
RESEARCH DATA IN SCIENTIFIC PUBLICATIONS: A CROSS-FIELD ANALYSIS

A PREPRINT

Puyu Yang

Institute for Logic, Language and Computation (ILLC)
University of Amsterdam
1098XH, Amsterdam, The Netherlands.
p.yang2@uva.nl

Giovanni Colavizza*

Department of Communication
University of Copenhagen
Karen Blixens Plads 8, Copenhagen, Danmark
colavizza@hum.ku.dk

February 4, 2025

ABSTRACT

Data sharing is fundamental to scientific progress, enhancing transparency, reproducibility, and innovation across disciplines. Despite its growing significance, the variability of data-sharing practices across research fields remains insufficiently understood, limiting the development of effective policies and infrastructure. This study investigates the evolving landscape of data-sharing practices, specifically focusing on the intentions behind data release, reuse, and referencing. Leveraging the PubMed open dataset, we developed a model to identify mentions of datasets in the full-text of publications. Our analysis reveals that data release is the most prevalent sharing mode, particularly in fields such as Commerce, Management, and the Creative Arts. In contrast, STEM fields, especially the Biological and Agricultural Sciences, show significantly higher rates of data reuse. However, the humanities and social sciences are slower to adopt these practices. Notably, dataset referencing remains low across most disciplines, suggesting that datasets are not yet fully recognized as research outputs. A temporal analysis highlights an acceleration in data releases after 2012, yet obstacles such as data discoverability and compatibility for reuse persist. Our findings can inform institutional and policy-level efforts to improve data-sharing practices, enhance dataset accessibility, and promote broader adoption of open science principles across research domains.

Keywords open science · data sharing · data reuse · research data

1 Introduction

Open Science, emerging from a diverse array of cultural and technological initiatives at the turn of the twenty-first century, has evolved into a transformative movement within the scientific community [Willinsky, 2005, Moore, 2017]. At its core, Open Science aims to make scientific research more transparent, accessible, and inclusive, fostering collaboration across disciplines and encouraging broader societal engagement. One influential definition characterizes Open Science as “transparent and accessible knowledge that is shared and developed through collaborative networks,” highlighting both the outputs of scientific endeavors and the processes involved in their creation [Vicente-Saez and Martinez-Fuentes, 2018, Leonelli, 2023]. Expanding on this perspective, UNESCO defines Open Science as “an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible, and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation, and communication to societal actors beyond the traditional scientific community” [Möller, 2023].

In practice, Open Science goes beyond providing open access to scientific publications; it also encompasses a wide range of activities, including the sharing of research data, software, and methodologies, all aimed at enhancing transparency and fostering collaboration [Ramachandran et al., 2021, Mauthner and Parry, 2013]. Among these practices, data

*Giovanni Colavizza is also affiliated at the University of Bologna, Department of Classical and Italian Philology, Italy

sharing has garnered significant attention. Specifically, data sharing refers to the release of data in formats that enable reuse by others [Pasquetto et al., 2017]. This practice can take many forms, ranging from private exchanges between researchers to more formal mechanisms such as depositing datasets in archives, repositories, domain-specific collections, or library collections. Additionally, researchers may share data by attaching supplemental materials to journal articles or posting datasets on laboratory websites [Wallis et al., 2013].

Data sharing practices offer substantial benefits to both the scientific community and individual researchers. For instance, the free availability of Landsat series data resulted in a twentyfold increase in downloads from the United States Geological Survey between 2009 and 2017, accompanied by a fourfold rise in its use in annual publications [Zhu et al., 2019]. This increased accessibility has advanced research applications in land monitoring, enabling studies on surface changes, coastal erosion rates, and glacier fluctuations [Kennedy et al., 2014, Roy et al., 2014, Wulder et al., 2012].

In addition to broadening research opportunities, data sharing correlates with higher citation rates. For example, Piwowar and Vision [2013] analyzed 10,555 studies utilizing gene expression microarray data and identified a 9% citation advantage for papers that shared data. This citation boost varies by discipline. In astronomy, articles linked to open datasets showed a 20% increase in citation rates [Henneken and Accomazzi, 2011], while paleoclimatology papers with publicly available data experienced a 35% citation advantage [Sears, 2011]. In the social sciences, Gleditsch et al. [2003] examined articles in the *Journal of Peace Research* and found that those providing data, regardless of format, were cited twice as often as similar articles without accessible data.

Beyond citation impacts, data sharing improves research productivity. A study of over 7,000 NSF and NIH-funded projects found that those with archived data produced a median of 10 publications, compared to only 5 for projects without archived data [Pienta et al., 2011]. Additionally, data sharing facilitates peer review and reproducibility, which are essential for verifying research findings and fostering scientific reliability [Peng, 2011].

Despite these advantages, there remain open questions on data sharing. One major challenge lies in detecting data-sharing behaviors within publications. Many studies focus on limited datasets or specific disciplines, failing to provide a comprehensive view of data-sharing practices across the scientific community [Zhao et al., 2018, Koesten et al., 2020, Khan et al., 2021, Stodden et al., 2018]. For instance, Cao et al. [2023] analyzed 1,062,586 arXiv papers in LaTeX format published between 2011 and 2021, but their study focused solely on computer science, physics, and mathematics, leaving other disciplines unexplored. Another limitation arises from the reliance on data availability statements (DAS) as the primary indicator of data sharing [Colavizza et al., 2020, Jiao et al., 2024, Strcic et al., 2022]. While useful, DAS are not universally required across fields or journals, creating substantial gaps in understanding the variations in data-sharing practices. Furthermore, as dataset reference standards evolve, mentions of datasets are no longer confined to DAS alone but may appear in other sections of publications [Cao et al., 2023].

To address these gaps, our study employs large-scale full-text analysis to investigate data sharing and reuse patterns comprehensively. We aim to answer the following questions:

- To what extent is research data released, reused, and referenced across scientific disciplines?
- How do releases, reuses, and references change across fields and over time?

Our analysis uses the PubMed Open Access (OA) collection, consisting of over 5.7 million full-text articles. To identify the datasets referenced in the publications, we relied on the repository list provided by the European Research Council (ERC) <https://zenodo.org/records/7728016> [Jahn et al., 2023]. This repository encompasses all research funded by the ERC, offering valuable insights into the availability and characteristics of data repositories across diverse research disciplines. For their work, the authors considered 220 repositories, identifying 137 trusted data repositories and 74 trusted literature repositories. For our investigation, we rely on the ERC data repositories list to detect and extract mentions to datasets in the full text of papers, as our primary emphasis lies in understanding the availability and nature of repositories across various research fields. Through natural language processing (NLP), this study categorizes data citation intent, such as release, reuse, and reference. This approach provides a nuanced understanding of data citation practices and offers an innovative methodology for analyzing data reuse intentions within scientific literature.

Our findings are expected to shed light on how research data repositories are utilized across diverse scientific fields. By providing insights into data citation patterns, this research aims to guide repository development strategies and contribute to the advancement of open science.

2 Previous Work

2.1 Open science and research data

Interest in open science has been growing steadily, with a noticeable increase in the adoption and enforcement of open science practices across disciplines. For instance, funding organizations such as the European Commission require grant recipients to comply with open-access publishing policies under frameworks like Horizon Europe, aiming to enhance the accessibility and dissemination of research outputs to broader audiences [Commission et al., 2021]. Similarly, numerous academic journals and institutions now mandate practices such as data sharing and methodological transparency as part of their publication and evaluation processes [Robson et al., 2021, Gorgolewski and Poldrack, 2016]. Moreover, open science communities play a pivotal role in facilitating the large-scale transition of researchers toward open science practices [Armeni et al., 2021].

Open science practices extend beyond open-access publishing and include the early sharing of research outputs. For example, platforms like arXiv and bioRxiv enable the dissemination of preprints, fostering early access to findings. Furthermore, open science encourages the public sharing of data and code, often hosted on online repositories such as Zenodo and GitHub, thereby improving research reproducibility and scalability. Open science also promotes rigorous and transparent research design, exemplified by practices like study preregistration [Gopal et al., 2018].

Substantial evidence indicates that open science practices offer significant advantages over traditional closed practices [McKiernan et al., 2016]. Open-access articles, for example, not only garner broader academic attention and higher citation rates [Huang et al., 2024] but also attract greater engagement from the general public and news media compared to paywalled articles [Schultz, 2021, Yang et al., 2024]. Furthermore, open science has been shown to accelerate scientific discovery in specific fields [Woelfle et al., 2011], enhance research transparency, and improve reproducibility [Besançon et al., 2021]. These benefits play a critical role in addressing challenges associated with the reproducibility crisis [Collaboration, 2015].

