Postprint of Empirical Research on Barriers to Open Sharing of Scientific Data

Authors: Sheng Xiaoping, Wu Hong, Hu Bingjie

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

Abstract

[Purpose/Significance] To summarize and validate perspectives on barriers to scientific data open sharing, thereby promoting theoretical research in this domain. [Method/Process] Based on theoretical analysis of barriers to scientific data open sharing, 64 research hypotheses were extracted and employed as questionnaire items. Through survey administration and data processing and analysis, the relevant research hypotheses were empirically validated. [Results/Conclusion] Barriers to scientific data open sharing primarily encompass 26 types of management and implementation barriers, 13 types of legal and funding barriers, 9 types of technical barriers, 8 types of cognitive barriers, 4 types of cost barriers, and 4 types of user barriers. We should formulate relevant solutions or countermeasures targeting these barriers to substantially enhance the level of scientific data open sharing in our country.

Full Text

Preamble

ChinaXiv Collaborative Journal
Volume 63, Issue 17, September 2019

An Empirical Study on Barriers to Scientific Data Open Sharing

Sheng Xiaoping[1], Wu Hong[2], Hu Bingjie[2]
[1]School of Library, Information and Archives, Shanghai University, Shanghai 200444
[2]School of Economics and Management, South China Normal University, Guangzhou 510006

Abstract

[**Purpose/Significance**] This paper summarizes and validates perspectives on barriers to scientific data open sharing to promote theoretical research in this

domain. [**Method/Process**] Based on theoretical analysis of barriers to scientific data open sharing, we derived 64 research hypotheses and used them as questionnaire items. Through questionnaire surveys, data processing, and analysis, we validated the relevant hypotheses. [**Result/Conclusion**] The barriers to scientific data open sharing primarily include 26 types of management and implementation barriers, 13 types of legal and financial barriers, 9 types of technical barriers, 8 types of cognitive barriers, 4 types of cost barriers, and 4 types of user barriers. We should develop targeted solutions and countermeasures to address these barriers and substantially improve the level of scientific data open sharing in China.

Keywords: scientific data open sharing; sharing barriers; empirical analysis
Classification Number: G203
DOI: 10.13266/j.issn.0252-3116.2019.17.003

Introduction

With the rapid development of Open Access and Open Science movements, scientific data open sharing has emerged as a new topic in the open domain. Although progress has been made in practice—evidenced by policies from organizations such as the OECD, European Research Council, UK Research Council, Wellcome Trust, U.S. National Science Foundation, U.S. Council on Science and Technology, Canadian Social Sciences and Humanities Research Council, and Irish Science Foundation [1], as well as the establishment of major open scientific data repositories like the UN Open Science Database, OECD Library Open Science Data, IMF Database, and Africa Open Data [2], and platforms including the Center for Open Science, BioMed Central, PubMed Central, China Earth System Science Data Sharing Platform, and China Population and Health Science Data Sharing Platform—scientific data open sharing still faces numerous problems and obstacles. These include legal gaps, privacy violations, inadequate IT infrastructure, and imperfect academic reward mechanisms [3-4], as well as cognitive barriers, governance barriers, cost barriers, skill barriers, ethical barriers, and socio-technical barriers [5-8].

Using “scientific data” or “research data” and “open data” as keywords, combined with “open access” or “sharing” and “barriers” (or “impediments”) as parallel keywords, we searched CNKI and full-text databases including Elsevier, Springer, EBSCO, Emerald, Taylor & Francis, SAGE Premier, Wiley, and ProQuest, as well as web literature. The results show that while existing literature has conducted theoretical [9-11] and survey analyses [12] of barriers to scientific data open sharing, empirical research remains scarce. Therefore, this paper aims to propose research hypotheses based on theoretical synthesis of relevant perspectives on barriers to scientific data open sharing, and then validate these hypotheses through questionnaires to identify specific barriers and provide references for overcoming them and promoting practice.

2. Research Hypotheses on Barriers to Scientific Data Open Sharing

Existing research offers various classifications of barrier types: (1) three categories—technical, behavioral/organizational, and legal barriers [9]; (2) six categories—institutional, task complexity, usage/participation, legal, information quality, and technical barriers [7]; or technical, motivational, economic, political, legal, and ethical barriers [11]; or economic, technical, legal, cultural, management, and risk-related barriers [8]; (3) seven categories—governance, economic issues, licensing/legal framework, data characteristics, metadata, access barriers, and skill barriers [6]; (4) ten categories—availability/access, discovery capability, applicability, comprehension, quality, linking/combining data, comparability/compatibility, metadata, interaction with data providers, and open/upload barriers [13].

