

Research on Publisher Policies in Scientific Data Open Sharing (Postprint)

Authors: Cui Yan

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose / Significance] In the open sharing ecosystem of scientific data, publishers represent key stakeholders. They collaborate with other stakeholders—including funders, data centers and repositories, and research institutions—to foster a culture of open sharing for research data. Although publishers have recognized the significant value of open scientific data sharing, their participation in open sharing practices for research data has not yet achieved scale. [Methods / Process] This study systematically reviews current publisher policies regarding open data sharing, analyzes their commonalities and gaps, and examines these policies within the context of existing scientific data sharing practices. [Results / Conclusion] The policy recommendations proposed for publishers will help them further advance the goal of achieving open sharing of research data.

Full Text

Research on Publishers' Policies for Open Sharing of Scientific Data

Cui Yan^{1, 2}

¹Beijing Jiaotong University Library, Beijing 100044

²National Science Library, Chinese Academy of Sciences, Beijing 100190

Abstract

[Purpose/Significance] Publishers are key stakeholders in the open sharing ecosystem for scientific data. They collaborate with other stakeholders—including funders, data centers, repositories, and research institutions—to foster a culture of open sharing for research data. Although publishers have recognized the significant value of open scientific data, their participation in open data sharing practices has not yet reached scale. [Method/Process] This study

systematically reviews current publisher policies on open data sharing, analyzing their commonalities and gaps to assess how publisher policies function in contemporary scientific data sharing practices. **[Result/Conclusion]** The paper proposes targeted policy recommendations for publishers to help them further advance the goals of open research data sharing.

Keywords: open sharing of scientific data, publishers, policy

Classification Number: G251

Open sharing of scientific data holds significant importance for enhancing both economic and academic value. As the European Commission has noted, “Open access to scientific data improves data quality, reduces the costs of duplication, accelerates scientific progress, and helps combat scientific fraud” [1]. A report from the Royal Society argues that openness is vital for science—it can identify errors, support or refute theories, and optimize them; data reuse enables deeper understanding and mastery of knowledge [2]. Open data is part of open science, characterized by being accessible, understandable, assessable, and usable. When combined with open access to scientific publications, open data can effectively facilitate communication and exchange of scientific research content.

Various stakeholders participate in the scientific research process. These stakeholders are diverse in type and nature, collectively forming the scientific research ecosystem. Based on their functions and centered around researchers, this system can be classified into four stakeholder groups: funders, data managers, research institutions, and publishers—all key actors in promoting and achieving open sharing of scientific research data. In this ecosystem, each stakeholder group has distinct functions, drivers, and interests, yet they share the overarching goal and motivation of advancing scientific progress. Stakeholder interests and behavioral motivations connect with their respective functions, forming the boundaries of the scientific research system. This system is essentially influenced by both external and internal factors: external factors may manifest as political, social, economic, scientific, and cultural systems; internal factors include political will, economic and academic competition, technical infrastructure, legal and ethical considerations, among others. Within this system, internal factors control the availability and allocation of resources. As shown in [Figure 1: see original paper]:

[Figure 1: see original paper]

In this open scientific system, stakeholder groups have numerous and occasionally overlapping functions. While all parties generally agree on the significance of open access to research data, they hold inconsistent views on how to achieve it. This study primarily examines the current state and challenges of publisher data policies in promoting open scientific data sharing, constructs corresponding models, and proposes recommendations.

1.2 Publishers' Objectives in the Open Scientific Data Sharing System

In promoting open sharing of research data, different stakeholders play distinct roles. Publishers provide data sharing policies to authors through journals, playing a crucial role in data openness. Researching publisher data policies is important because this is the point in the research process where data policies are closely linked to researchers. To publish their findings, researchers have both the motivation and possibility to comply with these policies. However, studies show that although funders and research institutions may have policies applicable to researchers, compliance rates remain low. Researchers lack direct incentives to follow data policies and may not have appropriate institutional repositories available when data archiving is required. Publisher data policies, proposed after research is completed but before publication, represent a point where researchers are both able and motivated to comply.

Many publishers recognize the Brussels Declaration, which states that “All researchers should have the freedom to access raw research data. Publishers encourage the public disclosure of raw research data results and the submission of relevant datasets and subsets to journals alongside papers, which should be made freely accessible to other researchers whenever possible” [6]. Many publisher data policies are mandatory, requiring authors to store research data supporting publications in certified repositories to achieve open access. Research indicates that if a journal has mandatory data policies and includes data availability statements, the likelihood of finding that data online is nearly 1,000 times greater than for journals without such policies [7].