In today’s data-driven research landscape, the collection, analysis, and interpretation of large datasets are critical to scientific discovery. Among the pillars of open science, research data is particularly vital for promoting transparency and reproducibility. Access to well-documented research data facilitates independent verification of results, supports secondary analyses, and fosters interdisciplinary collaboration, thereby amplifying the impact of scientific inquiry [Hossain et al., 2016, Milham et al., 2018]. There is evidence that integrated data sets have been instrumental in driving biomedical discoveries and drug development [Shahin et al., 2020].

The advantages of sharing research data are far-reaching, enhancing both the visibility and reuse of research outputs while maximizing the impact of funding agencies’ investments [Los, 2010]. Recognizing these benefits, governments and funding bodies worldwide have implemented policies to incentivize open data practices. The United States pioneered such efforts as early as 1991 [Bromley, 1991], with countries like China, the United Kingdom, and Australia subsequently strengthening their data management frameworks [General Office of the State Council of the People’s Republic of China, 2018, UK, 2016, Service, 2011]. In Europe, the Horizon 2020 initiative introduced the Open Research Data Pilot (ODP) to improve data accessibility and establish credibility in data-sharing practices. Leading funding agencies, including the NSF, NIH, and the UK’s Economic and Social Research Council, now require grant applicants to submit data management plans as part of their application process [Smith, 2012, Spengler, 2012]. Publishers such as Elsevier, PLOS, Springer, and Nature have also adopted policies that encourage or mandate data citation within reference lists, promoting transparency and accountability in scientific research [Cousijn et al., 2018, Walton, 2010, PLOS One, 2019, Springer Nature, 2016].

For researchers, open data practices offer additional benefits: they facilitate the development of scientific software [Niemeyer et al., 2016], increase research productivity [McNaught, 2015], and promote a collaborative data-sharing culture within the scientific community [Belter, 2014]. By aligning incentives for researchers, funders, and publishers, these policies collectively strengthen the foundation for transparent, reproducible, and impactful research.

However, significant barriers continue to hinder the widespread adoption of open data. These include limited incentives, inconsistent citation practices, concerns about data quality, and researchers’ reluctance to relinquish control over their data. Additionally, a lack of awareness and insufficient support mechanisms exacerbate these challenges [Chawinga and Zinn, 2019, Gajbe et al., 2021]. Practical issues such as time constraints, inadequate funding, and insufficient institutional support further impede progress [Tenopir et al., 2011, 2020]. Deficiencies in archival standards and infrastructure also contribute to low rates of data sharing [Markiewicz et al., 2021]. For example, studies that sought to obtain data directly from authors reported low success rates—ranging from 27% to 59%, depending on the discipline and geographical context [Tedersoo et al., 2021]. Even among papers with data availability statements claiming “data available upon request,” compliance remains low. A 2018 study revealed that only 44% of authors shared their data

when requested [Stodden et al., 2018], a finding corroborated by subsequent research [Strcic et al., 2022, Danchev et al., 2021].

2.2 Sharing and reuse of research data

Research on data sharing and reuse remains in an exploratory stage, with scholars using various data sources and quantitative methods to analyze and discuss data reuse and sharing behaviors in publications.

Disciplinary differences in data citation practices have been a focal point. For instance, Park et al. [2018] examined samples from biological and biomedical sciences in the Data Citation Index (DCI), revealing that informal citations within article text are more prevalent than formal citations in reference sections. Similarly, Robinson-García et al. [2016] also utilized DCI data to examine the varying uses of datasets and data studies across disciplines. Their analysis found that datasets were most frequently cited in the fields of science and engineering & technology, whereas data studies played a more prominent role in the social sciences and arts & humanities. Park and Wolfram [2017] analyzed 148 articles from the Web of Science Data Citation Index to identify factors influencing data sharing and reuse. They found that formal data citation remains relatively uncommon, while informal references in the main text are more typical.

Certain factors have been found to influence researchers' willingness and ability to share and reuse datasets. Studies suggest a correlation between dataset sharing and higher citation rates [Piwowar and Vision, 2013, Piwowar et al., 2007]. Authors also tend to reuse their own shared data, resulting in higher self-citation rates [Robinson-García et al., 2016]. Data-sharing practices vary notably by discipline, suggesting a need for tailored approaches for each field [Helbig et al., 2015, Torres-Salinas et al., 2014]. Furthermore, data-sharing rates vary by scientific field [Tenopir et al., 2011], and researchers' data-sharing behaviors and perceptions differ across age groups and geographical locations [Tenopir et al., 2015]. Certain data types, such as survey, aggregated, and sequence data, receive more frequent citations and higher altmetric scores [Peters et al., 2015].

Studies of data-sharing behavior highlight the impact of shared data on research practices. For instance, an analysis of 600 articles across PLOS journals showed that 74% of studies rely on datasets created by authors, with fewer reusing prior datasets [Zhao et al., 2018]. In biodiversity research, studies using Global Biodiversity Information Facility (GBIF) data demonstrate a rise in open data use, though best practices for data citation remain underutilized [Khan et al., 2021].

In addition, journal compliance policies for data sharing have improved, with an increase in the use of repositories instead of supplementary materials for data storage [Jiao et al., 2024]. However, data availability statements (DAS) remain inconsistent, especially in COVID-19 research, where only a quarter of preprints provide explicit data-sharing statements [Strcic et al., 2022].

Despite these findings, certain gaps in data-sharing and reuse research remain, particularly in the context of cross-disciplinary data-sharing practices. Most studies are based on samples or case studies from specific fields or repositories, lacking comprehensive cross-disciplinary insights [Kafkas et al., 2015, Piwowar et al., 2011, Zhao et al., 2018, Khan et al., 2021, Cao et al., 2023]. Furthermore, the use of data availability statements to accurately identify datasets in academic publications remains limited [Jiao et al., 2024, Strcic et al., 2022]. Some studies suggest that datasets are more commonly cited informally within the text, as opposed to formal citations in references [Belter, 2014, Kafkas et al., 2015]. While some researchers use the Data Citation Index (DCI) to examine dataset usage, the DCI's focus on natural sciences results in limited coverage across disciplines [Silvello, 2017, Park et al., 2018, Park and Wolfram, 2017], with citation patterns that remain incomplete [Robinson-García et al., 2016].

One related study, Cao et al. [2023] investigated the adoption of data and method-sharing practices by analyzing a dataset of 1.1 million arXiv papers, concentrating on physics, mathematics, and computer science. They utilized regular expression matching to extract URLs from the LaTeX-formatted full text of these papers, classifying the URLs as "data URLs" or "method URLs" using manual annotation and a fine-tuned SciBERT model. Their findings highlighted a growing trend in link-sharing for methods and data, with an increasing number of papers incorporating such URLs over time. They also noted a rise in the reuse of the same links across papers, particularly in computer science, indicating a possible expansion of reproducibility efforts. Furthermore, the analysis revealed a consolidation of links within fewer web domains, such as GitHub, over time. Importantly, papers featuring shared links tended to have a higher citation impact, especially when the links remained active, underscoring the practical benefits of data-sharing practices.

While this study represents a valuable contribution by leveraging full-text analysis on a large dataset, it has notable limitations. Its focus on preprint articles and specific disciplines (physics, mathematics, and computer science) may restrict the generalizability of its findings. Preprints are not universally utilized across all academic disciplines, meaning the dataset may not adequately capture fields where preprint culture is less established. Moreover, the exclusion of formally published articles leaves unanswered questions about potential differences in data-sharing practices between

preprints and peer-reviewed publications. Considering the diverse adoption rates of data-sharing practices across scientific disciplines, expanding this research to include formally published articles and additional fields would offer a more comprehensive understanding of how data-sharing practices vary and evolve.

To deepen our understanding of data-sharing and reuse practices, further work across disciplines is essential. Our study seeks to provide a more comprehensive perspective on data use across scientific fields, filling gaps left by previous research that focused on specific disciplines or datasets. By broadening the scope of analysis, the study aims to offer practical insights into the factors influencing data-sharing practices and the variability observed across disciplines. These insights can help foster the adoption of standardized practices and promote a more widespread culture of data-sharing within the research community, ultimately enhancing collaboration, reproducibility, and the overall impact of scientific research.

3 Methodology

The data utilized in this study was obtained from the PubMed Open Access collection² as of March 2024. The total number of publications considered in this dataset is $N = 5,704,648$ (3,772,464 of oa_comm, 1,502,488 of oa_noncomm, 429,696 of oa_other)³. To enhance the dataset with additional information such as publication dates, citation counts, and disciplines, we queried the Dimensions API (March 2024).

To extract the relevant data repositories mentioned in each paper, we implemented a series of processing steps.

