Research on Open Research Data Policies of Foreign Publishers

Authors: Peng Yunyuan, Liu Jingyu, Huang Jinxia

Date: 2017-09-29T00:00:00+00:00

Abstract

[Purpose/Significance] Based on existing policy content, this paper discusses areas that still require improvement in research data policies, aiming to provide reference and guidance for the publishing community to formulate unified and comprehensive open data policies.

[Method/Process] In a data-intensive environment, this study selected eight internationally renowned, mainstream, and representative publishing institutions as research objects. Using publisher platforms, journal websites, and the re3data.org data repository directory system as data sources, the data policy content of publishing institutions was summarized and analyzed according to three aspects: submission policies, storage policies, and utilization policies.

[Results/Conclusion] Future efforts should actively improve data submission policies, establish peer review models, standardize repositories, and clarify data reuse methods.

Full Text

Research on Open Scientific Data Policies of Foreign Publishing Institutions

Peng Yuanyuan[1,2], Liu Jingyu[1], Huang Jinxia[1,2] [1]. National Science Library, Chinese Academy of Sciences, Beijing, 100190, China [2]. University of Chinese Academy of Sciences, Beijing, 100049, China

Abstract

[Purpose/Significance] Based on existing policy content, this paper discusses areas for future improvement in scientific data policies, aiming to provide reference for the publishing industry to develop unified and comprehensive open

data policies. [**Method/Process**] In the context of data-intensive research environments, this study selected eight renowned, mainstream, and representative foreign publishing institutions as research subjects. Using publishing platforms, journal websites, and the re3data.org data repository directory system as data sources, we summarized and analyzed the content of these institutions' data policies according to three aspects: submission policies, storage policies, and utilization policies. [**Results/Conclusions**] Future efforts should focus on improving data submission policies, establishing peer review models, standardizing repositories, and clarifying data reuse methods.

Keywords: Scientific Data, Open Data, Data Policy, Publishing Institutions

1. Introduction

Research data constitutes an essential strategic resource for scientific research in the digital information age—a valuable collection of data generated during processes such as data collection, experimental analysis, and research output, forming the foundation and support for knowledge construction in science, technology, and medicine. Since the Budapest Open Access Initiative (BOAI) in December 2001 [1], the open access movement has gained momentum, and the open sharing of research data has gradually gained recognition. In October 2003, the Max Planck Society proposed the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, advocating for the open sharing of scholarly papers, educational resources, and research data, and calling upon major research institutions to provide open access to scientific resources [2]. In 2006, the Organization for Economic Co-operation and Development (OECD) [3] advocated for improving the timeliness, accessibility, and interconnectivity of publicly funded research data. In May 2012, the Royal Society published the report *Science as an Open Enterprise*, which stated that the immediate publication of scientific theories and specific experimental data forms the basis for others to understand and evaluate research, and serves as the foundation for falsification and innovation [4]. Open research data, as a new approach and concept for sharing data and information resources in the digital network environment—also known as "open scientific data" or "open data" —is expressed in English as "Open Data," "Open Scientific Data," or "Open Research Data" [5].

To improve the open access system for research data, various stakeholders including government departments, international organizations, and funding and educational institutions at home and abroad have introduced research data management and sharing policies to promote the accessibility, shareability, and reusability of research data. These efforts aim to achieve broader academic exchange and research integration through open data sharing, enabling new breakthroughs in scientific research. In April 2011, Research Councils UK (RCUK) released the *RCUK Common Principles on Data Policy* [6], establishing rules for research data usage. In February 2013, the U.S. White House Office of Science and Technology Policy (OSTP) [7] issued a government research data open policy stating that "digital research data generated from federally funded research

projects must be stored and made publicly accessible and shared through search, retrieval, and usage functions." The United States also promulgated the Open Data Charter through G8 meetings, with G8 members including the United Kingdom, Japan, and France responding actively. Canada, India, New Zealand, and other governments issued open data declarations, while Australia, Brazil, and Chile introduced relevant policies and regulations on data openness and protection. The Chinese Academy of Sciences' *Measures for Scientific Database Data Sharing (Trial)* also standardized issues such as data revenue distribution and property rights between the Academy and various units establishing databases. Fudan University' s Social Science Research Center established the China Social Science Research Data Sharing Center using Harvard University' s Dataverse network platform, while Peking University established the Peking University Open Data Research Platform and Management Science Data Center based on open research data sharing mechanisms.

