Postprint: Recent Advances in Domestic and International Scientific Data Management and Open Sharing

Authors: Zhang Lili, Wen Liangming, Shi Lei, Xiaohuan Zheng, Jianhui Li

Date: 2023-03-19T00:00:00+00:00

Abstract

As the soul of scientific research activities, scientific data serves as both the starting point for stimulating scientific innovation and an indispensable component of the rich outcomes of research endeavors. Scientific data management and sharing, both domestically and internationally, are pursued through two dimensions: an "actively moderate orientation in scientific data policy" and "comprehensive and meticulous practices in scientific data management." Through a comparative analysis of current developments at home and abroad, the author argues that the macro-level framework of domestic scientific data policies remains to be expanded, and the implementation of these policies still requires continued accumulation of experience; numerous disciplinary fields still need to enhance their awareness and capabilities in data management; while the overall environment for scientific research is conducive to fostering a culture of open scientific data, comprehensive coordination among multiple stakeholder groups remains necessary. Looking ahead, the actively moderate sharing trend will continue to dominate the mainstream, debates over the public and private rights of scientific data will intensify, and the re-conceptualization of the boundaries between information technology transformation and scientific data management will persistently drive data openness.

Full Text

Scientific Data Management and Open Sharing: Policies and Mechanisms

As the lifeblood of research activities, scientific data serves as both the starting point for stimulating scientific innovation and an indispensable component of research outcomes. Domestic and international efforts in scientific data management and sharing have developed along two dimensions: "proactive yet mea-

sured scientific data policy guidance" and "comprehensive and meticulous scientific data management practices." Through comparative analysis of current developments at home and abroad, we observe that China's scientific data policy framework remains in need of expansion, with policy implementation requiring continued accumulation. Many disciplines still need to enhance their awareness and capabilities in data management. While the overall research environment is conducive to nurturing an open scientific data culture, comprehensive coordination among multiple stakeholder groups remains essential. Looking ahead, the proactive and measured trend toward sharing will continue to dominate, debates over public versus private rights in scientific data will intensify, and technological transformations coupled with renewed understanding of scientific data management boundaries will continuously drive data openness.

Keywords: scientific data management, scientific data sharing, scientific data policy, open data

Regardless of research field or stakeholder group, effective scientific data management and open sharing benefit scientific research, the broader public, and individuals alike: they drive scientific progress, reduce redundant labor while boosting productivity, establish efficient boundaries for science policy, advance long-term research and education, bring new solutions to societal problems, and shorten product incubation cycles while satisfying public information demands. However, in complex research contexts, data cannot be treated and managed simply as a knowledge commons, as its effective flow requires additional incentives, quality controls, and more complex strategic choices and balancing mechanisms. A better grasp of global trends in scientific data management and sharing helps us identify and analyze problems, compare and reflect on current conditions, and form reasonable expectations for the future. Through extensive investigation, we have organized relevant topics in scientific data management and sharing (Table 1).

2.1.1 FAIR Principles and "Full Open" Models

Since 2000, international organizations such as the Organisation for Economic Co-operation and Development (OECD), Group on Earth Observations (GEO), and the Committee on Data for Science and Technology (CODATA) have promoted "full and open" scientific data sharing policies aimed at facilitating free, unrestricted cross-boundary flow and reuse of scientific data resources. In 2014, a multi-stakeholder academic workshop titled "Jointly Building a Fair Port for Data" in Leiden, Netherlands, proposed the FAIR Guiding Principles—Findable, Accessible, Interoperable, and Reusable—that further interpreted the fundamental philosophy of modern scientific data sharing and rapidly gained popularity. The FAIR principles categorize scientific data resources into six types based on openness status, with four categories considered primary forms of open data: "FAIR metadata," "FAIR limited-open data," "FAIR open data," and "FAIR enhanced open data." These principles have spread throughout the European Union, United States, Australia, and beyond. Research on measuring

FAIR data assets and the FAIR movement themed "Go change, Go build, Go train" have further advanced implementation of these principles.