Firstly, we employed regular expression (regex) matching to identify repositories from the full text based on their URLs. Our approach involved applying a unified rule across all URLs to strip away the protocol (http, https) and subdomain (www). For example, from the URL 'https://meertens.knaw.nl/en/collections/', we retained 'meertens.knaw.nl/en/collections'. Preliminary evidence suggests that this approach enhances resource availability compared to relying solely on data availability statements [Federer, 2022, Cao et al., 2023]. The comprehensive list of repository links is available in our repository⁴. Secondly, to ensure consistency in the repository names or URLs, we converted the entire text of the paper and our domain list to lowercase during the matching process.

After matching, we successfully extracted 69,090 articles (1.2%) from the PubMed Open Access collection dataset that included at least one repository link. Figure 1 illustrates the top 10 repositories appearing in the dataset, ranked by frequency.

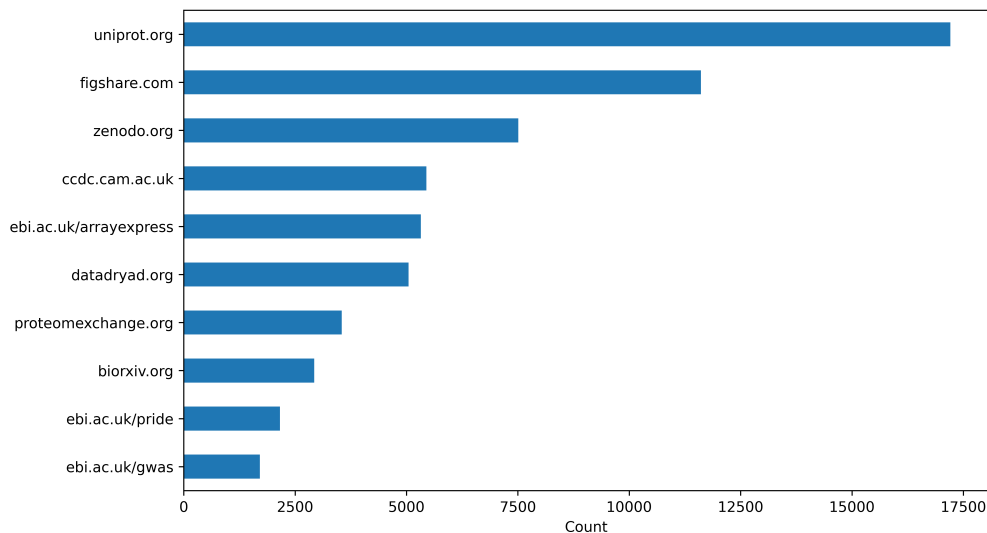


Figure 1: Top 10 repositories by frequency.

²<https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>

³Commercial Use Allowed (oa_comm): CC0, CC BY, CC BY-SA, and CC BY-ND licenses; Non-Commercial Use Only (oa_noncomm): CC BY-NC, CC BY-NC-SA, CC BY-NC-ND; Other (oa_other): no machine-readable license, no license, or a custom license. <https://pmc.ncbi.nlm.nih.gov/tools/ftp/>

⁴https://github.com/alsowbdxa/Research_Data_in_Scientific_Publications/blob/main/Codes/dataset_urls.xlsx

Subsequently, these 69,090 papers served as the foundation for constructing our annotation dataset used in model training. Based on our observations and experience, we found that the intention of usage of the repository is typically conveyed within the core sentence (the sentence containing the repository link) and one or two sentences around it. To maximize the information captured while maintaining simplicity, we adopted the same method as a previous study, which incorporates the sentences immediately preceding and following the core sentence to provide additional context [Zhang et al., 2022]. Specifically, we retained the two sentences preceding and the two sentences following the core sentence as the context. This context, often regarded as crucial, provides additional information to better identify and classify the intended use of the mentioned datasets [Koesten et al., 2020]. Using this method, we processed all the papers and randomly selected 1,000 articles (1.4% of the 69,090 papers) for annotation.

The manual annotation of these contexts was carried out using Label Studio. We classified the intention of contexts into four categories:

- **Release** The context of the repository mentioned indicates that the paper releases a new dataset on the repository, or generates a dataset by integrating diverse published datasets and releasing it on the repository. A repository ID is typically provided alongside the mention or in the paper.

Example 1: *The datasets are currently for private access during this review period, which can be accessed through: https://datadryad.org/stash/share/yRDf1KmJ9_hR_IIGg_vukBVNUmmB9tm_j8v1BZ721A.*

Example 2: *Raw microarray data have been deposited in compliance to MIAME guidelines at ArrayExpress database (<http://www.ebi.ac.uk/arrayexpress>), with accession number E-TABM-1215 release date June 11, 2012. Gene subsets corresponding to each combination of responses analyzed by microarrays were defined from Venn diagrams indicating the number of the included genes.*

- **Reuse** The context surrounding the repository mentioned reveals that the paper directly employed a published dataset hosted on the repository. A repository ID is typically provided alongside the mention or in the paper.

Example 1: *The European Commission do not accept any responsibility for use that may be made of the information it contains. All data used in the current study is publicly available. Summary statistics for IBS can be download from European Bioinformatics Institute GWAS Catalog (<https://www.ebi.ac.uk/gwas/>). Summary statistics for neuroticism can be downloaded from https://ctg.cncr.nl/software/summary_statistics and <http://www.ccace.ed.ac.uk>. Summary statistics for depression can be downloaded from <https://datashare.ed.ac.uk/handle/10283/3203>.*

Example 2: *The land use data were obtained from the 30-m annual land cover datasets and its dynamics in China from 1990 to 2020 (<https://zenodo.org/record/5210928#.Y9TDU3ZBxD>).*

- **Reference** The context surrounding the repository mentioned indicates that the paper references the repository, possibly to compare different datasets or for context. Importantly, the authors' work is not reliant on this dataset, nor have they produced a new dataset based on it.

Example 1: *Furthermore, in Table S3, we also list the top 20 ranked potential phosphorylation sites for MAPK1, in which Tyr325 and Tyr331 of FOS (P01100) has been confirmed to be modified by this kinase (http://www.uniprot.org/uniprot/P01100#ptm_processing).*

Example 2: *Some of the resources used an ontology, e.g., Disease Ontology, a taxonomy such as MeSH [24], or cross-referenced another resource such as OMIM. Diseases and phenotypes are often mixed in the same resource and sometimes in the same category annotation. For example, the European Variation Archive (EVA <http://www.ebi.ac.uk/eva/>) [25] trait names' labeling uses a mixed set of vocabularies from HP, SNOMED-CT, OMIM, and non-standardized local identifiers used internally at source from the ClinVar records. The identifiers of the record's cross-references for each trait name are not equivalently represented - e.g., trait name 'congenital adrenal hyperplasia' in EVA contains identifiers for SNOMED-CT, HP, but not for OMIM. This trait name also links to a non-standardized internal identifier used at the Office of Rare Disease.*

- **Nothing** Occasionally, we encountered erroneous or non-related hits. While using repository links to identify repositories within the full text, we found that sometimes these links did not solely indicate repositories but could also convey other meanings. In such cases, we labelled it as 'Nothing.'

Example 1: *The following link will take you to the Dryad record for your article, so you won't have to re-enter its bibliographic information, and can upload your files directly: <http://datadryad.org>.*

org/submit?journalID=pgenetics&manu=PGENETICS-D-19-01831R2 More information about depositing data in Dryad is available at <http://www.datadryad.org/depositing>.

Example 2: *Assessing the impact of autologous neutralizing antibodies on viral rebound in postnatally SHIV-infected ART-treated infant rhesus macaques 14 9 2023 2023.07.22.550159 <http://biorxiv.org/lookup/doi/10.1101/2023.07.22.550159> Abstract While the benefits of early antiretroviral therapy (ART) initiation in perinatally infected infants are well documented, early ART initiation is not always possible in postnatal pediatric HIV infections, which account for the majority of pediatric HIV cases worldwide. The timing of onset of ART initiation is likely to affect the size of the latent viral reservoir established, as well as the development of adaptive immune responses, such as the generation of neutralizing antibody responses against the virus.*

Following the definition of these four intentions, we manually annotate the annotation subset (1,000 articles with 1328 contexts), and we get 670 contexts with the label ‘Release,’ 119 contexts with ‘Reference,’ 453 contexts with ‘Reuse’ and 86 contexts with ‘Nothing.’ Then we use this subset to train the model.

For model training, we utilized pre-trained models from Hugging Face, specifically BertForSequenceClassification ‘bert-base-uncased’ for BERT and RobertaForSequenceClassification ‘roberta-base’ for RoBERTa, this model has been validated as delivering optimal performance in most NLP tasks [Devlin, 2018]. Before training, we mapped the original labels to distinct integers, assigning ‘Release’ as label 0, ‘Reuse’ as label 1, ‘Reference’ as label 2, and ‘Nothing’ as label 3. The dataset was then partitioned into training, testing, and validation subsets in an 80-10-10% split.