As stakeholders in the open access movement and important media for research data exchange and dissemination, the publishing industry has also introduced relevant policies for the open sharing of research data. Starting with data policies proposed by individual journals, major publishing groups have sparked a new wave of policy development in the publishing world, striving to build a data management and sharing model that is associated with publications, high-quality, and suitable for reuse. This approach aims to enhance the credibility of publications by verifying data authenticity, thereby improving their academic value and contributing to the development of open science. However, existing research has not yet systematically studied the content of open research data policies of foreign publishing institutions or future policy development and improvements. To comprehensively understand the current state of open research data policies and specific areas requiring future improvement, this study investigated the open research data policies of foreign publishing institutions. We selected eight renowned and influential traditional publishing groups and emerging open access publishers, categorized their data policies according to three aspects—submission policies, storage policies, and utilization policies—and discussed their future development trends. This provides reference for improving data policies in the publishing industry and promoting the free access, republication, and reuse of research data.

2.1 Research Subjects

Journals such as *Science* and *Nature* pioneered research data storage and dissemination policies, leading the rise and development of open data policies in the publishing industry. Subsequently, publishers including PLOS and BioMed Central proposed more general and explicit data sharing and management policies for their journals. Therefore, this study referenced the publisher members listed by the International Association of Scientific, Technical & Medical Publishers (STM) [8], publishing institutions mentioned in the 2016 report on the world's 52 largest book publishers [9], and emerging open access publishers in the

open access publishing environment. Considering whether these institutions had established research data policies, we ultimately selected eight renowned and influential foreign publishing institutions as our research and analysis subjects: Springer Nature, Elsevier, Wiley, Taylor & Francis Group, Oxford University Press (OUP), Hindawi, Public Library of Science (PLOS), and BioMed Central (BMC).

It should be noted that BMC was acquired by Springer in 2008, but has remained committed to the open access publishing industry, while Springer merged with Nature Publishing Group in 2015. Consequently, this study lists them separately as traditional publishing institutions and open access publishing institutions for comparative analysis from two perspectives.

2.2 Data Sources

The data for this study primarily originated from the official websites of each publishing institution, the *Scientific Data* journal (which provides lists of research data storage repositories), and the comprehensive repository registry and directory system re3data.org, supplemented by literature research. The investigation cutoff date was April 25, 2017.

Based on the research data sharing and management policies of each publishing institution, we elaborate in detail on specific policy content elements and future development and improvement of these policies.

3. Analysis of Publishers' Open Research Data Policy Elements

Across the eight publishing institutions selected for this study, most data policies are still in trial phases. Some institutions are in the initial stages of policy formulation, with preliminary policy frameworks but not yet fully determined content. Others have data policies applicable only to some of their journals, not their entire portfolio. Among the selected publishers, Springer Nature and Elsevier have relatively detailed data policy content. While PLOS and BMC introduced data policies earlier, their policy terms are fewer and more general. Wiley and OUP pilot their data policies for certain open access journals. Taylor & Francis and Hindawi are in the early stages of policy construction, having developed some terms and overall frameworks. Based on the content of these publishers' data policies, this study summarizes their elements and provides a more detailed analysis from three aspects: submission, review and storage, and utilization policies.

3.1 Data Submission Policies

Publishers' research data submission policies primarily focus on whether data submission is optional or mandatory, submission methods, specified data types, and format requirements [10].