Synthesizing these perspectives, we analyze barriers to scientific data open sharing from six dimensions: cognitive, economic, legal, technical, management, and implementation.

2.1 Cognitive Barriers and Research Hypotheses

Cognitive barriers refer to misconceptions or concerns that prevent participation in scientific data open sharing. These include: (1) various concerns—fear of having research ideas challenged, disrupting original research plans, being burdened with assistance requests, inability to guarantee data quality/quantity, and reduced competitive advantage [4]; fear of criticism, loss of data control, being scooped, and data misuse [11,14-15]; fear of privacy breaches and losing publication opportunities [16]; (2) cognitive deficiencies—lack of awareness about accessing others' research data [17]; lack of understanding about data's potential uses [9]; ignorance of survey data sources [18]; lack of knowledge for understanding and utilizing scientific data, mining data value, and statistical knowledge [7]; lack of knowledge for creating and preserving metadata [19].

Based on these, we propose 12 hypotheses:
A1: People lack awareness of sharing scientific data.
A2: People fear reduced competitive advantage.
A3: Researchers fear their scientific data being misunderstood.
A4: Researchers fear their scientific data being misused.
A5: Researchers fear criticism for research errors.
A6: Researchers fear disrupting research plans.
A7: Researchers fear interference from others.
A8: Researchers fear losing control of data.
A9: Researchers fear privacy breaches.
A10: Researchers fear losing publication opportunities.
A11: Data users lack knowledge for properly utilizing scientific data.
A12: Data users lack knowledge for creating and preserving metadata.

2.2 Economic Barriers and Research Hypotheses

Economic barriers refer to cost issues (time, labor, financial) or revenue loss. These include: (1) high costs—time costs for collecting, providing, storing, and transmitting data; high labor costs for providing follow-up research; high time/labor costs for using shared data [4,7,15,20]; (2) treating data as fiscal revenue and refusing to share [7]; (3) potential economic losses [11]; (4) lack of funding for processing and publishing data [9].

Based on these, we propose 8 hypotheses:
B1: High time costs for collecting, submitting, and storing scientific data.
B2: High labor costs for collecting, submitting, and storing scientific data.
B3: High time costs for providing, transmitting, maintaining, and managing scientific data.
B4: High labor costs for providing, transmitting, maintaining, and managing scientific data.
B5: Lack of reward/incentive mechanisms for sharing scientific data.
B6: Insufficient funding for scientific data open sharing.
B7: Need to increase research costs beyond budget.
B8: Potential economic losses from scientific data open sharing.

2.3 Legal Barriers and Research Hypotheses

Legal barriers refer to issues in intellectual property, privacy, data security, and protection. These include: (1) IP control, difficulty obtaining sharing licenses, imperfect IP protection, unresolved multi-author data ownership, and litigation risks [9,15,21-22]; (2) privacy protection issues—data sharing may violate respondent privacy [4], providers require privacy removal [15], participant privacy unresolved [20], and broad legal privacy scope [9]; (3) data security concerns—national security considerations preventing disclosure [14] or unresolved security issues [7]; (4) inadequate laws—lack of legal frameworks, contradictory clauses, existing IP/commercial secrecy laws discouraging open science [10], and researchers lacking rights to use open data [19].

Based on these, we propose 7 hypotheses:
C1: Imperfect legal system for scientific data open sharing.
C2: Security issues in scientific data open sharing.
C3: Privacy issues in scientific data open sharing.
C4: Risk of losing ownership and copyright control.
C5: IP issues in scientific data.
C6: Licensing issues in scientific data.
C7: Risk of disputes and litigation.

2.4 Technical Barriers and Research Hypotheses

Technical barriers refer to issues with data submission, storage, sharing, use, and maintenance. These include: (1) data quality/software problems—poor accuracy/timeliness, lack of quality descriptions, missing metadata or lack of agreed

standards [9,13], data mismatches [23], incompatible software, non-standard formats, incomplete/complex data [11,14], and lack of standard software [7]; (2) lack of standards—non-standardized dataset preservation/maintenance [4] and lack of metadata standards [7]; (3) access barriers—difficult access [6], mandatory registration/login [6], lack of public databases [22], and language barriers [9]; (4) processing barriers—lack of technology for large datasets [22], data protection, quality assurance, standardization [18], and effective technical solutions [11]; (5) lack of infrastructure/tools—lack of sharing infrastructure [9], auxiliary tools/systems, search indexes, and platforms [7].