To prepare the textual data for modeling, we performed tokenization using BERT and RoBERTa tokenizers respectively, considering a maximum sequence length of 512 tokens, for each sentence. If the total number of tokens is less than 512 (the model’s maximum limit), the entire sentence is retained. However, if it exceeds 512 tokens, we employ four different truncation methods:

- Method 1 If it exceeds 512 tokens, we truncate it by retaining the first 512 tokens.
- Method 2 If it exceeds 512 tokens, we truncate it by preserving the last 512 tokens.
- Method 3 If it exceeds 512 tokens, we truncate it by keeping the central 512 tokens. For instance, if it has 1000 tokens, we remove the first 244 and last 244 tokens.
- Method 4 If it exceeds 512 tokens, we truncate it by keeping the first 256 tokens and the last 256 tokens.

For each model, we employ each truncation method and evaluate the model based on the F1 score. Additionally, we incorporate the early stopping mechanism in the training process. Specifically, if the F1 score on the validation set shows no improvement over 10 epochs, and the model’s performance starts to degrade, we terminate the training

We present an overview of the performance results achieved by various methods when applied to either the RoBERTa and BERT models¹. We see that RoBERTa, specifically when employed with method 2, outperforms the other configurations, boasting a F1 score of 0.902. Building upon this outcome, we further enhance the model’s efficacy by merging the training and test subsets. Leveraging the RoBERTa model in conjunction with truncation method 2, we conduct fine-tuning to optimize its performance. The resultant refined model is subsequently subjected to testing on the validation dataset.

		Accuracy	Precision	Recall	F1
RoBERTa	Method 1	0.896	0.897	0.896	0.894
	Method 2	0.902	0.902	0.902	0.902
	Method 3	0.886	0.885	0.886	0.885
	Method 4	0.885	0.890	0.890	0.886
BERT	Method 1	0.876	0.882	0.876	0.876
	Method 2	0.876	0.874	0.876	0.874
	Method 3	0.870	0.873	0.870	0.871
	Method 4	0.855	0.864	0.855	0.857

Table 1: Performance of models by methods

Following the training phase, we deploy the trained model to predict the intention for each context within the entire dataset (69,090 articles with 92,267 contexts). The distribution of intentions across the entire dataset is illustrated in Figure 2.

Specifically, we observe that 55,680 contexts (60.3%) are labeled as ‘Release,’ 24,809 contexts (26.9%) as ‘Reuse,’ 8,597 contexts (9.3%) as ‘Reference,’ and 3,181 contexts (3.5%) as ‘Nothing.’

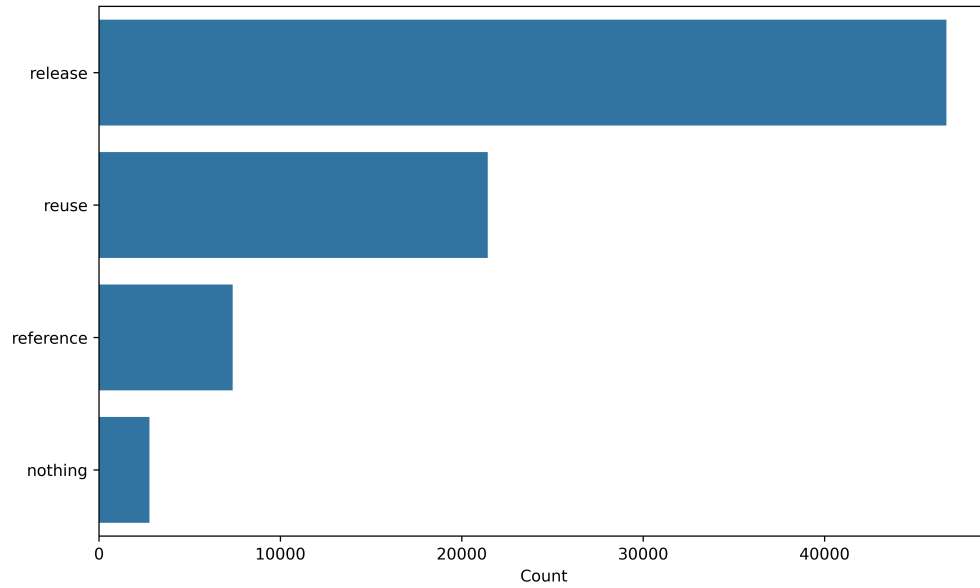


Figure 2: Predicted label distribution

We visualize the top 20 repositories along with their respective usage intentions in the figure below:

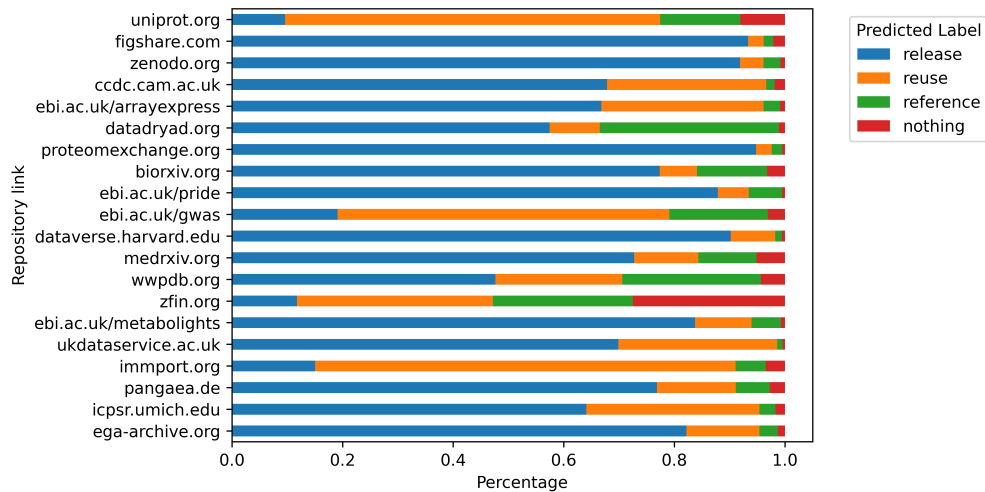


Figure 3: Distribution of intentions by repository.

The patterns of using intentions across repositories are highlighted in Figure 3. The figure further validates the accuracy of our classification to some extent. Repositories such as Figshare and Zenodo, often utilized for publishing datasets,

demonstrate a higher frequency of the ‘Release’ type. Conversely, repositories like Uniport⁵ and ebi.ac.uk⁶, dedicated to supplying datasets for research analysis, display a predilection for the ‘Reuse’ type.

4 Results

We start by providing a description of our findings, followed by an in-depth discussion in the subsequent section.

Figure 4 presents the proportional distribution of three intentions across various academic disciplines. The horizontal bar plot includes 22 disciplines, each represented on the y-axis, while the x-axis shows the proportion of each intention relative to the total papers for that field, which is calculated based on fractional counting.

In most disciplines, the ‘release’ intention dominates, indicating a strong preference for openly sharing data. Specifically, the top five disciplines for releasing datasets are ‘Commerce, Management, Tourism and Services’, ‘Studies in Creative Arts and Writing’, ‘Studies in Human Society’, ‘Psychology and Cognitive Sciences’ and ‘Economics’. Conversely, the proportions of released datasets are lowest in ‘Biological Sciences’, ‘Information and Computing Sciences’ and ‘Agricultural and Veterinary Sciences’.

Regarding the intention of reuse, STEM-related fields generally exhibit a higher proportion of reuse. Notably, ‘Agricultural and Veterinary Sciences’, ‘Technology’, ‘Chemical Sciences’, ‘Biological Sciences’, ‘Medical and Health Sciences’ have over 30% of mentions indicating dataset reuse.

For reference intention, datasets are referenced less frequently across all disciplines, with two exceptions: ‘Information and Computing Sciences’ and ‘Philosophy and Religious Studies’.