Springer Nature categorizes all its journals into four types, with research data submission being optional or mandatory depending on the journal. Specific requirements can be found in the "Author Guidelines," "Editorial Instructions," and "Data Policy" modules [11]. The characteristics of these data policies are classified as mandatory, recommended/optional, or with no special requirements. Specific policy details are shown in Table 1 . Among the four policy types, three encourage or require authors to provide a Data Availability Statement (DAS) when submitting articles and datasets. This statement should indicate where the data supporting the article' s conclusions can be found, including links to publicly released datasets analyzed or generated during the research period. The DAS should clearly specify: (1) that datasets generated and/or analyzed during the current study are available in a specified repository; (2) if datasets are not publicly available, the reasons should be explained and conditions under which they may be provided upon reasonable request; (3) that datasets are available upon reasonable request from the authors; (4) if data sharing is not applicable, reasons should be stated (e.g., no datasets were generated or analyzed); and (5) that all data generated or analyzed in the study are included in the article and/or its supplementary information files. Additionally, Springer Nature requires authors to upload data in three main ways: (1) encouraging authors to store research data in appropriate public repositories without requiring submission to the journal; (2) requiring descriptive materials (all relevant raw data) in original manuscripts to be submitted with the article while encouraging dataset storage in public repositories; and (3) providing all datasets related to the article' s conclusions and storing them in public repositories before peer review as supplementary materials [12].

Elsevier recognizes that research data forms the foundation of scientific, technical, and medical knowledge but acknowledges challenges in data access and sharing. In line with the STM Brussels Declaration, Elsevier envisions a future where data can be easily and effectively stored, shared, discovered, and used to support researchers' work and advance science and health [13]. To address potential challenges, Elsevier has formulated corresponding recommended research data policies and implemented a pilot program for open data publication and sharing in 2015, encouraging authors to upload raw research data to Mendeley Data [14] and link it to their ScienceDirect articles. Wiley has launched open data sharing services for 13 of its journals, each with its own data policy. Authors must consult the detailed guidelines on each journal' s homepage during submission [15]. These 13 journals have no mandatory requirements for data file upload, defining them only as optional data sharing agreements with no format requirements. Articles with data sharing agreements need only upload datasets to public repositories and provide accurate links. OUP' s submission policy is similar to Wiley' s, but OUP emphasizes the need to provide a Data Availability Statement that clarifies data validity and specific applications. Datasets must be stored in standard public repositories with corresponding accession numbers [16]. Taylor & Francis only requires provision of the data source URL [17]. The three open access publishers—Hindawi, PLOS, and BMC—particularly em-

phasize submission of Data Availability Statements [18]. Among these, BMC explicitly states that data file types should be XML, CSV, XLS/XLSX, or RDF [19] to facilitate unified management and subsequent data mining and text content analysis.

3.2 Data Review and Storage Policies

Among the publishers selected in this study, only Springer Nature has established a review mechanism for submitted datasets. Elsevier treats raw research data as supplementary material for editorial and reviewer examination. Although Springer Nature' s review mechanism has not been implemented across all journals, selected journals require peer review of datasets and Data Availability Statements to assess whether authors have complied with Springer Nature' s policy requirements regarding research data availability and have made efforts to ensure that supporting data can be replicated or reused by other researchers. Reviewers have the right to request access to underlying data (or code) to evaluate manuscripts.

Reviewers of Data Availability Statements should consider: (1) Can readers access the data? (2) Do the links provided in the DAS function properly? (3) When data access is restricted, are the conditions reasonable? (4) Are descriptions of data included in the manuscript and/or supplementary information files accurate?

For data files, reviewers should consider: (1) Is the data stored in the most appropriate repository? (2) Was the data generated using rigorous methodology? (3) Do the data and metadata comply with academic formatting and reporting standards? (4) Are the data documents stored by authors complete and consistent with manuscript descriptions? (5) Do the datasets include personal, sensitive, or inappropriate information?