2.1.2 The Dynamically Evolving Boundary of Scientific Data Openness

The depth and breadth of scientific data sharing—that is, the boundary of openness—are constrained by factors including economic interests (such as data ownership and intellectual property rights), privacy rights, and public safety, as clearly stated in the Royal Society's report *Science as an Open Enterprise*. After four years of preparation, the European Union's General Data Protection Regulation (GDPR) was approved on April 14, 2016, and formally implemented on May 25, 2018, aiming to protect European citizens from privacy data breaches in the data era. Its core provisions establish five citizen rights: the right to be informed, right of access, right to object, right to data portability, and right to be forgotten, making it the most important data privacy regulation in nearly two decades. Due to the complexity of research contexts, the rights confirmation of data assets warrants continuous exploration. The evolving boundary of scientific data openness represents a dynamic game as scientific data moves from closed to open contexts. Boundary delineation will continue to be a focal point and challenge for scientific data sharing, requiring responses from both top-down administrative mandates and bottom-up frontline data production, particularly through technological applications, training and education, citizen science development, and comprehensive impact measurement.

2.1.3 The Maturing Comprehensive Policy System

From an organizational perspective, the policy system for scientific data management and sharing has expanded comprehensively (Figure 1 [Figure 1: see original paper]), spanning international and national levels, as well as regional, disciplinary, and institutional levels or smaller organizational units. Among these, domain and institutional-level scientific data policies are closer to actual research and data contexts, making them crucial forces for extending and enriching the entire policy chain. Beyond vertically integrated policy system construction, connections between different policy levels are increasingly close, such as international policies interfacing with national policies through data diplomacy, and domain/institutional policies adjusting and calibrating in response to national policies. Some organizations have archived existing data policies: the U.S. Department of Energy's Systems Biology Knowledgebase (KBase) includes bioinformatics data policy resources primarily from the United States; the FAIRsharing platform has collected metadata information on 112 data policies across multiple domains; and the EU and OECD jointly established the STIPCompass database to collect and publish science and technology policies from 51 countries including China, covering scientific data management policies.

2.2.1 Data Management Plans: From Concept to Practice

In 1995, the UK's Economic and Social Research Council (ESRC) established Data Management Plans (DMPs), requiring that data generated from ESRC-funded research be shared as much as possible with proper long-term preservation and high-quality management. The U.S. National Science Foundation (NSF) mandated in January 2011 that project proposals must include data management plans. In recent years, data management has gradually moved from paper plans to practice, focusing on data types, data or metadata format and content standards, access and sharing/reuse policies, and data archiving plans. Numerous libraries, scientific data centers, research institutions, government agencies, and international and regional organizations have participated in technical support, policy interpretation, and training for DMP implementation.

2.2.2 Emerging Technologies Driving Continuous Innovation

Examples of emerging technologies driving scientific data openness and sharing are numerous. Here we highlight three areas: blockchain-enabled data sharing, citizen science-driven data production, and human-machine network interoperability promoted by the Data Documentation Initiative (DDI).

(1) Blockchain-Enabled Data Sharing. The multi-level evolution and pipeline processing characteristics of big scientific data throughout its lifecycle pose new challenges for data transmission, processing, and sharing. Blockchain technology offers solutions: using encryption algorithms and consensus mechanisms to ensure security, tracing origins and "filtering" to guarantee data quality, and distributed decision-making to remove intermediaries, thereby significantly improving data sharing efficiency. Healthcare data has experimented with using blockchain to store and share personal health data. Additionally, distributed edge computing will play a greater role in rapidly achieving data collection, processing, and analysis through blockchain integration.

(2) Citizen Science-Driven Data Production. As a new source of data collection, citizen science is flourishing. Over the past 22 years, nearly 30,000 whale shark images provided by ecotourists have helped researchers effectively identify 20 whale shark aggregation sites. The value of citizen science data is substantial: the Citizen Science Association (CSA) now has members from over 80 countries, with more than one million volunteers participating in over 1,000 major scientific programs.

(3) Human-Machine Network Interoperability. To promote understandability between human and machine networks, the DDI Alliance launched DDI 3.3, covering technical content such as classification management, non-survey data collection, sampling and weighting, questionnaire design, supporting DDI as a property graph, and quality statement optimization, primarily applied to archiving, discovery, and interoperability guidance for data in sociology, behavioral science, economics, and public health.

2.2.3 Data Publishing and Trusted Repositories

Data publishing provides new platforms for open scientific data management. Publishing datasets and data papers has become popular in recent years, with practices such as *Earth System Science Data* (2008), *GigaScience* (2012), *Nature Scientific Data* (2015), and *China Scientific Data* (2015). In a broader sense, data publishing also includes data repository construction. Repositories provide storage and access platforms for datasets, supporting standardized data quality control and complete lifecycle management, and are categorized as general-purpose repositories, institutional repositories, domain repositories, publication repositories, libraries/archives/museums, and research project repositories. Trusted repositories, as stable and reliable data infrastructure, provide technical and management resource guarantees for open data work including data publishing.

