

Post-print: Survey and Analysis of Scientific Data Management in International Journals

Authors: Jiang Hong, Liu Jingyi

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] To provide insights and references for domestic journals to implement scientific data management, and to propose a framework for journal scientific data management along with sound development recommendations based on the current domestic landscape. [Method/Process] Through survey research, 15 foreign journals with exemplary scientific data management practices were selected as research subjects, and their characteristics in four aspects were analyzed: scientific data management policies, data submission procedures, repository platforms, and open access. [Results/Conclusion] The following implications are drawn: formulate comprehensive scientific data management policies; establish detailed scientific data submission workflows; integrate with specialized scientific data repository platforms; actively promote open access to scientific data, among other recommendations.

Full Text

Foreign Journal Scientific Data Management: Investigation and Analysis

Hong Jiang¹, Jingyi Liu^{1,2}

¹Wuhan Library, Chinese Academy of Sciences, Wuhan 430071

²Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract

[Purpose/Significance] This study provides insights and references for domestic journals to implement scientific data management, and proposes a framework for journal scientific data management along with reasonable development suggestions based on the current domestic situation. [Method/Process] Through

investigation, 15 foreign journals with well-established scientific data management practices were selected as research objects, and their characteristics were analyzed in four aspects: scientific data management policies, data submission requirements, repository platforms, and open access. **[Result/Conclusion]** The findings offer the following recommendations: formulate comprehensive scientific data management policies; establish detailed scientific data submission workflows; associate with professional scientific data repository platforms; and actively promote open access to scientific data.

Keywords: journal; scientific data; scientific data management; framework

2. Research Status at Home and Abroad

2.1 Foreign Research Status

Foreign theoretical research on journal scientific data management began relatively early, and with the establishment of relevant models in practice and the exploration of an increasing number of journals, theoretical discussions have become more comprehensive and specific. These can be mainly divided into three aspects: (1) Research on relevant systems: For example, K.R. Wilhemus proposed through investigation and analysis that domain journals should mandate authors to submit datasets [3]; G. Santori suggested actively promoting data reproducibility and signing relevant declarations [4]. (2) Research on existing theoretical models: For instance, P. Sturges presented and described models related to journal scientific data management such as the “Journal Data Sharing Policy Model” [5]. (3) Research on specific journals: For example, N.P. Gleditsch et al. used the *Journal of Peace Research*, which publishes datasets in the form of “special data features,” as a case study for elaboration and analysis [6]. In terms of practical implementation, many well-known foreign journal publishers such as Elsevier [7] and Wiley [8] have put forward corresponding scientific data management requirements for articles submitted by researchers to promote the preservation and utilization of scientific data. The International Association of STM Publishers also established a data publishing pyramid model to provide a common approach for relevant personnel to make data accessible. This pyramid model consists of four layers, with meanings from bottom to top as follows (where and both belong to the second layer): Data stored on authors’ and institutions’ hard drives; Data cited from articles and preserved in data centers and repositories; Description of dataset availability during data publication; Further explanation of data in the form of supplementary files; Data included and explained in articles. The specific model is shown in Figure 1 [Figure 1: see original paper] [9].

2.2 Domestic Research Status

In China, some scholars have also recognized the importance of scientific data management in journals. However, since domestic journal scientific data management is still in its infancy, existing theoretical research mostly selects rep-

representative foreign journals as research objects to explore and analyze their characteristics from different perspectives, thereby providing references for domestic journals. This research can be mainly divided into two categories: (1) Discussion on policies supporting journal scientific data management: For example, Huang Ruhua and Li Nan investigated the policies of five foreign data journals and proposed that China should establish policies based on the research data lifecycle [10]; Han Mingyang and Jiang Xin studied BioMedCentral and concluded that relevant Chinese journals should strengthen data standardization [11]. (2) Analysis of specific aspects of journal scientific data management, such as data citation [12], or analysis of the overall process to identify existing problems and specific areas needing improvement in Chinese journal scientific data management [13]. In terms of practical implementation, domestic journals lag behind foreign journals in promoting scientific data management and are still in the initial stage. *Data Analysis and Knowledge Discovery* (formerly *Modern Library and Information Technology*) began requiring authors to submit scientific data supporting their conclusions in December 2016, with appropriate data sharing mechanisms [14]. In the same month, *Library Journal* launched a data management platform (test version) to explore functions such as storage and management of journal article data, which officially entered trial operation in 2017 [15]. Overall, foreign journals started scientific data management work earlier than domestic journals and have relatively mature development, with more theoretical research in this area that is worthy of study and reference by domestic researchers.