Regarding data storage policies, authors can generally choose repositories for most disciplinary data types, either uploading to comprehensive repositories like Figshare and Dryad that accommodate all research data types, or to discipline-specific repositories. Springer Nature only provides a reference list of specialized repositories for life science journals and offers a recommended repository list for authors based on its open access data journal *Scientific Data* [21]. Elsevier requests but does not require authors to add datasets to the Mendeley Data management space, where datasets can be used by others under licenses chosen by authors. Mendeley Data assigns DOIs to datasets and associates them with ScienceDirect articles, enabling readers to access datasets through the articles [22]. Other publishing institutions have not specified data storage repositories or mentioned recommended repositories, leaving the choice to authors' discretion.

3.3 Data Utilization Policies

Springer Nature and BMC stipulate that data utilization follows the CC0 waiver agreement, with specific usage licenses determined by the data repository where the dataset is stored. Elsevier' s data files follow the CC BY license agreement. Wiley adopts the CC0 agreement for uploaded data and permits data mining and large-scale meta-analyses. Other publishers have not explicitly stated usage licenses for data documents on their websites, primarily referring to the usage licenses of the associated papers. Therefore, users must also consult the specific usage license restrictions of the data repository before utilization.

The emergence and development of open scientific data repositories reflect the growing volume of open scientific data and increasing demand for its utilization. Repository construction entities include individual universities, research institutions, and publishers, as well as cross-organizational and cross-regional collaborations such as the EU CESSDA Data Portal. Data sources mainly come from two channels: (1) mandatory submission of research data involved in studies, as required by publishers like Springer Nature; and (2) encouragement or recommendation for researchers to submit relevant scientific data, as practiced by PLOS One. Due to the proliferation of data storage repositories, researchers face difficulties discovering and selecting needed data sources, leading to the development of repository registries and directory systems such as OAD (Open Access Directory), re3data.org, and Databib.

As *Scientific Data* is an early data journal involving authoritative and standardized repositories, and re3data.org contains a large number of repositories with detailed disciplinary classifications and numerous subject themes, this study uses *Scientific Data* journal and re3data.org as data sources to comparatively analyze usage licenses for research data in data repositories.

The open research data repositories recommended by *Scientific Data* cover various disciplines and serve as the open data storage and sharing platform recommended by Springer Nature Publishing Group for its various journals, also endorsed by other publishers. The number of repositories 收录 in re3data.org has shown a yearly increasing trend. Repository types mainly include national, governmental, institutional, and publisher-built repositories. Construction entities comprise universities, research institutions, and publishers. The primary purposes of repository construction are open sharing and providing research data management tools, resulting in uneven disciplinary distribution with larger volumes in life sciences. Exchange protocols for data sharing also vary, with the United States being most prominent in building open data repositories.

Statistics from these data sources show that approximately 2,000+ data repositories have been recorded, with data licenses and database licenses in re3data.org [23] mainly falling into three categories (see Figure 1 [Figure 1: see original paper]):

(1) **Data Access & Database Access**: The openness of data/databases is

divided into four situations: open, restricted access (requiring registration or authentication), prohibited access (read-only), and completely invisible (limited to institutional or project team members). Openly accessible data accounts for 56.4% of the total (2,768); data with restricted access accounts for 32%; and inaccessible data accounts for 11.6%. Database openness is divided into three situations: completely open (94.5%), restricted (5%), and closed (0.5%).

(2) **Restricted Data Access & Restricted Database Access**: Reasons for restricted access mainly include: registration required, institutional membership, payment required, and other reasons. The primary restriction is registration-based access, with institutional membership requirements being minimal. Some reasons remain unknown.

(3) **Data Licenses/Database Licenses**: Primarily copyright restrictions; followed by CC license agreements, with CC0 comprising a larger proportion, indicating lower restrictions on reuse. Other licenses include Apache License 2.0, BSD (Berkeley Software Distribution), ODC (Open Data Commons), OGL (Open General License), public domain, and others.

Note: Data in Figure 1 are derived from the re3data.org platform and compiled through statistical analysis.