Despite current attention to open access publications as a business model, publisher engagement in research data publication—particularly open access to research data—has not yet reached scale. Publishers focus on research data and its openness because data adds value to their primary product: publications. Data helps verify research findings, thereby enhancing the credibility of published results, which is crucial for research [2]. Additionally, funder policies requiring open access to research data essentially exert pressure on publishers, who must participate in open research data access to help authors and institutions meet funder requirements. Leading publishers are now collaborating with other stakeholders to unlock the potential of research data, develop new data-based products and services, conduct peer review of research data, and provide other services to improve data quality.

2 Major Publishers' Data Management Policies

Based on STM's annual report [8], this study identifies the following publishers as primary analysis objects (see). These ten publishers accounted for 45.2% of all journals published in 2014. Additionally, this study includes the data openness policies of open access publishers Public Library of Science (PLOS) and BioMed Central (BMC).

2.1 Current State of Publisher Data Policies

2.1.1 Open Access as the Default Publishers establish open access to data as the foundation for policy development. For example, PLOS' s mandatory policy for open access to research data [9] states that, with rare exceptions, all research data supporting PLOS publications must be openly accessible. When submitting to PLOS, authors must simultaneously submit a data availability statement indicating compliance with PLOS policy requirements. After manuscript acceptance, data are published as part of the final manuscript. PLOS requires authors to store data in recommended, certified data centers or repositories; small datasets may be uploaded with the manuscript. PLOS editorial and submission guidelines provide researchers with instructions to help them comply with journal open data policies. In cases where data access is restricted, PLOS reserves the right to publish corrections, contact the author' s institution or funder, or even retract the publication.

2.1.2 Emergence of New Publication Formats Growing attention to open data in the publishing world has given rise to a new product: data journals. The emergence of data journals is closely related to the ability to publish data separately, ensuring that data—as a fundamental component of the scientific record—are provided to the scientific community in an understandable form. Data journals are peer-reviewed, open access platforms for publishing, sharing, and disseminating data across disciplines. Published data papers include specific information about datasets, such as collection and processing methods. Data papers are linked to recognized repositories, with the paper citing the dataset stored in the repository or data center. As the Australian National Data Service (ANDS) notes in its data journal guidelines, “Fundamentally, data journals seek to promote scientific certification and reuse, improve the transparency of scientific methods and results, support good data management practices, and provide an accessible, permanent, and resolvable path to datasets” [10]. ANDS points out that the data paper publication process includes peer review of datasets, maximizing opportunities for data reuse and providing researchers with possibilities for academic recognition.

2.1.3 Peer Review Open access to research data and data publications highlights the scientific quality of research data and the importance of peer review. In this context, some publishers have expanded the scope of peer review to include data. M. S. Mayernik et al. (2014) conducted research on “peer review of datasets,” proposing that because publications or resource types differ, peer review methods must also vary. The researchers distinguished several types of reviewable resources, including data analyzed in traditional scientific papers, data articles published in traditional scientific journals, and open access repositories and datasets published through data journals. Mayernik et al. identified common factors for dataset peer review and quality assurance processes, including: datasets must be accessible through data centers or repositories; datasets must contain sufficient information for review; journals must have clear guide-

lines indicating review points to guide reviewers in data examination [11].

The Journal of Open Archaeology Data (JOAD) [12-13] employs peer review for all submitted data papers, evaluating both the paper's content and the stored data. The paper's content refers to information related to dataset creation and reuse, as well as dataset descriptions. Stored data refers to data submitted to repositories in a sustainable manner, including licensing arrangements.

2.1.4 Data Citation In addition to peer review of data, publishers are gradually introducing data citation policies to promote standardized use of research data. Research data holds significant value for the research process. To expand the dissemination of high-quality research data and establish normative practices for data utilization, FORCE11 [14] developed main principles for data citation. FORCE11's citation principles are premised on the need for data citations to be both human- and machine-readable. These principles may not be comprehensive, but their main purpose is to encourage disciplines to develop citation methods that reflect their own characteristics.

FORCE11 principles include: explaining the importance of data citation; promoting academic credit through data citation; data citations must be machine-actionable, including globally persistent unique identifiers; data citations must facilitate access to the data itself; data citations should be persistent; they should be easily identifiable, accessible, and verifiable; and they should be interoperable and flexible.

2.1.5 Content Discovery and Linking Services Publishers are gradually shifting their development focus toward content discovery and linking services, concentrating on text and data mining (TDM) tools to develop content and provide related services. Publishers' growing attention to data mining directly responds to researchers' needs to utilize large databases of content, data, and publications. A report from an expert panel on text and data mining notes that TDM is an important technology for analyzing and extracting new insights and knowledge from exponentially growing digital data [15]. The report concludes that as researchers' skills and technologies continue to improve and the complexity, diversity, and scale of studied datasets expand, TDM is likely to become even more important. However, whether restrictions should be placed on the use of text and data mining tools remains controversial.