2.2.4 Flourishing Data Management Training

Data management training guides research practice through practical short-term skill development. The European intergovernmental organization ELIXIR, covering 20 national nodes, comprehensively promotes scientific data management training across Europe. UK domain-specific training includes DCC (general), CAiRO (arts), DataTrain (archaeology, anthropology), DATUM (health), DMTpsych (psychology), and Research Data MANTRA (geoscience, social science, and clinical psychology). CODATA provides annual data management technical training for researchers in developing countries. Data Carpentry, derived from software training, collaborates with multiple countries to conduct training outreach. Additionally, professional degree programs in data science are increasingly flourishing.

2.2.5 Comprehensive Impact Measurement

(1) Starting with Data Citation. Since 2010, CODATA's Data Citation and Practice Working Group has discussed "data citation standards and norms" in detail. In 2014, the American Society for Information Science and Technology (ASIS&T) Data Access and Preservation Summit focused on data citation, metadata, and data reuse. Harvard University's Institute for Quantitative Social Science (IQSS) launched a data citation research project in 2014. University libraries and non-profit organizations (such as DataCite and ICPSR) have also participated in developing and promoting data citation standards.

(2) Altmetrics from a Social Perspective. Altmetrics comprehensively evaluate the social impact of academic achievements based on mass social media, traditional mainstream media, academic social media, blogs, and reference management software (including views, saves, discussions, recommendations, and citations).

(3) Advancing Data Metrics. Moving beyond traditional citations between literature and references to multiple relationships between data and literature,

data and data, and data and datasets, with greater focus on "data," "scholarly records," and "scholarly individuals."

Scientific Data Management and Open Sharing in China

2.3.1 Overview of National Scientific Data Policy System

Scientific data management has accompanied research activities without interruption, particularly flourishing after 2000. China has formed a data policy system led by government, industry institutions, and domain data centers (Table 2). Among these, the *Measures for the Management of Scientific Data* took effect on March 17, 2018. For the first time at the national level and across multiple domains, this regulation proposed open sharing as the guiding principle, marking an epoch-making significance. Additionally, typical sectoral institutional systems include both data management measures and policy guidelines, such as the State Oceanic Administration's *Opinions on Standardizing Marine Ecological Environment Monitoring Data Management Work* (February 2015) and the Ministry of Transport's *Implementation Opinions on Promoting Data Resource Open Sharing in the Transportation Industry* (September 2016). Cross-departmental cooperative sharing has gradually advanced, such as the 2015 data sharing agreement between the State Forestry Administration and Ministry of Land and Resources establishing a long-term sharing mechanism. Domain scientific data centers that parallel data practice with data policy are particularly noteworthy.

2.3.2 Concurrent Scientific Data Open Sharing Practices

Scientific data management has always preceded policy and serves policy, with most domain data policies rooted in data practice. While scientific data management is well-established, open sharing remains exploratory, with related practices still dominated by exchanges and discussions. Although open data demonstration platforms exist, widespread data sharing practices remain to be developed. Figure 2 [Figure 2: see original paper] reviews representative events in China's scientific data open sharing history, with major domestic practices from January 2017 to July 2018 listed in Table 3 . These practices involve government, research institutions, and enterprise/social forces, covering data infrastructure construction, big data project-driven scientific data management and opening, scientific data exchange seminars, and international exchanges and cooperation.

Comparative Reflection and Development Prospects

3.1 China's Scientific Data Policy and Practice Compared with Developed Countries

(1) Development Level. The *Measures for the Management of Scientific Data* was recently promulgated, and its implementation will require years of

exploration and accumulation. Based on existing scientific systems and data resource volumes, the macro-level data policy management system remains to be expanded.

(2) Development Breadth. Domestic scientific data management practices are concentrated primarily in natural and engineering sciences. While there are exemplary social science data practices, such as national statistical data and research institution survey data platforms (e.g., the China Survey and Data Center at Renmin University of China), significant progress remains to be made relative to the scale of disciplinary research activities. The overall level of scientific data open sharing still needs improvement, with data black holes formed by dispersion among individual researchers continuing to exist objectively.

(3) Development Drivers. While scientific data sharing represents the general trend, matching evaluation and incentive mechanisms for data sharing remain immature. Motivation for data sharing work comes primarily from spontaneity or administrative constraints. Better integration of tangible and intangible forces to mobilize stakeholders across the entire lifecycle is crucial to the future of research data management.