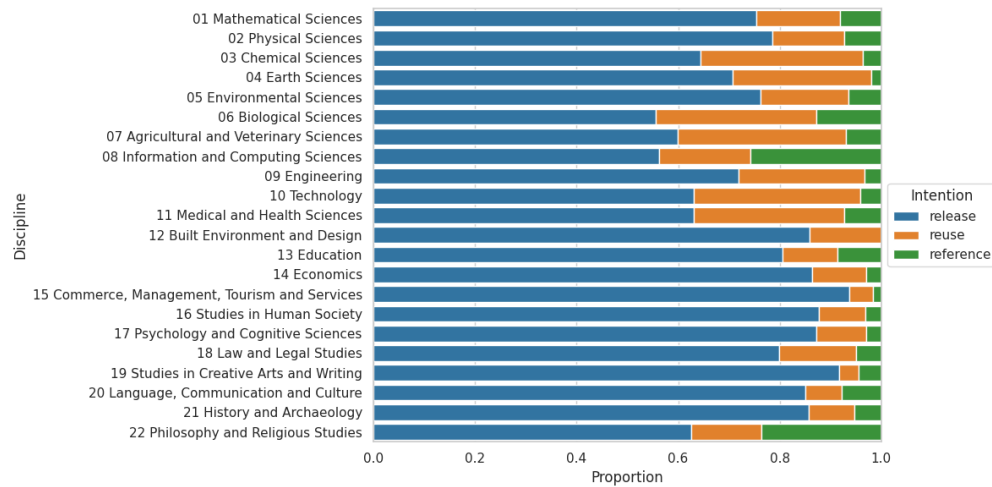


Figure 4: Distribution of intentions across disciplines.

Furthermore, figure 5 shows the distribution of different mention intentions in the dataset (‘reuse,’ ‘release,’ and ‘reference’) over time, with the x-axis representing the years and the y-axis showing the percentage of each intention.

The figure reveals trends in dataset usage across publications. From 2007 to 2012, the intention to reuse datasets increased, while the intention to release datasets remained relatively low and even declined slightly. However, starting in 2012, the trends shifted. The intention to release datasets sharply increased and consistently remained high (around 60%), while the intention to reuse datasets decreased significantly, dropping from 50% to approximately 30%.

⁵A prominent free-access collection of protein sequences and their annotations, supporting fields like biology, medicine, and biotechnology. <https://pmc.ncbi.nlm.nih.gov/articles/PMC4384041/>

⁶The world’s most comprehensive suite of freely available data resources and tools for life science research. <https://www.ebi.ac.uk/about>

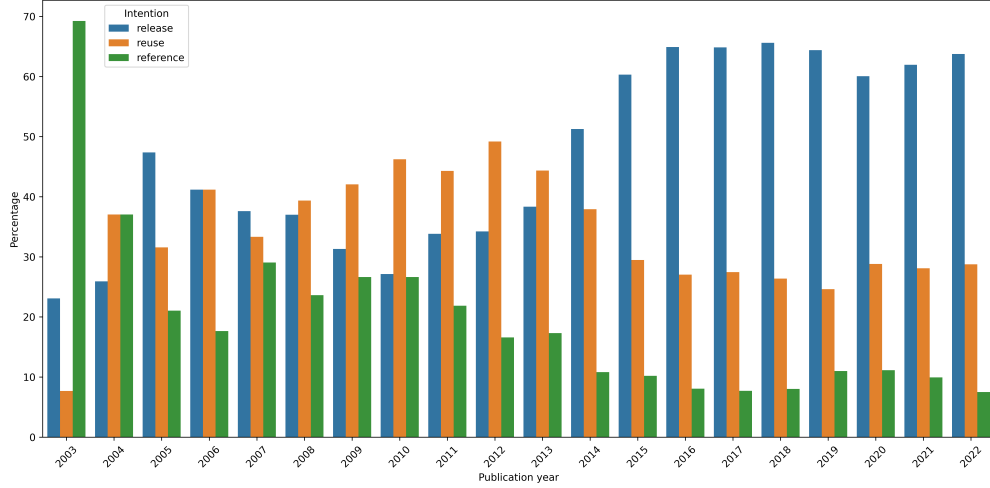


Figure 5: Distribution of intention over time.

To explore the disciplinary interconnections in the contexts of release, reuse, and reference intentions, we constructed co-occurrence networks for each intention via VOSviewer. Specifically, a publication may encompass multiple disciplines, the co-occurrence of macro disciplines associated with that publication is treated as a co-occurrence instance. The co-occurrence distribution for each intention network is subsequently calculated using fractional counting. In these networks, nodes represent disciplines, edges represent the co-occurrence of two disciplines within the same article, and the color means the cluster identified based the VOSviewer[Van Eck and Waltman, 2010]. The strength of connections is determined by the frequency of such co-occurrences.

The co-occurrence network for the release intention (Figure 6) underscores the pivotal role of Biological Sciences and Medical and Health Sciences, which form strong connections with related disciplines such as Agricultural and Veterinary Sciences, Environmental Sciences, and Chemical Sciences.

For the reuse intention (Figure 7), the network displays a more dispersed pattern of connections. While Biological Sciences and Medical and Health Sciences remain central, the network highlights strong links with Information and Computing Sciences and Mathematical Sciences, illustrating the growing importance of computational and data-driven approaches in reused research. Additionally, significant connections to social science disciplines, such as Studies in Human Society and Education, suggest that data reuse is becoming increasingly relevant across diverse academic contexts.

The reference intention network (Figure 8) exhibits a distinct structure, with Biological Sciences maintaining a central role but with more dispersed connections compared to the other two intentions.

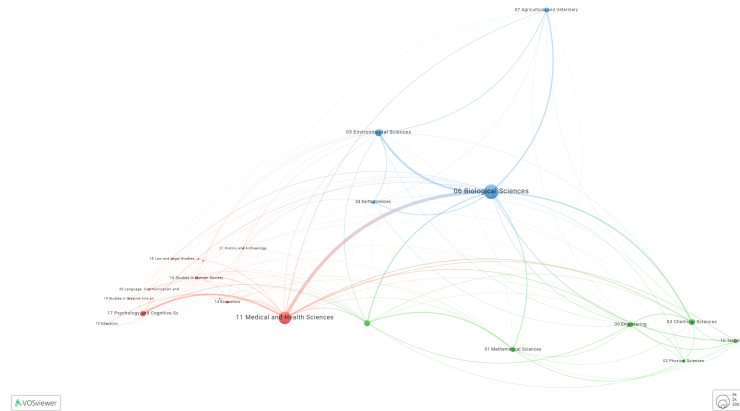


Figure 6: Co-occurrence network of release

to ensure compliance. Thus, while the overall trend supports open science initiatives, significant variation persists across disciplines due to these domain-specific barriers and norms.

In contrast, the reuse intention is more prevalent in STEM-related fields such as ‘Agricultural and Veterinary Sciences’, ‘Technology’, ‘Chemical Sciences’, ‘Biological Sciences’, ‘Medical and Health Sciences’, where over 30% of mentions involve the reuse of existing datasets. This prevalence can be attributed to the availability of shared databases, such as UniProt and CCDC, and the methodological reliance on pre-existing data in these disciplines [Jiménez-Contreras et al., 2015]. However, the relatively low reuse proportions in humanities and social sciences suggest that data reuse practices are less institutionalized in these fields. This discrepancy likely arises from variations in data availability, research methodologies, and the perceived value of reusing datasets [Kim and Yoon, 2017].

The reference intention remains consistently low across most disciplines, with notable exceptions in ‘Information and Computing Sciences’ and ‘Philosophy and Religious Studies’. This overall low level of dataset referencing highlights a critical issue in academic publishing: datasets are not yet widely recognized as formal research outputs in many fields [Silvello, 2017]. While dataset citation practices are gaining traction in Information and Computing Sciences [Force, 2014], other disciplines lag behind due to a lack of standardized citation practices and limited awareness of the benefits of dataset citation [Kratz and Strasser, 2014]. As previous studies suggest, data citation not only provides credit to data creators but also enhances transparency and reproducibility, underscoring its significance in advancing open science [Altman et al., 2015, Piwowar and Vision, 2013].

In addition, Biological Sciences and Medical and Health Sciences play a pivotal role in the connection among disciplines, which consistently serve as hubs across all intentions. These fields highlight the inherently interdisciplinary nature of research, particularly in data release practices, as evidenced by strong connections within the life and health sciences cluster. This observation aligns with previous findings that 86% of research data published in biological sciences journals are cited by articles from disciplines outside the biological sciences domain in the Web of Science [Park, 2022].

5.2 Temporal trends in releases, reuses, and references

Temporal trends in data-sharing practices show a clear evolution. From 2007 to 2012, the rise in reuse intentions marked the early stages of data-sharing adoption, driven by large-scale repositories and a growing emphasis on data-driven research [Tenopir et al., 2015]. The post-2012 surge in data release intentions may align with global open science initiatives, such as the U.S. White House memorandum in 2013, which required federal agencies to increase public access to research results, and the Plan S initiative in 2018, which set standards for immediate open access publication across Europe [Holdren et al., 2013, Schiltz, 2018]. These policies have had a substantial societal impact, making millions of academic publications freely accessible to the public and fostering a shift toward collaborative, open science.

Since 2018, all disciplines have seen significant growth in data release activity, with STEM fields showing steady growth in reuse and reference intentions. This signals the increasing normalization of data-driven research practices. However, the delayed adoption of these practices in the humanities suggests ongoing cultural and infrastructural shifts, compounded by challenges such as non-standardized data formats and discipline-specific attitudes toward open science [Führ and Bisset Alvarez, 2021].