2.2 Problems with Journal Data Policies

2.2.1 Differences in Data Formats and File Sizes Journals often find it difficult to include all data supporting article results. Different research methods generate vastly different data, with significant variations in data formats and file sizes. Data from qualitative research often exist as text, such as field observation notes or interview transcripts. Quantitative research data are often preserved as spreadsheets. A single study may generate multiple data types, while a paper may include additional text, numerical datasets, and digital images—all of

which can increase paper size. Consequently, publishers express concern about the size of datasets integrated into each paper. Some publishers have begun experimenting with publishing online journal articles that include multiple data types, such as Elsevier's Article of the Future initiative [16]. However, not every journal has the capacity to include various data types. Therefore, journal policies should clearly state the extent to which data can be included as paper components.

2.2.2 Costs of Institutional Repositories To address publisher server overload, linking important journal article data to a specific institutional repository may be a reasonable option, but this transfers related long-term operational costs to institutions. Current funder grants do not include these costs, and institutions may be unwilling to add such expenses to existing management budgets. This raises questions about the sustainability of storing scientific data in institutional repositories.

2.2.3 Researchers' Awareness of Data Openness Many researchers are unaware of relevant repositories. Therefore, journal data policies should specify whether data should be stored in recognized repositories, whether permanent URLs should be used, and whether certain forms of data citation are required. The timing of data publication is also a concern. Researchers are less concerned with when data become publicly accessible during the publication process than with when data should be made publicly accessible during the research process. Research papers are not produced only after all results are generated but are developed gradually throughout the research process. Whether it is appropriate to publish research data simultaneously with paper publication depends on multiple factors, such as data sensitivity and the need to protect subjects.

3 Framework Model for Journal Research Data Policies

3.1 Basic Requirements for Publisher Journal Policies

An effective policy development process must necessarily focus on the perspectives of all stakeholders in the data sharing process. Current digital infrastructure is constantly evolving, and there are no strong measures encouraging data sharing among publishers, repositories, and systems. Common points include: (1) publishers vary greatly in how they share data; (2) publishers provide unclear instructions on data types accepted during publication, where data should be stored, and for how long; (3) researchers generally support data sharing but do not know how to overcome sharing obstacles; (4) researchers believe clear publisher and journal policies on data formats and storage locations would be beneficial; (5) publishers also acknowledge barriers in data linking and embedding.

Many issues in scientific data sharing can be resolved through strong and clear journal policies during the publication process. Therefore, this study aims to

identify policy models that can be recommended for journals. Collected journal policy information was analyzed to identify current main policy requirements, as shown in :

3.2 Framework Model

Based on the observations listed above, a basic model framework for journal research data policies was developed, with funders and research institutions collaborating as one party with publishers as the other to develop data policies. presents the constructed framework model for journal research data policies.

3.3 Recommendations

3.3.1 Gradually Develop Mandatory Open Sharing Policies for Supporting Data Publishers should support journal editors in developing mandatory research data policies to enhance research process transparency and unlock the potential of research data. Editorial policies should address issues such as documentation, metadata, data publication formats, licensing, and citation. These policies should require authors to submit data availability statements during manuscript submission and peer review. Policies should propose corresponding measures for non-compliant submissions, such as retracting published articles.

3.3.2 Collaborate with Certified Repositories and Data Centers to Simplify Data Access Publishers can assess the credibility of repositories and data centers according to relevant standards and collaborate with those meeting data certification criteria. Such repositories may be domain-specific disciplinary repositories, institutional repositories, or comprehensive repositories. This collaboration should have positive impacts, providing researchers and institutions with high-quality, operable, and technically seamless products and services to achieve open dissemination and preservation of high-quality academic products and research data. For disciplines without recognized data centers or repositories, publishers should provide researchers with guidance and assistance, recommending suitable institutional repositories for data storage and access, or providing commercial data services. In this regard, publishers should primarily facilitate academic communities in establishing repository evaluation criteria to help researchers select appropriate repositories. Once the academic community establishes recognized repository standards, publishers can enforce these standards through journal policies.

3.3.3 Treat Data as First-Class Scholarly Output Subject to Peer Review One of publishers' primary responsibilities is ensuring publication quality, and establishing a research data peer review system helps enhance product quality. The peer review process should detail evaluation criteria, assessing both technical and quality aspects of research data. Technical aspects include dataset completeness and consistency, data collection standards, and software

used. Scientific quality of research data is evaluated by the research community through pre-publication and post-publication peer review. Publishers should also explore incentives for reviewers, including payment, invitations to write special features, editorial board membership, or even hiring experts for content review.