3.2 Mainstream Trends in Future Scientific Data Management and Sharing

(1) Proactive and Measured Sharing Will Continue as the Mainstream. From open science to open access to open data and FAIR practices, open data still requires flexible adaptation to different research contexts. For example, the International Committee of Medical Journal Editors (ICMJE), representing 282 clinical researchers from 33 countries, opposed free sharing of clinical trial data within 6 months of publication by 14 medical journals, arguing it was unrealistic. This demonstrates that open data cannot be achieved overnight but requires gradual, proactive, and measured strategies.

(2) Public vs. Private Rights in Scientific Data Will Be Increasingly Contested. Scientific data should be openly shared for public benefit while protecting specific interests from infringement. Therefore, confirming data rights is crucial. Effectively balancing public and private rights requires legal wisdom, technological support (such as technical explorations enabling fine-grained data sharing while reducing privacy infringement risks), and societal understanding and participation in building a sharing culture.

(3) Information Technology Transformation Cannot Be Underestimated. Information and communication technologies have ushered us into a new data era and impact research data assets. Scientific data open sharing depends on technical support while continuously posing new challenges to information technology, such as blockchain applications and the flourishing of citizen science. Embracing new technologies with an open mind is a new tool for driving mature open data management.

(4) Reconceptualizing Scientific Data Management. Mature scientific data management involves not only working with data but also participation from multiple stakeholder groups. Efficient scientific data management activities require clear division of responsibilities, such as professional refinement of institutional data assets and assigning responsibilities to individuals, to ensure data management meets expectations. The foundation for promoting effective scientific data management includes but is not limited to institutional macro-level data management functions, data governance bodies (policy makers and practitioners), team culture, and outcome measurement and evaluation.

In summary, through literature review and practical exchanges, we have summarized major progress in research and practice on scientific data management and sharing both domestically and internationally. Based on comparative analysis, we conclude that China's scientific data management practices are maturing, but macro-level development still requires more accumulation, significant differences in data management levels persist across disciplines, and flexible application of information technology and expansion of scientific data management boundaries will be important driving forces for enhancing scientific data management development.

Acknowledgments

The authors thank the China Scholarship Council for supporting the first author's visiting research in the United States.

References

1. Pfenninger S, De Carolis J, Hirth L, et al. The importance of open data and software: Is energy research lagging behind? Energy Policy, 2017, 101: 211-215.
2. NSF National Science Board. Long-lived digital data collections: Enabling research and education in the 21st century. Washington DC: NSF, 2005.
3. Schrier B. Government open data: Benefits, strategies, and use. The Evans School Review, 2014, 4(1): 12-27.
4. Fecher B, Friesike S, Hebing M. What drives academic data sharing? PLoS ONE, 2015, 10(2): e0118053.
5. Institute of Medicine. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington, DC: The National Academies Press, 2015.
6. Boulton G, Campbell P, FREng B C, et al. Science as An Open Enterprise. London: The Royal Society Science Policy Centre, 2012.
7. Wilkinson M D, Dumontier M, Aalbersberg I J, et al. The FAIR Guiding Principles for scientific data management and stewardship. Nature Scientific Data, 2016, 3: 167-172.

8. European Commission. H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020. [2018-07-01]. http://ec.europa.eu/research/participants/data/ref/h2020/g hi-oa-data-mgt_{en}.pdf.

9. Briney K, Goben A, Zilinsk L. Do You Have an Institutional Data Policy? A review of the current landscape of library data services and institutional data policies. Journal of Librarianship and Scholarly Communication, 2015, 3(2): eP1232.

10. Kanous A, Brock E. Contractual Limitations on Data Sharing. [2016-08-18]. http://deepblue.lib.umich.edu/bitstream/2027.42/123016/1/ContractualLimitationsonDataSharing1 1.pdf.

11. Tumwesigye B T, Nakanjako D, Wanyenze R, et al. Policy development, implementation and evaluation by the AIDS control program in uganda: A Review of the Processes. Health Res Policy, 2013, (11): 7.

12. Wilkinson M D, Dumontier M, Aalbersberg I J, et al. The FAIR Guiding Principles for scientific data management and stewardship. Nature Scientific Data, 2016, 3: 167-172.

13. European Commission. H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020. [2018-07-01]. http://ec.europa.eu/research/participants/data/ref/h2020/g hi-oa-data-mgt_{en}.pdf.

14. KBase. Data policies. [2018-07-01]. https://kbase.us/data-policy-and-sources/.

15. FAIRsharing.org. Policies. [2018-07-01]. https://fairsharing.org/policies/.

16. Economic and Social Research Council. Research Data Policy. [2018-07-01]. https://esrc.ukri.org/funding/guidance-for-grant-holders/research-data-policy/.