Despite the success of open access policies, the decline in reuse intentions highlights ongoing challenges in data discoverability, compatibility, and a lack of incentives for reuse. Issues such as insufficient metadata, unclear licensing terms, and the technical complexity of integrating datasets may continue to hinder effective reuse in new research contexts [Borgman, 2012, Mayernik, 2017].

We acknowledge certain limitations in our study. First, while we attempted to match dataset mentions in the full text with the repository list provided by the European Research Council, some dataset mentions may have been missed. These could include instances where datasets were not associated with URLs or were not included in the repository list. Additionally, our analysis relies on full-text data, and although we worked with a substantial corpus of millions of publications across disciplines, it should be noted that the dataset mostly represents biomedical and life sciences journal literature. As a result, our findings may not fully capture data-sharing practices in fields where such datasets are less prevalent or not yet widely integrated into the research ecosystem.

6 Conclusion

This study highlights the evolving landscape of data-sharing practices across disciplines, focusing on the intentions of data release, reuse, and reference. The findings indicate that the release intention is the dominant mode of data sharing, with notable variations across disciplines. Fields such as Commerce, Management, and Creative Arts show

high levels of data release, reflecting a strong alignment with open science principles, while disciplines like Biological Sciences and Agricultural Sciences face unique challenges related to ethical, legal, and privacy concerns. The reuse intention is particularly prevalent in STEM-related disciplines, emphasizing the growing reliance on shared datasets and computational methodologies in research. However, humanities and social sciences show a delayed adoption of data reuse practices, likely due to factors like limited data availability and infrastructure.

The analysis also reveals a low proportion of dataset referencing across most fields, suggesting that datasets are not yet fully recognized as formal research outputs, despite their increasing role in the research process. Temporal trends indicate that recent open science initiatives have accelerated data release practices, particularly post-2012, yet challenges persist, especially in terms of data discoverability and compatibility for reuse. The findings further highlight the central role of the Biological and Medical Sciences in fostering interdisciplinary data sharing.

This study provides a comprehensive understanding of how data is utilized across different scientific disciplines and offers valuable insights to help institutions and publishers develop better data policies. By identifying trends in data release, reuse, and citation, it can inform strategies to enhance data sharing practices and improve the accessibility and discoverability of datasets. These findings will assist in creating more effective support systems for researchers and encourage broader adoption of open science practices across various fields.

References

- John Willinsky. The unacknowledged convergence of open source, open access, and open science. *First Monday*, 2005.
- Samuel A Moore. A genealogy of open access: negotiations between openness and access to research. *Revue française des sciences de l'information et de la communication*, 11(2), 2017.
- Ruben Vicente-Saez and Clara Martinez-Fuentes. Open science now: A systematic literature review for an integrated definition. *Journal of business research*, 88:428–436, 2018.
- Sabina Leonelli. *Philosophy of open science*. Cambridge University Press, 2023.
- Lutz Möller. Unesco recommendation on open science. In *66. Helmholtz Open Science online seminar*, 2023.
- Rahul Ramachandran, Kaylin Bugbee, and Kevin Murphy. From Open Data to Open Science. *Earth and Space Science*, 8(5):e2020EA001562, May 2021. ISSN 2333-5084. doi:10.1029/2020EA001562. URL <https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2020EA001562>.
- Natasha Susan Mauthner and Odette Parry. Open Access Digital Data Sharing: Principles, Policies and Practices. *Social Epistemology*, 27(1):47–67, January 2013. ISSN 0269-1728. doi:10.1080/02691728.2012.760663. URL <http://www.tandfonline.com/doi/abs/10.1080/02691728.2012.760663>.
- Irene V. Pasquetto, Bernadette M. Randles, and Christine L. Borgman. On the Reuse of Scientific Data. *Data Science Journal*, 16:8, March 2017. ISSN 1683-1470. doi:10.5334/dsj-2017-008. URL <https://datascience.codata.org/articles/10.5334/dsj-2017-008>.
- Jillian C. Wallis, Elizabeth Rolando, and Christine L. Borgman. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLOS ONE*, 8(7):e67332, July 2013. ISSN 1932-6203. doi:10.1371/journal.pone.0067332. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0067332>.
- Zhe Zhu, Michael A Wulder, David P Roy, Curtis E Woodcock, Matthew C Hansen, Volker C Radeloff, Sean P Healey, Crystal Schaaf, Patrick Hostert, Peter Strobl, et al. Benefits of the free and open landsat data policy. *Remote Sensing of Environment*, 224:382–385, 2019.
- Robert E Kennedy, Serge Andréfouët, Warren B Cohen, Cristina Gómez, Patrick Griffiths, Martin Hais, Sean P Healey, Eileen H Helmer, Patrick Hostert, Mitchell B Lyons, et al. Bringing an ecological view of change to landsat-based remote sensing. *Frontiers in Ecology and the Environment*, 12(6):339–346, 2014.
- David P Roy, Michael A Wulder, Thomas R Loveland, Curtis E Woodcock, Richard G Allen, Martha C Anderson, Dennis Helder, James R Irons, David M Johnson, Robert Kennedy, et al. Landsat-8: Science and product vision for terrestrial global change research. *Remote sensing of Environment*, 145:154–172, 2014.
- Michael A Wulder, Jeffrey G Masek, Warren B Cohen, Thomas R Loveland, and Curtis E Woodcock. Opening the archive: How free data has enabled the science and monitoring promise of landsat. *Remote Sensing of Environment*, 122:2–10, 2012.
- Heather A. Piwowar and Todd J. Vision. Data reuse and the open data citation advantage. *PeerJ*, 1:e175, October 2013. ISSN 2167-8359. doi:10.7717/peerj.175. URL <https://peerj.com/articles/175>.

- Edwin A Henneken and Alberto Accomazzi. Linking to data-effect on citation rates in astronomy. *arXiv preprint arXiv:1111.3618*, 2011.
- JR Sears. Data sharing effect on article citation rate in paleoceanography. *EOS, Transactions, American Geophysical Union*, 92(53):IN53B–1628, 2011.
- Nils Petter Gleditsch, Claire Metelits, and Havard Strand. Posting your data: Will you be scooped or will you be famous. *International Studies Perspectives*, 4(1):89–97, 2003.
- Amy Pienta, George Alter, and Jared Lyle. The enduring value of social science research. In *8 th International Conference on Preservation of Digital Objects*, page 215, 2011.
- Roger D. Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, December 2011. ISSN 0036-8075. doi:10.1126/science.1213847. URL <https://www.science.org/doi/10.1126/science.1213847>.
- Mengnan Zhao, Erjia Yan, and Kai Li. Data set mentions and citations: A content analysis of full-text publications. *Journal of the Association for Information Science and Technology*, 69(1):32–46, 2018. ISSN 2330-1643. doi:10.1002/asi.23919. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23919>.
- Laura Koesten, Pavlos Vougiouklis, Elena Simperl, and Paul Groth. Dataset Reuse: Toward Translating Principles to Practice. *Patterns*, 1(8), November 2020. ISSN 2666-3899. doi:10.1016/j.patter.2020.100136. URL [https://www.cell.com/patterns/abstract/S2666-3899\(20\)30184-7](https://www.cell.com/patterns/abstract/S2666-3899(20)30184-7).
- Nushrat Khan, Mike Thelwall, and Kayvan Kousha. Measuring the impact of biodiversity datasets: data reuse, citations and altmetrics. *Scientometrics*, 126(4):3621–3639, February 2021. ISSN 0138-9130. doi:10.1007/s11192-021-03890-6. URL <https://link.springer.com/10.1007/s11192-021-03890-6>.
- Victoria Stodden, Jennifer Seiler, and Zhaokun Ma. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 115(11):2584–2589, March 2018. ISSN 0027-8424. doi:10.1073/pnas.1708290115. URL <https://pnas.org/doi/full/10.1073/pnas.1708290115>.
- Hancheng Cao, Jesse Dodge, Kyle Lo, Daniel A. McFarland, and Lucy Lu Wang. The Rise of Open Science: Tracking the Evolution and Perceived Value of Data and Methods Link-Sharing Practices, October 2023. URL <http://arxiv.org/abs/2310.03193>.
- Giovanni Colavizza, Iain Hrynaszkiewicz, Isla Staden, Kirstie Whitaker, and Barbara McGillivray. The citation advantage of linking publications to research data. *PLOS ONE*, 15(4):e0230416, April 2020. ISSN 1932-6203. doi:10.1371/journal.pone.0230416. URL <https://dx.plos.org/10.1371/journal.pone.0230416>.
- Chenyue Jiao, Kai Li, and Zhichao Fang. Data sharing practices across knowledge domains: A dynamic examination of data availability statements in PLOS ONE publications. *Journal of Information Science*, 50(3):673–689, June 2024. ISSN 0165-5515. doi:10.1177/01655515221101830. URL <https://doi.org/10.1177/01655515221101830>.
- Josip Strcic, Antonia Civiljak, Terezija Glozinic, Rafael Leite Pacheco, Tonci Brkovic, and Livia Puljak. Open data and data sharing in articles about COVID-19 published in preprint servers medRxiv and bioRxiv. *Scientometrics*, 127(5):2791–2802, March 2022. ISSN 0138-9130. doi:10.1007/s11192-022-04346-1. URL <https://link.springer.com/10.1007/s11192-022-04346-1>.
- Najko Jahn, Mikael Laakso, Emma Lazzeri, and Peter McQuilton. Study on the readiness of research data and literature repositories to facilitate compliance with the Open Science Horizon Europe MGA requirements, March 2023. URL <https://doi.org/10.5281/zenodo.7728016>.
- European Commission, Directorate-General for Research, and Innovation. *Horizon Europe, open science : early knowledge and data sharing, and open collaboration*. Publications Office of the European Union, 2021. doi:doi/10.2777/18252.
- Samuel G Robson, Myriam A Baum, Jennifer L Beaudry, Julia Beitner, Hilmar Brohmer, Jason M Chin, Katarzyna Jasko, Chrystyna D Kourou, Ruben E Laukkanen, David Moreau, et al. Promoting open science: a holistic approach to changing behaviour. *Collabra: Psychology*, 7(1):30137, 2021.
- Krzysztof J Gorgolewski and Russell A Poldrack. A practical guide for improving transparency and reproducibility in neuroimaging research. *PLoS biology*, 14(7):e1002506, 2016.
- Kristijan Armeni, Loek Brinkman, Rickard Carlsson, Anita Eerland, Rianne Fijten, Robin Fondberg, Vera E Heininga, Stephan Heunis, Wei Qi Koh, Maurits Masselink, et al. Towards wide-scale adoption of open science practices: The role of open science communities. *Science and Public Policy*, 48(5):605–611, 2021.
- Anand D Gopal, Joshua D Wallach, Jenerius A Aminawung, Gregg Gonsalves, Rafael Dal-Ré, Jennifer E Miller, and Joseph S Ross. Adherence to the international committee of medical journal editors’(icmje) prospective registration policy and implications for outcome integrity: a cross-sectional analysis of trials published in high-impact specialty society journals. *Trials*, 19:1–13, 2018.

