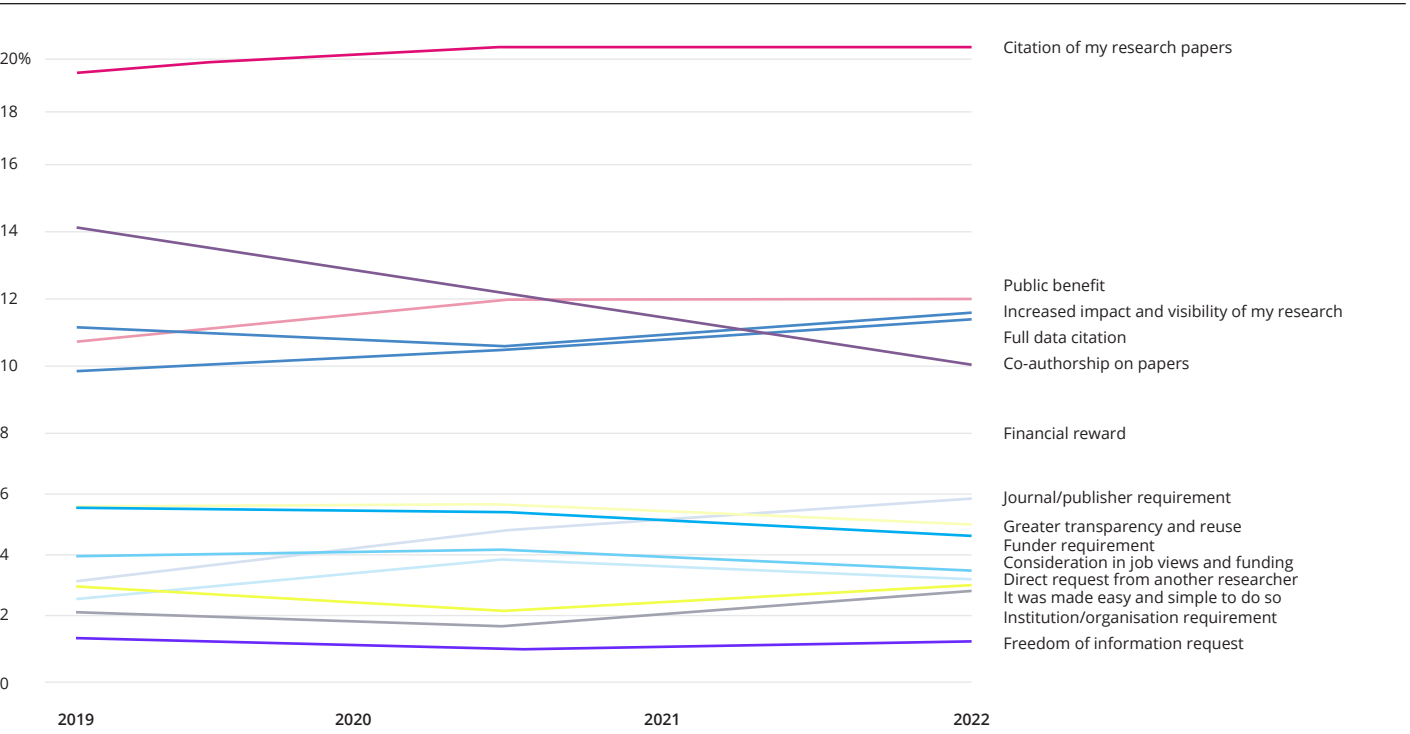


Figure 16. Longitudinal graph of State of Open Data responses to the question “What circumstances would motivate you to share your data?” from years 2019-2022

The goal is to foster more widespread and high-quality sharing of scientific data. [The DataWorks! Prize](#) organized by the Federation of American Societies for Experimental Biology (FASEB) and NIH, is a challenge that encourages researchers to propose impactful secondary data analysis projects using existing biomedical data.

Being able to measure the impact of researchers sharing their data and ultimately reward them for doing so means that we are on the verge of having both carrots and sticks to enable open data sharing.

## 2. Resource disparities are holding back progress

From the DAS analysis we see positive trends in open science practices, such as modest increases in repository sharing and reductions in “on request” sharing. These trends are more pronounced in developed regions like Europe and the US. This observation raises concerns about a potential divide where open science becomes the preserve of better-resourced research environments, potentially marginalizing researchers in low- and middle-income countries (LMICs).

Resource disparities, including access to infrastructure, funding, and training, may hinder the adoption of open science practices in LMICs. Researchers in these regions may face challenges such as limited internet connectivity, lack of institutional support for data management, or insufficient awareness of open science benefits and requirements.

Addressing these disparities is crucial to ensuring that the benefits of open science - such as enhanced collaboration, innovation, and equitable access to knowledge - are globally realized. Initiatives that provide resources, training, and infrastructure support can help bridge the gap and promote inclusive participation in open science.