Based on these, we propose 11 hypotheses:
D1: Poor scientific data integrity.
D2: Poor scientific data accuracy.
D3: Poor scientific data timeliness.
D4: Non-standard scientific data formats.
D5: Incompatible data software.
D6: Lack of scientific data open sharing platforms/systems.
D7: Lack of auxiliary tools (indexes, search engines).
D8: Lack of metadata.
D9: Language barriers in international data sharing.
D10: Lack of data protection technology.
D11: Lack of scientific data open sharing standards.

2.5 Management Barriers and Research Hypotheses

Management barriers refer to organizational and managerial issues. These include: (1) leadership/policy gaps—lack of leadership [8], coordination mechanisms [21], and unified policies [24]; (2) procedural obstacles—cumbersome data management procedures and chaotic measures [4,7]; (3) publication barriers—lack of unified publication policies and channels [7]; (4) input barriers—poor user input management and quality assurance [7], lack of metadata quality mechanisms [6]; (5) lack of incentives—ineffective open data incentives [7] and lack of reciprocity between producers and managers [11]; (6) lack of trust—information blockades/bureaucracy [25] and lack of open sharing culture [7].

Based on these, we propose 14 hypotheses:
E1: Lack of effective scientific data open sharing policies.
E2: Lack of coordination mechanisms.
E3: Lack of management mechanisms.
E4: Conflicting management systems.
E5: Restrictive commercial contracts.
E6: Lack of effective sharing plans.
E7: Lack of sharing channels.
E8: Lack of interaction experience with data users.
E9: Lack of user feedback processing.
E10: Lack of incentive mechanisms.
E11: Lack of open sharing organizational culture.

E12: Chaotic data management measures.
E13: Complex data management procedures.
E14: Lack of reciprocity between producers and managers.

2.6 Implementation Barriers and Research Hypotheses

Implementation barriers refer to operational issues in practice. These include: (1) data acquisition barriers—providers restricting access [15] and lack of understanding/judgment about data value [9,26]; (2) linking barriers—difficulty connecting data [13]; (3) conflicting results—similar data producing different results across systems [7]; (4) lack of guidelines [11] and upload difficulties [13]; (5) lack of cooperation—inter-institutional sharing difficulties [9]; (6) lack of professionals [24].

Based on these, we propose 12 hypotheses:
F1: Lack of trust between providers and users.
F2: Lack of scientific data open sharing guidelines.
F3: Data providers restrict data use.
F4: Lack of cooperation within/between institutions.
F5: Market regulation failure for scientific data.
F6: Uneven development of open sharing across countries.
F7: Small proportion of open shared data.
F8: Scientific data monopolization.
F9: Lack of professionals in data resource sharing.
F10: Users lack ability to judge/understand data value.
F11: Users lack data acquisition skills.
F12: Users lack data processing skills.

3. Questionnaire Survey

3.1 Preliminary Survey

On December 8-9, 2017, we distributed 50 questionnaires titled "Survey on Benefits, Barriers, and Status of Scientific Data Open Sharing" to faculty and librarians with intermediate-level or master's degrees at a Hunan university during academic exchanges, and interviewed 6 experts. Since the preliminary survey covered three topics (benefits, barriers, status) with many items but incomplete questions for each theme, the data couldn't adequately support the research. Therefore, based on expert feedback, we revised the questionnaire in April 2018, creating separate questionnaires for barriers and benefits.

3.2 Questionnaire Revision and Main Survey

We redesigned the "Scientific Data Open Sharing Barriers Questionnaire" based on the 64 hypotheses (A1, A2, A3…F12), using a five-point Likert scale: "strongly agree," "agree," "uncertain," "disagree," and "strongly disagree" (scored 5, 4, 3, 2, 1 respectively). From May 23 to June 30, 2018, we distributed

200 paper questionnaires to faculty and students at 6 universities in Guangzhou Higher Education Mega Center, receiving 132 responses. We also invited faculty and graduate students from other regions (unrestricted by major) via Wenjuanxing, receiving 128 responses, for a total of 260 questionnaires. After removing invalid responses, we obtained 252 valid questionnaires. Invalid criteria were: (1) missing answers to one or more questions; (2) obvious careless or random responses.

The survey combined on-site paper distribution (where we first confirmed respondent identity before invitation) and online distribution (targeting familiar faculty, graduate students, and their networks). Since respondents regularly use scientific data for research, their responses are representative.

4. Empirical Analysis

Using SPSS 19.0, we conducted empirical analysis from validity, reliability, and factor analysis perspectives.