4. Future Development Trends and Recommendations for Publishers' Data Policies

Publishers have made attempts to promote research data management and sharing. Based on current policy formulation and implementation trends, publishers' open research data policies will become increasingly comprehensive. However, existing policies are primarily recommendatory and encouraging. There is a need to gradually develop mandatory and essential policies and incorporate data citation guidelines. The rise of data journals has introduced peer review mechanisms for research data publication, sharing, and dissemination. Correspondingly, when research data serves as article appendices, supplementary materials, or data support, peer review processes should also be established to recognize their value as academic achievements. With open storage and utilization models for research data not yet fully defined, implementing text and data mining (TDM) services and conducting meta-analyses represent future practical directions.

4.1 Trend Toward Increasingly Comprehensive Publisher Data Policies

Current data open sharing policies released by publishers and proposed initiatives for establishing open scientific data policies demonstrate that many foreign publishers have begun to attach importance to research data management. Driven by the open science environment, they continuously encourage authors

to open data from their research results. Springer Nature, BMC, and PLOS [24] have established data sharing agreements. Elsevier' s data policy primarily serves research data management, with open sharing requiring further analysis based on specific journal circumstances. Elsevier encourages authors to open data and has conducted "pilot projects" for open data, but does not mandate it. Wiley and OUP have introduced data sharing policies for some journals. Hindawi, as a fully open access journal publisher, stated in 2017 that developing an open data policy would be an important task for that year.

4.2 Establishing Mandatory Data Submission Policies

Research data submission policies should gradually move toward mandatory upload mechanisms for documentation and metadata. Following the unified regulations implemented by PLOS for all submitted articles, publishers should standardize formats, storage locations, and usage licenses for published data. Data Availability Statements should be further refined to clarify submission requirements, with penalties for policy violations (such as article retraction) even after publication.

4.3 Establishing a Peer Review System for Open Data

Establishing a peer review system for research data is a method to ensure high-quality output and verify the academic value of research results. The peer review process should specify that the technology, standards, dataset compatibility, and experimental environment completely match those used when drawing experimental conclusions, ensuring dataset integrity, consistency, relevant standards, and appropriate software. Scientific quality and value should be evaluated by reviewers, researchers, and practitioner communities through pre-publication and post-publication peer review.

Research data peer review can also collaborate with primary data managers (data centers and repositories), who are the main holders of research data content and provide access links. Data managers can ensure technical quality, security management, and preservation of research data. As the linkage between articles and data continues to increase, quality control of research data through data managers can better ensure access to high-quality data and long-term preservation.

4.4 Standardization of Open Data Repositories

The number of data repositories is increasing annually, including disciplinary, institutional, and comprehensive repositories. Distributing open data by discipline can facilitate further analysis of repository data. When specifying data sharing policies, publishers also recommend suitable repositories, but repository construction varies, with different openness levels and storage formats that should be standardized in the future. Publishers should also strengthen cooperation with high-trust repositories and data centers with strict storage standards.

For cases where neither disciplinary data centers nor repositories are specified, publishers could plan for contractual storage and access to facilitate resource interoperability and content sharing.

4.5 Clarifying Reuse Methods for Research Data

Different data repositories and datasets have varying standards for accessing and utilizing existing open datasets. To further promote research data reuse, the primary task should be clarifying reuse methods and restrictions for datasets. During the utilization phase, data citation policies should be explicit, including DOIs, license information (such as CC agreements), etc., to provide users with clear guidance on the purpose and norms of research data use. Research datasets involving privacy, confidentiality, or biological genetic patterns require more specific reuse methods and norms. Management of such data should be strictly limited, with clear and detailed conditions for use, which is also an important step in promoting open data development. Reuse methods for basic research data should be as universal and general as possible. Additionally, attempts can be made to establish licensing policies for text and data mining (TDM), providing users with deeper data services based on authors' data usage licenses and the regulations of publishing institutions and data repositories.