## 3. There is a need for practical support beyond policy

The reductions in “on request” and “in manuscript” sharing are generally not complemented by significant rises in repository sharing. In some cases, we observe increases in data being declared as “not applicable” or shared through non-repository venues. This pattern suggests that policy changes alone are insufficient to drive the desired shift towards open data practices.

Researchers require practical support, including training on data management and sharing, access to user-friendly repositories, and clear guidelines on data policies and expectations. For example, [in the Nature portfolio](#), editorial interventions and repository integration are correlated with a decrease in “on request” sharing and growth in repository

use. This illustrates the impact of targeted support and infrastructure in facilitating open data practices.

## 4. Sustained efforts are required to respond to the challenges in diverse research areas

Within Springer Nature, the expansion of DAS policy has led to a growing number of statements from a wider range of research areas, including disciplines where data sharing norms are less established. Fields such as humanities or certain social sciences may have different methodological approaches, data types, and sensitivities that pose challenges for data sharing. Researchers in these areas may lack established community practices, appropriate repositories, or may be dealing with data that is inherently difficult to share due to confidentiality or ethical concerns. As more articles are published with DAS in these areas we could reasonably expect to see a greater proportion of “on request” sharing and less sharing in repositories. Against this backdrop the relatively stable use of repositories globally, and increases in certain regions, can be seen as a positive sign. It suggests that open data practices are gaining traction even in diverse research contexts, but tailored support and resources are needed to address discipline-specific challenges.

Additionally, the trends observed in Springer Nature’s broader journal portfolio can be contrasted with those in the Nature Portfolio and PLOS journals, both of which have had robust data policies in place for a longer duration. These policies have been associated with significant increases in repository sharing in recent years (see for example PLOS’s [Open Science Indicators](#) and the [Nature journal repository uptake](#)). This suggests that the longevity and enforcement of data policies, coupled with practical support and infrastructure, are crucial factors in promoting open data practices. The sustained commitment by journals like those in the Nature Portfolio and PLOS highlights the positive impact that comprehensive data policies can have on enhancing data accessibility and fostering a culture of openness in scientific research.

Report DOI:  
[10.6084/m9.figshare.27337476](https://doi.org/10.6084/m9.figshare.27337476)

2024 Survey data:  
[10.6084/m9.figshare.27291627](https://doi.org/10.6084/m9.figshare.27291627)

Country, Funder, and Affiliation dataset:  
[10.6084/m9.figshare.27900828](https://doi.org/10.6084/m9.figshare.27900828)

Springer Nature DAS analysis data:  
[10.6084/m9.figshare.27886320](https://doi.org/10.6084/m9.figshare.27886320)



What’s next?



**I am very excited about open data right now.**

If we look at open research in general and closed vs open publishing, there is now more open publishing than closed. Open research is now officially an inevitability.

We’re now in a place where we consistently see around 2 million datasets being published every year; this is the same amount of articles that we saw published annually in the year 2000.

This report was a great opportunity to look into what is really driving this data sharing and try to understand what is working and what we need to do more of to both sustain these figures and increase them.

By slicing the data by countries, by funders and by universities we can see successful examples to follow but also identify where further intervention is still needed.

Owing to new developments in the space, such as the aforementioned S-Index and the creation of the MDC DCC, both monitoring data sharing success rates and attributing credit and recognition for data sharing are now becoming possible, creating a very exciting time for open data and open research more generally.

At Digital Science, we believe that research is the single most powerful transformative force for the positive development of humanity, and as such, knowledge and research outcomes should be shared for common good. By making research and research data as open as possible, society derive maximal benefit.

**Mark Hahnel is the VP Open Research at Digital Science. He is the founder of Figshare, which he created while completing his PhD in stem cell biology at Imperial College London. Figshare 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.**



**With this quantitative analysis, we set out not only to understand but to inspire action.**

These insights mark our joint effort to stimulate broader conversations—conversations that look beyond researcher intentions and dive into real-world practices. By examining the uptake of open science in diverse contexts, we aim to show how its implementation varies widely across regions. This calls for tailored, region-specific interventions.

Importantly, the differences we’ve observed across our data sources highlight that the measurement of open science is still evolving. Defining and refining this field will take continued research and collaboration. As an organization, Springer Nature is committed to complementing policy with practical support, recognizing that lasting change requires resources, tools, and community engagement just as much as mandates. We invite the community to join us in building on these findings, bridging policy and practice for a future where open science truly thrives.

**Graham Smith 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.**

Contributors

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

Alongside the main authors of the report, Mark Hahnel and Graham Smith, we would also like to thank Ann Campbell, Technical Product Solutions Manager at Digital Science for her support in our data analysis.

Acknowledgements

Figshare, Digital Science and Springer Nature extend their thanks to survey respondents who continue to provide us with unparalleled insights and enable us to bring reports of this nature to the wider research community.



A collaboration between Figshare, Digital Science,  
and Springer Nature