4.1 Questionnaire Validity and Reliability Analysis

We assessed validity through correlation analysis between each item (A1, A2, A3…F12) and total scores. SPSS 19.0 results showed correlation coefficients ranging from 0.176-0.787, all significant at p=0.000 (two-tailed), indicating high validity. Reliability was tested using Cronbach's alpha, yielding 0.981 for all 64 items—well above the 0.800 threshold—confirming satisfactory reliability.

4.2 First and Second Factor Analysis and Reliability

As shown in , KMO was 0.954 and Bartlett's test was significant (p=0.000), indicating high common factors among the 64 variables suitable for factor analysis. Using principal component analysis with eigenvalues >1, SPSS extracted 10 factors, but the items were disorganized with some factor loadings <0.400, making this approach unsuitable.

We then specified 6 factors. Principal component analysis showed these explained 62.518% of total variance (see), exceeding the 50% threshold. The rotated component matrix showed all loadings >0.400. Initial reliability analysis (see) showed Cronbach's alpha >0.797 for each factor. However, for Factor 2, item A2's "alpha if item deleted" was 0.948 > 0.947, indicating unreliability, so A2 was removed.

After deleting A2, we conducted a second factor analysis on the remaining 63 items. KMO was 0.954 (p=0.000). The 6 factors explained 62.825% of variance, but item A1 in Factor 2 showed "alpha if item deleted" as 0.949 > 0.948, so A1 was also removed.

4.3 Third Factor Analysis and Reliability

After removing A1, we analyzed the remaining 62 items. KMO was 0.955 (p=0.000). Using principal component analysis with Kaiser normalized orthogonal rotation and specifying 6 factors, SPSS converged after 10 iterations. The 6 factors explained 63.165% of variance. The rotated component matrix (see) showed all loadings >0.400, with each factor containing >3 items.

Factor 1 comprised D10, D11, E1-E14, and F1-F10, reflecting management and implementation issues, named "Management and Implementation Barriers." Factor 2 comprised B5-B8 and C1-C7, reflecting economic and legal issues, named "Legal and Financial Barriers." Factor 3 comprised D1-D9, reflecting technical issues, named "Technical Barriers." Factor 4 comprised A3-A10, reflecting cognitive deficiencies, named "Cognitive Barriers." Factor 5 comprised B1-B4, reflecting cost concerns, named "Cost Barriers." Factor 6 comprised A11, A12, F11, and F12, reflecting user issues, named "User Barriers."

Reliability analysis (see) showed Cronbach's alpha >0.797 for all factors. All "corrected item-total correlations" exceeded 0.400, and no "alpha if item deleted" exceeded the factor's alpha, confirming reliability. Thus, of the original 64 hypotheses, all were validated except A1 and A2.

5. Conclusions and Recommendations

The results validate six main barrier types: management and implementation (26 barriers), legal and financial (13), technical (9), cognitive (8), cost (4), and user barriers (4) (see).

These empirical findings differ somewhat from the initial theoretical analysis, primarily in that: (1) "Management Barriers" and "Implementation Barriers" merged into "Management and Implementation Barriers," likely because implementation requires management support. "Lack of data protection technology" and "Lack of open sharing standards" were added to this category, originally considered "Technical Barriers," possibly because some view these as management deficiencies. (2) "Legal Barriers" and some "Economic Barriers" combined into "Legal and Financial Barriers," as both are essential supports. (3) Four economic propositions became "Cost Barriers," showing concern about unnecessary costs. (4) User-related propositions from "Cognitive" and "Implementation" barriers formed "User Barriers," indicating successful sharing must address user problems.

We recommend: (1) Establishing a legal system for scientific data open sharing, strengthening data legislation, and developing detailed rules like the Chinese Academy of Sciences' "Scientific Data Management and Open Sharing Measures (Trial)" [27] to provide legal and policy support; (2) Creating scientific data open sharing funds to support major national projects and data maintenance; (3) Improving national platforms (e.g., National Basic Science Data Sharing Service Platform, National Population and Health Science Data Sharing Service Plat-

form) and building new platforms tied to national priorities; (4) Strengthening institutional platforms and training to address cognitive, cost, and technical issues, eliminating barriers to improve China's scientific data open sharing level.