5. Conclusion

This study examines publishers' open research data policies as its entry point. By investigating and analyzing current data policies proposed by mainstream open access journal publishers abroad, we systematically discuss future areas for expansion in open research data policies, such as establishing mandatory data submission systems, creating research data peer review systems, standardizing open data repository content, and clarifying data reuse methods. This research aims to provide reference and guidance for the academic publishing industry to develop more comprehensive and complete open research data policies in the future. However, this study selected a limited number of publishing institutions and could not conduct broader and more in-depth investigations of numerous publishers. Future research should continue to track policy developments at other publishing institutions.

References

[1] Budapest Open Access Initiative [EB/OL]. [2009-10-27] http://www.budapestopenaccessinitiative.org/

[2] Berlin Declaration. Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities [EB/OL]. [2009-10-27] http://oa.mpg.de/openaccess-berlin/berlindeclaration.html.

[3] Organisation for Economic Co-operation and Development (OECD). Principles and Guidelines for Access to Research Data from Public Funding [EB/OL]. [2013-05-16] http://www.oecd.org/dataoecd/9/61/38500813.pdf.

[4] The Royal Society. Science as an open enterprise [EL/OL]. [2012-06-21]. http://royalsociety.org/policy/projects/science-public-enterprise/report/

[5] Huang Yongwen, Zhang Jianyong, Huang Jinxia, Wang Fang. Review of Foreign Open Data Research [J]. *New Technology of Library and Information Service*. 2013(5): 21-27

[6] RCUK. RCUK common principles on data policy [EB/OL]. [2015-01-14]. http://www.rcuk.ac.uk/research/datapolicy/.

[7] Dong Kun, Gu Liping. Research on the Framework of National Scientific Research Data Open Policies [J]. *China Science and Technology Resources Review*, 2016, 48(3): 51-57.

[8] STM Members List [EL/OL]. [2017-04-10] http://www.stm-assoc.org/membership/our-members/

[9] The World' s 52 Largest Book Publishers, 2016 [EL/OL]. [2016-08-26] https://www.publishersweekly.com/pw/by-topic/international/international-book-news/article/71268-the-world-s-52-largest-book-publishers-2016.html

[10] Research data policies and services [EL/OL]. [2017-04-10] http://www.springernature.com/gp/group/data-policy/

[11] Research Data Policy Types [EL/OL]. [2017-04-10] http://www.springernature.com/gp/group/data-policy/policy-types

[12] Over 600 Springer Nature journals commit to new data sharing policies [EL/OL]. [2016-12-06] http://www.springernature.com/br/group/media/press-releases/over-600-springer-nature-journals-commit-to-new-data-sharing-policies/11111248

[13] Research data policy [EL/OL]. [2017-04-10] https://www.elsevier.com/about/open-science/research-data

[14] Make your research data citable [EL/OL]. [2017-04-10] https://data.mendeley.com/

[15] Wiley's Data Sharing Service [EL/OL]. [2017-04-10] https://authorservices.wiley.com/author-resources/Journal-Authors/licensing-and-open-access/open-access/data-sharing.html

[16] When do we share information? [EL/OL]. [2017-04-10] https://www.oup.com.au/privacy-policy

[17] The Open Data movement: international consolidation [EL/OL]. [2014-07-14] http://blog.scielo.org/en/2014/07/14/the-open-data-movement-international-consolidation/#.WPOvArglG00

[18] Data Availability at Hindawi [EL/OL]. [2017-04-10] https://about.hindawi.com/opinion/data-availability-at-hindawi/

[19] Open Data [EL/OL]. [2017-04-10] https://www.biomedcentral.com/about/policies/open-data

[20] Research Data Policies FAQs [EL/OL]. [2017-04-10] http://www.springernature.com/gp/group/data-policy/faq

[21] Recommended Repositories [EL/OL]. [2017-04-10] http://www.springernature.com/gp/group/data-policy/repositories

[22] Open Data [EL/OL]. [2017-04-10] https://www.elsevier.com/authors/author-services/research-data/open-data

[23] Data access. [EL/OL]. [2017-04-10] http://www.re3data.org/search?query=

[24] PLOS Data Policy [EL/OL]. [2014-03-03] http://journals.plos.org/plosone/s/file?id=c4aa/PLOSONE_data

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