This article selects 15 foreign journals, including G3 (abbreviation for Geochemistry, Geophysics, Geosystems), BioMed Central, Geology, Geosphere, GSA Bulletin, Lithosphere, Environmental and Engineering Geoscience, PLOS ONE, Earth System Science Data, Molecular Ecology, PANGAEA journals, Nature, Geoscience Data Journal, Scientific Data (a pure data journal), and Elsevier journals, as research objects to explore and analyze the current status of their scientific data management from four aspects: policies, submission requirements, repository platforms, and open access, with the aim of providing food for thought for domestic journals to carry out related work.

3. Current Status of Foreign Journal Scientific Data Management

3.1 Scientific Data Management Policies

The establishment of relevant policies facilitates effective management of scientific data by stakeholders including journal managers, data producers, relevant institutions, and funders, promoting the maximum value of scientific data at every stage of its lifecycle and thereby advancing scientific communication and development. The scientific data management policies of the investigated journals are compiled into a table (see Table 1, sorted by the publication date of relevant policies from earliest to latest).

As shown in Table 1, all investigated journals have relevant scientific data management policies. Among them, G3 (established in 1999 [29]) adopts the AGU Publications Data Policy formulated by the American Geophysical Union (AGU), which was adopted by the publishing association as early as 1993, indicating that both policy makers and adopters recognized the importance and necessity of scientific data management in journals at an early stage. In terms of content elements, all policies include four basic aspects: data preservation, sharing, access, and reuse. Specifically, journals such as *Geoscience Data Journal* and *Scientific Data* provide clear explanations of data citation requirements and principles, and set up “quick Q&A” sections and editorial contact information on their websites to answer authors’ questions, thereby further improving and ensuring more effective scientific data management. Additionally, to serve authors in different fields, some policy makers collaborate with journals to provide the most suitable policies. For example, Springer Nature divides its research data policy into four types according to the scale and website style of its journals [30], aiming to provide clear guidance for authors in different fields (see Table 2 for details).

3.2 Scientific Data Submission Requirements

3.2.1 Data Description To improve the transparency and usability of datasets related to papers, journals generally require authors to provide complete descriptions of submitted data and offer data description templates for download to help draft and complete relevant content. The investigation found that the Data Descriptor content specification from DataCite is widely accepted and adopted by many journals. It requires that data descriptor manuscripts should include nine elements: Title: maximum 110 characters including spaces; Abstract: maximum 170 words, no references; Background and summary: maximum 700 words; Methods: no length limit; Data records: no length limit; Technical validation: no length limit; Usage notes: no length limit; Figures: usually no more than 3 per manuscript; Tables: usually no more than 3 per manuscript [31]. Additionally, some journals have specific requirements. For example, *Geoscience Data Journal* requires authors to provide details on data collection, processing, and file formats in papers describing datasets, and recommends using the journal’s online system to complete the drafting and submission of relevant content [26].

3.2.2 Availability Statement The investigation found that most journals require authors to submit a data availability statement along with the data. The statement should describe key points including data accessibility, accession numbers, applicable conditions, and restrictions, thereby ensuring that data in the manuscript can be properly cited and accessed by other researchers. Table 3 lists the requirements of some journals for data availability statements. Notably, *Nature* has adopted a guidance policy specifically for data availability—the “Data Availability Statements and Data Citation Policy: Guidance for Authors”—to better help authors complete the submission of data availability statements in

their manuscripts. This policy also includes detailed examples to assist authors in further understanding its content.