References

[1] OECD. Making open science a reality [R/OL]. [2019-01-29]. http://wiki.lib.sun.ac.za/images/0/02/Open-science-oecd.pdf.
[2] Xiong Yi. Research on international open scientific data resources and utilization [J]. Library Theory and Practice, 2016(1): 12-14.
[3] Björk B. Open access to scientific publications: an analysis of the barriers to change? [EB/OL]. [2019-01-29]. http://www.informationr.net/ir/9-2/paper170.html.
[4] Hedrick TE. Justifications for and obstacles to data sharing [M]// Stephen EF, Margaret EM, Miron LS. Sharing research data. Washington, DC: National Academy Press, 1985: 123-147.
[5] Childs S, McLeod J, Lomas E, et al. Opening research data: issues and opportunities [J]. Records management journal, 2014, 24(2): 142-162.
[6] Martins S, Foulonneau M, Turkis, et al. Open data: barriers, risks, and opportunities [C]// Ferrarie, Castelnove W. Proceedings of the 13th European conference on e-government. Como: Academic Conferences and Publishing International Limited, 2012: 258-268.
[7] Janssen M, Charalabidis Y, Zuiderwijk A. Benefits, adoption barriers and myths of open data and open government [J]. Information systems management, 2012, 29(4): 258-268.
[8] Barry E, Bannister F. Barriers to open data release: a view from the top [J]. Information policy, 2014, 19(1): 129-152.
[9] Stagars M. Open data in Southeast Asia [M]. Singapore: Palgrave Macmillan, 2016: 17-20.
[10] Sui D. Opportunities and impediments for Open GIS [J]. Transactions in GIS, 2014, 18(1): 1-24.
[11] Panhuis WG van, Paul P, Emerson C, et al. A systematic review of barriers to data sharing in public health [EB/OL]. [2019-01-29]. https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-14-1144.
[12] Vanden Eynden V, Knight G, Vlad A, et al. Towards open research: practices, experiences, barriers and opportunities [R/OL]. [2019-01-29]. https://ndownloader.figshare.com/files/6739038.
[13] Zuiderwijk A, Janssen M, Choenni S, et al. Socio-technical impediments of open data [J]. Electronic journal of e-government, 2012, 10(2): 156-172.
[14] Fienberg SE, Martin ME, Straf ML. Sharing research data [J]. Medical care, 1986, 24(10): 879-880.
[15] Borgman CL, Qing Xiuling. Challenges of research data sharing [J]. New Technology of Library and Information Service, 2013(5): 1-20.
[16] Savage CJ, Vickers AJ. Empirical study of data sharing by authors publishing in PLOS journals [EB/OL]. [2019-01-29]. http://journals.plos.org/plosone/article?id=10.1371/journal.pone

[17] Borgman CL, Wallis JC, Enyedy N. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries [J]. International journal on digital libraries, 2007, 7(1/2): 17-30.
[18] Sayogo DS, Pardota. Exploring the determinants of scientific data sharing: understanding the motivation to publish research data [J]. Government information quarterly, 2013, 30(S1): 19-31.
[19] Tenopir C, Allard S, Douglass K, et al. Data sharing by scientists: practices and perceptions [EB/OL]. [2018-10-9]. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.00211
[20] Huang Ruhua, Qiu Chunyan. Review of foreign research on scientific data sharing [J]. Information and Documentation Services, 2013(4): 25-31.
[21] Peng C, Song X, Jiang H, et al. Towards a paradigm for open and free sharing of scientific data on global change science in China [EB/OL]. [2019-01-29]. https://esajournals.onlinelibrary.wiley.com/doi/epdf/10.1002/ehs2.1225.
[22] Sedransk N, Young LJ, Kelner KL, et al. Make research data public? Not always so simple: a dialogue for statisticians and science editors [J]. Statistical science, 2010, 25(1): 41-50.
[23] ?Tiemei, Gu Liping, Dong Jie, et al. Research on open access to foreign scientific data [M]. Beijing: China Financial & Economic Publishing House, 2017: 52.
[24] Wan Wanghui, Cui Chenzhou, Qiao Cuilan, et al. Analysis of astronomical observation data open sharing policies and strategies [J]. Astronomical Research & Technology, 2015, 12(3): 364-373.
[25] Chen Chuanfu. Mechanisms for public access to scientific data in China: characteristics, obstacles, and optimization suggestions [J]. China Soft Science, 2004(2): 8-13.
[26] Wessels B, Finn RL, Linde P, et al. Issues in the development of open access to research data [J]. Prometheus, 2014, 32(1): 49-66.
[27] Chinese Academy of Sciences. "Scientific Data Management and Open Sharing Measures (Trial)" [EB/OL]. [2019-02-21]. http://m.cas.cn/tzgg1/201902/P020190220358041915907.pdf.

Author Contributions

Sheng Xiaoping: Responsible for paper writing and revision.
Wu Hong: Participated in questionnaire design, survey, and data processing/analysis.
Hu Bingjie: Participated in questionnaire survey and data collection.

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