- Erin C McKiernan, Philip E Bourne, C Titus Brown, Stuart Buck, Amye Kenall, Jennifer Lin, Damon McDougall, Brian A Nosek, Karthik Ram, Courtney K Soderberg, Jeffrey R Spies, Kaitlin Thaney, Andrew Updegrove, Kara H Woo, and Tal Yarkoni. How open science helps researchers succeed. *eLife*, 5:e16800, July 2016. ISSN 2050-084X. doi:10.7554/eLife.16800. URL <https://elifesciences.org/articles/16800>.
- Chun-Kai Huang, Cameron Neylon, Lucy Montgomery, Richard Hosking, James P Diprose, Rebecca N Handcock, and Katie Wilson. Open access research outputs receive more diverse citations. *Scientometrics*, 129(2):825–845, 2024.
- Teresa Schultz. All the research that’s fit to print: Open access and the news media. *Quantitative Science Studies*, 2(3): 828–844, 2021.
- Puyu Yang, Ahad Shoaib, Robert West, and Giovanni Colavizza. Open access improves the dissemination of science: insights from wikipedia. *Scientometrics*, pages 1–24, 2024.
- Michael Woelfle, Piero Olliaro, and Matthew H Todd. Open science is a research accelerator. *Nature chemistry*, 3(10): 745–748, 2011.
- Lonni Besançon, Nathan Peiffer-Smadja, Corentin Segalas, Haiting Jiang, Paola Masuzzo, Cooper Smout, Eric Billy, Maxime Deforet, and Clémence Leyrat. Open science saves lives: lessons from the covid-19 pandemic. *BMC Medical Research Methodology*, 21(1):117, 2021.
- Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.
- Mohammad Alamgir Hossain, Yogesh K Dwivedi, and Nripendra P Rana. State-of-the-art in open data research: Insights from existing literature and a research agenda. *Journal of organizational computing and electronic commerce*, 26(1-2):14–40, 2016.
- Michael P. Milham, R. Cameron Craddock, Jake J. Son, Michael Fleischmann, Jon Clucas, Helen Xu, Bonhwang Koo, Anirudh Krishnakumar, Bharat B. Biswal, F. Xavier Castellanos, Stan Colcombe, Adriana Di Martino, Xi-Nian Zuo, and Arno Klein. Assessment of the impact of shared brain imaging data on the scientific literature. *Nature Communications*, 9(1):2818, July 2018. ISSN 2041-1723. doi:10.1038/s41467-018-04976-1. URL <https://www.nature.com/articles/s41467-018-04976-1>.
- Mohamed H Shahin, Sanchita Bhattacharya, Diego Silva, Sarah Kim, Jackson Burton, Jagdeep Podichetty, Klaus Romero, and Daniela J Conrado. Open data revolution in clinical research: opportunities and challenges. *Clinical and Translational Science*, 13(4):665–674, 2020.
- Wouter Los. *Riding the wave How Europe can gain from the rising tide of scientific data Final report of the High Level Expert Group on Scientific Data A submission to the European Commission*. European Union, January 2010.
- Allan Bromley. Policy Statements on Data Management for Global Change Research, February 1991. URL <https://digital.library.unt.edu/ark:/67531/metadc11862/>.
- General Office of the State Council of the People’s Republic of China. Notification by the general office of the state council on the issuance of scientific data management practices., 2018. URL https://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.
- Research Councils UK. Concordat on open research data, 2016. URL <https://www.ukri.org/wp-content/uploads/2020/10/UKRI-020920-ConcordatOnOpenResearchData.pdf>.
- Australian National Data Service. Outline of a research data management policy for australian universities / institutions, 2011. URL <https://alliancecan.ca/sites/default/files/2022-03/institutional-research-data-management-policies.pdf>.
- Mackenzie Smith. Institutional perspectives on credit systems for research data. In *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop*, pages 77–80, 2012.
- S Spengler. Data citation and attribution: A funder’s perspective. In *For attribution: Developing scientific data attribution and citation practices and standards: Summary of an international workshop*, pages 177–188, 2012.
- Helena Cousijn, Amye Kenall, Emma Ganley, Melissa Harrison, David Kernohan, Thomas Lemberger, Fiona Murphy, Patrick Polischuk, Simone Taylor, Maryann Martone, and Tim Clark. A data citation roadmap for scientific publishers. *Scientific Data*, 5(1):180259, November 2018. ISSN 2052-4463. doi:10.1038/sdata.2018.259. URL <https://www.nature.com/articles/sdata2018259>.
- David W. H. Walton. Data Citation - Moving to New Norms. *Antarctic Science*, 22(4):333–333, August 2010. ISSN 0954-1020. doi:10.1017/S0954102010000520. URL https://www.cambridge.org/core/product/identifier/S0954102010000520/type/journal_article.
- PLOS One. Plos one data availability, December 2019. URL <https://journals.plos.org/plosone/s/data-availability>.