3.3 Scientific Data Repository Platforms

All investigated journals encourage authors to deposit data in appropriate repositories when submitting data and assign DOIs to make data more discoverable, reusable, and preserved for the long term, thereby further expanding their influence. For example, GSA encourages all its authors to permanently preserve new data used in published papers in trusted repositories and lists five benefits to promote this practice: Maintaining open access to data; Providing long-term preservation; Using persistent and unique identifiers; Enabling metadata registration; Ensuring quality assurance [18]. Table 4 compiles some of the scientific data repository platforms most recommended by the investigated journals.

The five data repository platforms listed in Table 4 span different disciplinary fields, providing services such as data preservation, open access, DOI assignment, proper citation, and discovery for journal authors submitting data. Notably, Dryad states that when older versions of preserved data are abandoned, long-term preservation will be achieved by migrating to generic file formats, and it cooperates with DataONE to ensure users have indefinite access to data content. Additionally, Mendeley Data proposes to provide data submitters with progress-sharing services, encouraging researchers to share unpublished data with peers or funding agencies to advance the research process, and submitters have complete control over who can access and download the submitted data.

3.4 Scientific Data Open Access

Journals generally advocate and require authors to make submitted scientific data openly accessible through two approaches: developing their own open access principles and instructions, or adopting recognized relevant principles and licenses, thereby promoting data flow, discovery, and citation.

3.4.1 Self-Developed Open Access Principles and Instructions In its 2012 position statement, AGU mentioned that “Earth and space science data should be widely accessible in multiple formats, and long-term data preservation is the responsibility of scientists and sponsoring agencies.” To follow this statement and promote scientific exploration and discovery, while allowing all researchers to comprehensively evaluate all results in AGU publications, AGU requires that all necessary data must be understandable, assessable, and reproducible, and that data underlying reported research must be available and accessible. Therefore, AGU encourages authors to identify and preserve their data in certified data repositories or centers. If no relevant public repository is available and such data cannot be easily included in a supplementary file, authors are expected to manage the data for five years after publication and provide a transparent process to make the data available to anyone upon request. Authors must also include a clear statement in the paper’s “Acknowledgments” section

explaining how other users can obtain the data from the article (through supplementary files, repositories, or other sources) and describing any restrictions on access. AGU publications do not cite datasets that are not managed or cannot be reliably provided to anyone requesting data use [16]. *PLOS ONE* requires data submitters to agree to share data, relevant metadata, and methods; if they refuse, the journal has the right to reject the submission. Two situations require special attention: If authors did not collect the data themselves but used other sources, those sources must be properly documented; If restrictions on data access are exposed after publication, the journal reserves the right to issue corrections, contact the author's institution and funders, or in extreme cases retract the publication [19].

3.4.2 Adoption of Recognized Principles and Licenses In terms of open access, BioMed Central adopts the Panton Principles for Open Data in Science, requiring authors to waive copyright and related rights for published data so that others can reuse the data without special permission. Under normal circumstances (unless otherwise stated in individual articles), all data in open access articles published in the journal must follow the Creative Commons CC0 1.0 Public Domain Dedication waiver. Although there is no legal requirement, anyone reusing data published in the journal should cite the data source in derivative works as much as possible [17]. *Molecular Ecology* (following Wiley's requirements) hopes that authors will sign an Open Access Agreement to provide a Creative Commons license (Wiley uses CC BY, CC BY-NC, and CC BY-NC-ND) for published works to comply with open access publishing terms and ensure the widest possible dissemination of published works [39].

4. Recommendations for Domestic Journal Scientific Data Management

Scientific data plays a vital role in today's research environment, and an increasing number of international journals require authors to submit scientific data related to their papers to promote more effective scientific communication. This indicates that data publishing is an important way to make data accessible, and domestic journals should follow this international trend by actively carrying out scientific data management and promoting data publishing. Compared with domestic journals, foreign journals have more mature development and a more complete scientific data management system, which is worthy of reference and learning. Based on the characteristics of the research objects in scientific data management, this article has drawn a framework for journal scientific data management (Figure 2 [Figure 2: see original paper]) and proposes recommendations for domestic journals to implement scientific data management in combination with this framework.