17. National Science Foundation. Dissemination and Sharing of Research Results. [2018-07-01]. https://www.nsf.gov/bfa/dias/policy/dmp.jsp.

18. National Science Foundation. Chapter II-Proposal Preparation Instructions. [2018-07-01]. https://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp.

19. Li Jianhui, Shen Zhihong, Meng Xiaofeng. Scientific Big Data Management: Concepts, Technologies, and Systems. Journal of Computer Research and Development, 2017, 54(2): 235-247.

20. Ding Wei, Wang Guocheng, Xu Aidong, et al. Research on Key Technologies and Information Security Issues of Energy Blockchain. Proceedings of the CSEE, 2018, 38(4): 1026-1034.

21. Zhou Zhou. Overview: Citizen Science is Flourishing in the United States. [2018-08-19]. http://www.xinhuanet.com/world/2018-06/14/c_{1122987853}.htm.

22. CSA. The power of Citizen Science. [2018-08-19]. http://www.citizenscience.org/.

23. The DDI Alliance. New DDI 3.3 specification available for public review and comment. [2018-06-30]. http://www.ddialliance.org/announcement/new-ddi-33-specification-available-for-public-review-and-comment.

24. CoreTrustSeal. Requirements. [2018-07-05]. https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_{{{Trustworthy}}}*{{Data}}*{{{Repositories}}}*{{Requirements}}*{{{01

25. ELIXIR. Training services. [2018-07-01]. https://www.elixir-europe.org/services/training.

26. Jisc. CAiRO: Curating Artistic Research Output. [2018-07-01]. https://www.webarchive.org.uk/wayback/archive/20140614073310/http://www.jisc.ac.uk/whatwedo/pr

27. Archaeology Data Service. Data train. [2018-07-01]. http://archaeologydataservice.ac.uk/learning/DataT

28. Mantra. Home. [2018-07-01]. https://mantra.edina.ac.uk/.

29. CODATA. Data Citation. [2018-07-03]. http://codata2012.tw/.

30. ASIS&T. Data Access & Preservation Summit. [2018-07-03]. https://www.asist.org/rdap/past-events/#comments.

31. The Institute for Quantitative Social Science. IQSS Data Science: Aiding Reproducible Research By Adding Provenance in Data Citations. [2018-07-03]. https://www.iq.harvard.edu/news/iqss-data-science-aiding-reproducible-research-adding-provenance-data-citations.

32. DataCite. Cite your data. [2018-07-03]. https://www.datacite.org/cite-your-data.html.

33. ICPS. Find & Analyze Data. [2018-07-03]. https://www.icpsr.umich.edu/icpsrweb/ICPSR/.

34. Wang Dandan. Research on the Application of Altmetrics at the Data Level. Information Studies: Theory & Application, 2018, 41(7): 60-64.

35. Liu Feng, Zhang Xiaolin. Data Publishing and Its Impact on Academic Publishing. Library and Information Service, 2015, 41(2): 56-71.

36. Song Ge, Hu Wenjing. Investigation and Analysis of Foreign Mandatory Open Scientific Data Policies. Library and Information Service, 2016, 60(9): 61-69.

37. Murray T. Researchers oppose data-sharing proposal. Canadian Medical Association Journal, 2016, 188(14): E336.

38. Anne Marie Smith. Foundations of data stewardship. [2018-07-01]. https://www.ewsolutions.com/foundations-data-stewardship/.

ZHANG Lili received her Ph.D. in Management from Peking University and is currently a Senior Engineer at the Computer Network Information Center, Chinese Academy of Sciences (CAS). She serves as a member of the CODATA Data Policy Committee and Deputy Director of the Editorial Office of *China*

Scientific Data. Her research focuses on scientific data management (data stewardship, open data) and information economics. E-mail: zhll@cnic.cn

LI Jianhui is a Professor at the Computer Network Information Center, Chinese Academy of Sciences (CAS), and an Executive Member of CODATA. He is also Director of the Beijing Engineering Laboratory for Big Data Application Service Technologies. Professor Li has long been dedicated to promoting scientific data openness, sharing, and application services. His major responsibilities include constructing the CAS Data Cloud Service System, developing cloud service platforms, and innovating data-intensive scientific applications. His current research focuses on big data resource sharing, management technologies, and computing and analysis technologies. E-mail: lijh@cnic.cn

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