3.3.4 Develop Strategies for Research Data Citation For research data to be reusable long-term, citation standards must be established and used. Publishers should require that publications and related data be citable and provide clear guidance for data citation. In developing data citation guidelines, publishers can reference existing practices such as Force11 principles [14] and participate in DataCite [17], joining ongoing discussions in the research and editorial communities. Data citations should include DOIs and licensing information, such as Creative Commons licenses. Data citation methods should preferably be machine-actionable, allowing users to understand how research data can be used.

3.3.5 Establish Licensing Policies to Encourage Text and Data Mining Publishers' editorial policies should clearly state how research data are published and accessed, such as default or recommended open access. Considering that text and data mining tools can generate significant economic benefits, publishers are encouraged to adjust their policies to allow researchers to use these technologies in their research.

4 Conclusion

Although publishers have strongly emphasized open access to publications as a business model, their participation in research data—particularly open research data—has not yet reached scale. Publishers focus on research data and its openness because data adds value to their primary product: publications. Through the verification function of research findings, data enhances the credibility of published results, which is crucial for research. Simultaneously, to respond to funder pressure for open access, publishers are gradually developing relevant policies.

Publishers and their data policies can play a key role in promoting broader adoption of data archiving and assisting in developing models for data archiving policies. Publisher data policies should embody transparency, enable reuse possibilities, propose data storage requirements, provide guidance on data structure and metadata, and direct authors to appropriate repositories with network links. Data policies benefit not only researchers themselves and their peers in the same or related fields but also promote data archiving and linking, improving data storage.

References

- [1] European Commission. Recommendation on access to and preservation of scientific information [EB/OL]. [2016-03-20]. http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf
- [2] BOULTON G, CAMPBELL P, COLLINS B, et al. Science as an open enterprise [EB/OL]. [2016-03-20]. https://royalsocietypublishing.org/~/media/royal_society_content/policy/projects/2016-20-saoe.pdf.
- [3] Council of European Social Science Data Archives. Depositing data -benefits [EB/OL]. [2016-03-20]. <http://www.cessda.org/sharing/depositing/1/>.
- [4] FRY J, LOCKYER S, OPPENHEIM C, et al. Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and Research Institutes[EB/OL]. [2016-03-20]. http://repository.jisc.ac.uk/279/2/JISC_data_sharing_finalreport.pdf.
- [5] United Nations Educational, Scientific and Cultural Organization. Policy guidelines for the development and promotion of open access [EB/OL]. [2016-03-20]. <http://unesdoc.unesco.org/images/0021/002158/215863e.pdf>.
- [6] STM. Brussels declaration on STM publishing[EB/OL]. [2016-03-20]. http://www.stm-assoc.org/2007_11_01/Brussels_Declaration.pdf.
- [7] VINES T H, ANDREW R L, BOCK D G, et al. Mandated data archiving greatly improves access to research data[J]. The FASEB journal, 2013, 27(4): 1304-1308.
- [8] WARE M, MABE M. The stm report. An overview of scientific and scholarly journal publishing [EB/OL]. [2016-04-16]. http://www.stm-assoc.org/2015_02_20/STM_Report_2015.pdf.
- [9] PLOS ONE. Materials and software sharing [EB/OL]. [2016-04-16]. <http://www.plosone.org/static/policies.action#sharing>.
- [10] Australian National Data Service. Data journal guide [EB/OL]. [2016-03-20]. <http://ands.org.au/guides/data-journals.html>.
- [11] MAYERNIK M S, CALLAGHAN S, LEIGH R, et al. Peer review of datasets: When, why, and how [J]. Bulletin of the American Meteorological Society, 2015, 96(2): 191-201.
- [12] The journal of open archaeology data (JOAD) [EB/OL]. [2016-03-20]. <http://openarchaeologydata.metajnl.com/>.
- [13] The Journal of Open Archaeology Data. Editorial policies, peer review process [EB/OL]. [2016-03-20]. <http://openarchaeologydata.metajnl.com/about/editorialPolicies#peerReviewProcess>.
- [14] FORCE11. Joint declaration of data citation principles -FINAL [EB/OL].

[2016-03-20]. <https://www.force11.org/group/joint-declaration-data-citation-principles-final>.

[15] European Commission. Standardization in the area of innovation and technological development, notably in the field of text and data mining [EB/OL]. [2016-03-20]. http://ec.europa.eu/research/innovation-union/pdf/TDM-report_{{{from}}}{{{the}}}{{{expert}}}_{{{group}}}-042014.pdf.

[16] The article of the future: creating an optimal way to communicate research [EB/OL]. [2016-03-20]. <http://www.articleofthefuture.com/>.

[17] DataCite. Why cite data? [EB/OL]. [2016-03-20]. <http://www.datacite.org/whycitedata>.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv –Machine translation. Verify with original.