4.1 Formulate Comprehensive Scientific Data Management Policies

When implementing scientific data management, journals need to formulate comprehensive policies to support the entire workflow, ensuring that every aspect has rules to follow and policies to rely on, while protecting the rights and interests of journals, data submitters, and other stakeholders. A complete scientific data management policy should include: **Preservation:** As an important product of scientific endeavors, data should be effectively preserved for long-term reuse. Data submitters should select appropriate repositories (recognized by the journal) for data storage, and journals have the responsibility to guide submitters in preserving data and recommend guaranteed and suitable data repositories. **Sharing:** In an open science environment, data sharing can better promote data flow and scientific communication. Data submitters should be aware that data sharing means their scientific research products can be shared with other researchers to create benefits and promote the development of related scientific fields; journals should play a supervisory role in urging submitters to share data with others to help advance the scientific research process. **Access:** Data accessibility ensures that data is effectively available, facilitating verification and use by other researchers and enhancing data value. Submitters should inform other researchers how to access the submitted data (e.g., through data repositories or relevant links), thereby saving time for other researchers in obtaining data for scientific research; journals should review whether submitted data can be effectively accessed. **Reuse:** Data reuse requires strict citation regulations to ensure the reasonable use of others' scientific products. Submitters should ensure data reusability, meaning submitted data must be authentic and well-documented; journals should adopt internationally accepted citation regulations to standardize the reuse process, such as the JDDCP [21]. The above four aspects should be formulated around the data lifecycle to create a complete scientific data management policy. Once established, all stakeholders must strictly comply to ensure that data can maximize its value at every stage.

4.2 Establish Detailed Scientific Data Submission Workflows

Scientific data should be submitted as part of or an appendix to the paper. Therefore, journals should establish detailed submission workflows to guide submitters in effective submission, while avoiding overly cumbersome processes that may affect the quality of completion. Journals should also equip dedicated personnel or set up online submission systems (with appropriate examples) to help submitters complete this work. The workflow should primarily include: **Clear data description:** Clear data description facilitates other researchers' understanding and selection of data. Description elements should cover title, abstract, background, methods, collection and processing, credibility verification, relevant figures and tables, file formats, etc. Journals can also set character limits for each element according to their own characteristics to avoid overly lengthy data descriptions, referring to DataCite's requirements for data description. **Completing availability statements:** The availability statement facilitates other

researchers in obtaining and citing the data. It should be placed at the end of the manuscript as a required component, containing accession numbers or DOIs through which other researchers can obtain the data, and specifying applicable conditions and restrictions to ensure more effective data reuse, referring to the data availability statements adopted by *PLOS ONE* and *Nature*. **Submission verification:** Journal personnel or online submission systems need to verify submitted data information to prevent omissions that may affect subsequent sharing and citation.

4.3 Associate with Professional Scientific Data Repository Platforms

Long-term preservation of data is essential for scientific development, as it helps data producers store large amounts of scientific data and conduct related management work. Currently, many well-established international repositories use advanced technology to achieve long-term preservation and large-capacity storage, providing great convenience for data producers. To better implement scientific data management, journals should cooperate with appropriate repositories or recommend recognized repositories to data submitters for reasonable selection. It is recommended to make data deposition in a repository a necessary condition for publishing papers supported by the data. When associating with scientific data repository platforms, journals should: Ensure that the repository can effectively achieve long-term preservation and secure storage, preferably through timely migration to generic file formats, such as Dryad; Understand the disciplinary fields accepted by the repository, and recommend corresponding repositories based on the journal's discipline, such as EarthChem [40] for earth and environmental sciences, Protein Data Bank for molecular biology, and general repositories like figshare or Harvard Dataverse [41] for journals covering broad disciplines; Clearly understand the relevant principles of associated repositories, with journals being more familiar than submitters with repository requirements and principles to better make recommendations and improve the efficiency of data storage.