- Springer Nature. Research data policy, 2016. URL <https://www.springernature.com/gp/authors/research-data-policy>.
- Kyle E. Niemeyer, Arfon M. Smith, and Daniel S. Katz. The Challenge and Promise of Software Citation for Credit, Identification, Discovery, and Reuse. *Journal of Data and Information Quality*, 7(4):1–5, October 2016. ISSN 1936-1955. doi:10.1145/2968452. URL <https://dl.acm.org/doi/10.1145/2968452>.
- Keith McNaught. The Changing Publication Practices in Academia: Inherent Uses and Issues in Open Access and Online Publishing and the Rise of Fraudulent Publications. *The Journal of Electronic Publishing*, 18(3), June 2015. ISSN 1080-2711. doi:10.3998/3336451.0018.308. URL <http://hdl.handle.net/2027/spo.3336451.0018.308>.
- Christopher W. Belter. Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets. *PLoS ONE*, 9(3):e92590, March 2014. ISSN 1932-6203. doi:10.1371/journal.pone.0092590. URL <https://dx.plos.org/10.1371/journal.pone.0092590>.
- Winner Dominic Chawinga and Sandy Zinn. Global perspectives of research data sharing: A systematic literature review. *Library & Information Science Research*, 41(2):109–122, April 2019. ISSN 0740-8188. doi:10.1016/j.lisr.2019.04.004. URL <https://www.sciencedirect.com/science/article/pii/S074081881830330X>.
- Sagar Bhimrao Gajbe, Amit Tiwari, Gopalji, and Ranjeet Kumar Singh. Evaluation and analysis of Data Management Plan tools: A parametric approach. *Information Processing & Management*, 58(3):102480, May 2021. ISSN 0306-4573. doi:10.1016/j.ipm.2020.102480. URL <https://linkinghub.elsevier.com/retrieve/pii/S0306457320309699>.
- Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE*, 6(6):e21101, June 2011. ISSN 1932-6203. doi:10.1371/journal.pone.0021101. URL <https://dx.plos.org/10.1371/journal.pone.0021101>.
- Carol Tenopir, Natalie M. Rice, Suzie Allard, Lynn Baird, Josh Borycz, Lisa Christian, Bruce Grant, Robert Olendorf, and Robert J. Sandusky. Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLOS ONE*, 15(3):e0229003, March 2020. ISSN 1932-6203. doi:10.1371/journal.pone.0229003. URL <https://dx.plos.org/10.1371/journal.pone.0229003>.
- Christopher J Markiewicz, Krzysztof J Gorgolewski, Franklin Feingold, Ross Blair, Yaroslav O Halchenko, Eric Miller, Nell Hardcastle, Joe Wexler, Oscar Esteban, Mathias Goncavles, Anita Jwa, and Russell Poldrack. The OpenNeuro resource for sharing of neuroscience data. *eLife*, 10:e71774, October 2021. ISSN 2050-084X. doi:10.7554/eLife.71774. URL <https://elifesciences.org/articles/71774>.
- Leho Tedersoo, Rainer Küngas, Ester Oras, Kajar Köster, Helen Eenmaa, Äli Leijen, Margus Pedaste, Marju Raju, Anastasiya Astapova, Heli Lukner, Karin Kogermann, and Tuul Sepp. Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, 8(1):192, July 2021. ISSN 2052-4463. doi:10.1038/s41597-021-00981-0. URL <https://www.nature.com/articles/s41597-021-00981-0>.
- Valentin Danchev, Yan Min, John Borghi, Mike Baiocchi, and John P. A. Ioannidis. Evaluation of Data Sharing After Implementation of the International Committee of Medical Journal Editors Data Sharing Statement Requirement. *JAMA Network Open*, 4(1):e2033972, January 2021. ISSN 2574-3805. doi:10.1001/jamanetworkopen.2020.33972. URL <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2775667>.
- Hyoungjoo Park, Sukjin You, and Dietmar Wolfram. Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. *Journal of the Association for Information Science and Technology*, 69(11):1346–1354, 2018. ISSN 2330-1643. doi:10.1002/asi.24049. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24049>.
- Nicolas Robinson-García, Evaristo Jiménez-Contreras, and Daniel Torres-Salinas. Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*, 67(12):2964–2975, 2016. ISSN 2330-1643. doi:10.1002/asi.23529. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23529>.
- Hyoungjoo Park and Dietmar Wolfram. An examination of research data sharing and re-use: implications for data citation practice. *Scientometrics*, 111(1):443–461, April 2017. ISSN 1588-2861. doi:10.1007/s11192-017-2240-2. URL <https://doi.org/10.1007/s11192-017-2240-2>.
- Heather A. Piwowar, Roger S. Day, and Douglas B. Fridsma. Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLOS ONE*, 2(3):e308, March 2007. ISSN 1932-6203. doi:10.1371/journal.pone.0000308. URL <https://dx.plos.org/10.1371/journal.pone.0000308>.
- Kerstin Helbig, Brigitte Hausstein, and Ralf Toepfer. Supporting Data Citation: Experiences and Best Practices of a DOI Allocation Agency for Social Sciences. *Journal of Librarianship and Scholarly Communication*, 3(2):1220,

- September 2015. ISSN 2162-3309. doi:10.7710/2162-3309.1220. URL <https://jlscc-pub.org/article/10.7710/2162-3309.1220/>.
- Daniel Torres-Salinas, Evaristo Jiménez-Contreras, and Nicolas Robinson-García. How many citations are there in the Data Citation Index?, September 2014. URL <http://arxiv.org/abs/1409.0753>.
- Carol Tenopir, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLOS ONE*, 10(8):e0134826, August 2015. ISSN 1932-6203. doi:10.1371/journal.pone.0134826. URL <https://dx.plos.org/10.1371/journal.pone.0134826>.
- Isabella Peters, Peter Kraker, Elisabeth Lex, Christian Gumpenberger, and Juan Gorraiz. Research Data Explored: Citations versus Altmetrics, April 2015. URL <http://arxiv.org/abs/1501.03342>.
- Senay Kafkas, Jee-Hyub Kim, Xingjun Pi, and Johanna R. McEntyre. Database citation in supplementary data linked to Europe PubMed Central full text biomedical articles. *Journal of Biomedical Semantics*, 6(1):1, 2015. ISSN 2041-1480. doi:10.1186/2041-1480-6-1. URL <http://www.jbiomedsem.com/content/6/1/1>.
- Heather A. Piwowar, Jonathan D. Carlson, and Todd J. Vision. Beginning to track 1000 datasets from public repositories into the published literature. *Proceedings of the American Society for Information Science and Technology*, 48(1): 1–4, 2011. ISSN 0044-7870. doi:10.1002/meet.2011.14504801337. URL <https://onlinelibrary.wiley.com/doi/10.1002/meet.2011.14504801337>.
- Gianmaria Silvello. Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, 69(1):6–20, September 2017. ISSN 2330-1635. doi:10.1002/asi.23917. URL <https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.23917>.
- Lisa M. Federer. Long-term availability of data associated with articles in PLOS ONE. *PLOS ONE*, 17(8):e0272845, August 2022. ISSN 1932-6203. doi:10.1371/journal.pone.0272845. URL <https://dx.plos.org/10.1371/journal.pone.0272845>.
- Yang Zhang, Rongying Zhao, Yufei Wang, Haihua Chen, Adnan Mahmood, Munazza Zaib, Wei Emma Zhang, and Quan Z Sheng. Towards employing native information in citation function classification. *Scientometrics*, pages 1–21, 2022.
- Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. URL <https://arxiv.org/abs/1810.04805>.
- Nees Van Eck and Ludo Waltman. Software survey: Vosviewer, a computer program for bibliometric mapping. *scientometrics*, 84(2):523–538, 2010.
- Evaristo Jiménez-Contreras, Daniel Torres-Salinas, and Nicolas Robinson-Garcia. Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*, 2015.
- Xiaolei Huang, Bradford A Hawkins, Fumin Lei, Gary L Miller, Colin Favret, Ruiling Zhang, and Gexia Qiao. Willing or unwilling to share primary biodiversity data: results and implications of an international survey. *Conservation letters*, 5(5):399–406, 2012.
- Mai H Oushy, Rebecca Palacios, Alan EC Holden, Amelie G Ramirez, Kipling J Gallion, and Mary A O’Connell. To share or not to share? a survey of biomedical researchers in the us southwest, an ethnically diverse region. *PLoS One*, 10(9):e0138239, 2015.
- Youngseek Kim and Ayoung Yoon. Scientists’ data reuse behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, 68(12):2709–2719, 2017.
- II Force. Force ii: The future of research communication and scholarship, joint declaration of data citation principles, 2014. URL <https://force11.org/info/joint-declaration-of-data-citation-principles-final/>.
- John Kratz and Carly Strasser. Data publication consensus and controversies. *F1000Research*, 3, 2014.
- Micah Altman, Christine Borgman, Mercè Crosas, and Maryann Matone. An introduction to the joint principles for data citation. *Bulletin of the Association for Information Science and Technology*, 41(3):43–45, 2015.
- Hyoungjoo Park. The interdisciplinarity of research data: How widely is shared research data reused in the stem fields? *The Journal of Academic Librarianship*, 48(4):102535, 2022.
- John P Holdren et al. Memorandum for the heads of executive departments and agencies: Increasing access to the results of federally funded scientific research, 2013. URL https://rosap.ntl.bts.gov/view/dot/34953/dot_34953_DS1.pdf.
- Marc Schiltz. Plan s, 2018. URL <https://www.coalition-s.org/plan-s-funders-implementation/>.

Fabiane Führ and Edgar Bisset Alvarez. Digital humanities and open science: initial aspects. In *International Conference on Data and Information in Online*, pages 154–173. Springer, 2021.

Christine L. Borgman. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6):1059–1078, April 2012. ISSN 1532-2882. doi:10.1002/asi.22634. URL <https://onlinelibrary.wiley.com/doi/10.1002/asi.22634>.

Matthew S Mayernik. Open data: Accountability and transparency. *Big Data & Society*, 4(2):2053951717718853, 2017.