4.4 Actively Promote Open Access to Scientific Data

Today's research environment has become more open, and journals should adapt to this trend by actively promoting open access to scientific data to improve researchers' timeliness in obtaining information, accelerate the scientific research process, and expand authors' influence to a certain extent. To better promote open access, the primary and most difficult issue for journals to address is intellectual property and ownership rights. Based on current international practices, it is recommended that journals require authors to waive copyright and related rights to facilitate unrestricted data reuse by other researchers and achieve complete open access. However, authors may request to set access restrictions for a certain period (with specific reasons listed and approved by editors), such as prohibiting public access to data for one year after publication. Exceptions include sensitive information involving human subjects or endangered species locations,

where journals may stipulate conditions for data access or decide whether to make it public. Additionally, it is recommended that journals develop or adopt relevant policies and principles for data submitters regarding open access, such as the Panton Principles for Open Data in Science, and sign open access agreements and intellectual property licenses (CC BY, CC BY-NC, CC BY-NC-ND, etc.) to ensure smooth implementation of open access, promote widespread dissemination of scientific data, and advance scientific endeavors.

References

- [1] ICSU World Data System. Strategic plans [EB/OL]. [2018-09-13]. <http://www.icsu-wds.org/publications/strategic-plans/strategic-plan-2014-2018>.
- [2] General Office of the State Council. Notice of the General Office of the State Council on issuing the Measures for the Management of Scientific Data [EB/OL]. [2018-09-14]. http://www.gov.cn/zhengce/content/2018-04/02/content_{5279272}.htm.
- [3] WILHELMUS K R. The analysis of clinical research: mandatory submission of datasets journals should have access to research data [J]. Archives of ophthalmology, 2000, 118(2): 273-274.
- [4] SANTORI G. Journals should drive data reproducibility [J]. Nature, 2016, 535(7612): 355.
- [5] STURGES P, BANKS M, ANDERS J H S, et al. Research data sharing: developing a stakeholder-driven model for journal policies [J]. Journal of the Association for Information Science & Technology, 2015, 66(12): 2445-2455.
- [6] GLEDITSCH N P, NORDÅS R, URDAL H. Peer review and replication data: best practice from journal of peace research [J]. College & research libraries, 2017, 78(3): 267-271.
- [7] Elsevier. Publishing with Elsevier: step-by-step [EB/OL]. [2018-09-18]. <https://www.elsevier.com/authors/journal-authors/submit-your-paper>.
- [8] Wiley. Open data [EB/OL]. [2018-09-18]. <https://authorservices.wiley.com/open-science/open-data/index.html>.
- [9] STM. Integration of data and publications [EB/OL]. [2018-09-21]. <https://www.stm-assoc.org/standards-technology/integration-of-data-and-publications/>.
- [10] Huang Ruhua, Li Nan. Research on foreign data journal policies based on the data lifecycle model [J]. Library and Information, 2017(3): 36-42+108.
- [11] Han Mingyang, Jiang Xin. Investigation and analysis of international open access journal data policies [J]. Journal of Intelligence, 2017, 36(8): 161-165.

- [12] Chi Yuzhuo, Wang Yanfei. A content analysis framework for scientific data citation in scientific data management [J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(1): 43-51.
- [13] Liu Can, Wang Ling, Ren Shengli. Analysis of the current status and trends of data journals [J]. Acta Editologica, 2018, 30(4): 344-349.
- [14] Editorial Office of *Data Analysis and Knowledge Discovery*. Supporting data submission requirements [EB/OL]. [2018-09-17]. http://manu44.magtech.com.cn/Jwk_{{infotech}}_{{wk3}}3467/NEWS/20161213090914.pdf.
- [15] *Library Journal* Data Publishing Management Platform. About us [EB/OL]. [2018-09-17]. <http://data.libraryjournal.com.cn/home/static/about.jsp>.
- [16] AGU. AGU publications data policy [EB/OL]. [2018-09-18]. <https://publications.agu.org/author-resource-center/publication-policies/data-policy/>.
- [17] BMC. Open data [EB/OL]. [2018-09-18]. <https://www.biomedcentral.com/about/policies/open-data>.
- [18] The Geological Society of America. GSA data policy for publications [EB/OL]. [2018-09-18]. <https://www.geosociety.org/gsa/pubs/dataPolicy.aspx>.
- [19] PLOS ONE. Data availability [EB/OL]. [2018-09-18]. <https://journals.plos.org/plosone/s/data-availability>.
- [20] FORCE11. Joint declaration of data citation principles [EB/OL]. [2018-09-18]. <https://www.force11.org/datacitationprinciples>.
- [21] Earth System Science Data. Data policy [EB/OL]. [2018-09-18]. https://www.earth-system-science-data.net/about/data_{{policy}}.html.
- [22] Wiley. Wiley's data sharing policies [EB/OL]. [2018-09-18]. <https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/data-sharing-policy.html>.
- [23] WILKINSON M D, DUMONTIER M, AALBERSBERG I J, et al. The FAIR guiding principles for scientific data management and stewardship [J]. Scientific data, 2016, 3: 167-172.
- [24] PANGAEA. About [EB/OL]. [2018-09-18]. <https://www.pangaea.de/about/>.
- [25] Nature. Availability of data, materials, code and protocols [EB/OL]. [2018-09-18]. <https://www.nature.com/authors/policies/availability.html#data>.
- [26] Geoscience data journal. Author guidelines [EB/OL]. [2018-09-18]. <https://rmets.onlinelibrary.wiley.com/hub/journal/20496060/about/author-guidelines>.
- [27] Scientific data. Data policies [EB/OL]. [2018-09-18]. <https://www.nature.com/sdata/policies/data-policies>.
- [28] Elsevier. Research data [EB/OL]. [2018-09-18]. <https://www.elsevier.com/about/policies/research-data>.

- [29] Yuan Honglin. International geoscience SCI journal *Geochemistry, Geophysics, Geosystems* [J]. Rock and Mineral Analysis, 2012, 31(5): 928.
- [30] Springer Nature. Research data policy types [EB/OL]. [2018-09-19]. <https://www.springernature.com/gp/authors/research-data-policy/data-policy-types/12327096>.
- [31] Scientific data. For authors [EB/OL]. [2018-09-19]. <https://www.nature.com/sdata/publish/for-authors#format>.
- [32] Molecular ecology. Author guidelines [EB/OL]. [2018-09-19]. <https://onlinelibrary.wiley.com/page/journal>.
- [33] Nature. Data availability statements and data citations policy: guidance for authors [EB/OL]. [2018-09-19]. <https://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf>.
- [34] Figshare. Key benefits [EB/OL]. [2018-09-20]. <https://figshare.com/services/publishers>.
- [35] Dryad. The repository: key features [EB/OL]. [2018-09-20]. <https://datadryad.org/pages/repository>.
- [36] Mendeley Data. Discover Mendeley data [EB/OL]. [2018-09-20]. <https://data.mendeley.com/>.
- [37] PDB. Features & highlights [EB/OL]. [2018-09-20]. https://www.rcsb.org/#Subcategory-deposit_{deposit}-data.
- [38] ScienceDaily. GenBank [EB/OL]. [2018-09-20]. <https://www.sciencedaily.com/terms/genbank.htm>.
- [39] Wiley. Wiley's best practice guidelines on publishing ethics copyright and intellectual property [EB/OL]. [2018-09-21]. <https://authorservices.wiley.com/ethics-guidelines/copyright-and-intellectual-property.html>.
- [40] EarthChem. EarthChem overview [EB/OL]. [2018-09-25]. <http://www.earthchem.org/overview>.
- [41] Harvard Dataverse. Subject: Medicine, Health and Life Sciences [EB/OL]. [2018-09-25]. [https://dataverse.harvard.edu/dataverse/harvard?q=&fq0=subject_{ss}%3A\"Medicine%2C+Health+and+Life+Sciences\"&types=dataverses%3Adatasets&sort=dateSort&order=desc](https://dataverse.harvard.edu/dataverse/harvard?q=&fq0=subject_{ss}%3A\).

Author Contributions:

Hong Jiang: Paper topic selection and design, proposed revision suggestions;

Jingyi Liu: Paper framework conception, writing and revision.

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

Source: ChinaXiv – Machine translation. Verify with original.