Postprint: A Review of Policy Research on Data Governance

Authors: Zhang Mengxia, Gu Liping, Gu Liping

Date: 2017-10-11T00:00:00+00:00

Abstract

【Purpose】 This study explores the implementation details of data governance policies to facilitate their establishment. 【Methods】 We review relevant theoretical achievements in data governance from both domestic and international sources, and systematically summarize three key dimensions: selection criteria for scientific data, storage specifications, and dissemination and communication mechanisms. 【Results】 The key policy elements encompass: data selection criteria (compliance with data submission workflows, priority selection principles, declarations of data authenticity and usability, and non-controversial data provenance); data storage specifications (adherence to relevant policies, assurance of data integrity, compliance with universal technical standards, and guarantee of long-term sustainability); and dissemination and communication mechanisms (compliance with laws, regulations, and industry guidelines, open access licensing agreements, disclaimers for dissemination activities, and documentation for data reuse). 【Limitations】 The policy framework requires further refinement in accordance with China' s specific context. 【Conclusion】 Research institutions, professional associations, funding agencies, and other stakeholders should actively promote and formulate data governance policies.

Full Text

Preamble

ChinaXiv Collaborative Journal, Issue 266, 2016, No. 1

A Review of Policy Research on Data Curation

Zhang Mengxia[1],[2], Ku Liping[1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract

[**Objective**] This study explores the implementation details of data curation policies to promote their establishment.
[**Methods**] We systematically review theoretical achievements in data curation both domestically and internationally, summarizing findings across three dimensions: scientific data selection criteria, storage standards, and communication mechanisms.
[**Results**] The key policy elements of data curation include: (1) Data selection criteria (compliance with submission workflows, priority selection principles, statements of data authenticity and usability, and non-controversial data sources); (2) Data storage standards (adherence to relevant policies, ensuring data integrity, meeting general technical standards, and guaranteeing long-term sustainability); and (3) Communication mechanisms (compliance with laws and industry guidelines, open access licensing agreements, disclaimers for dissemination activities, and documentation for data reuse).
[**Limitations**] The policy framework requires further refinement in detail to align with China' s specific context.
[**Conclusions**] Research organizations, associations, and funding agencies should actively promote and formulate data curation policies.

Keywords: Data curation; Data management service; Digital archives management; Data rights management; Data selection; Long-term preservation
Classification Number: G302

1. High-Quality Reuse of Scientific Data Under Digital Curation

Sharing scientific data through data papers lays the foundation for data discovery, reuse, and recognition of researchers' contributions within their fields. In the e-Science environment, there is growing demand for high-quality, assured datasets to support data-driven research. Consequently, the scientific community widely recognizes the need for proactive curation throughout the data lifecycle to better align data with utilization and reuse requirements while promoting sharing and exchange within academic communities. Digital curation refers to the active management of scientific data throughout its lifecycle through processes such as data annotation, evaluation, selection, and transformation to add value and enable broader data sharing.

In 2012, the Association of College and Research Libraries (ACRL) identified data curation as a mainstream development trend for libraries. The library and information science community has subsequently implemented various data curation initiatives, including: (1) constructing data repositories such as Edinburgh DataShare, Dryad, and figshare; (2) developing data curation education programs, with the University of Illinois at Urbana-Champaign launching a data curation course in 2006, followed by North Carolina State University, the University of Michigan, and others; (3) creating lifecycle-based data management

planning tools like the California Digital Library' s DMP Online; and (4) establishing data management infrastructure and practices such as Johns Hopkins University' s Data Conservancy, Purdue University Research Repository, Rutgers University' s RUresearch Data Portal, and Cornell University Library' s DataStar project.

While these efforts have achieved notable progress, new challenges have emerged, including quality control issues for curated data, formulation of institutional data policies, and development of reference guides to support data curators. These challenges will intensify as data curation work continues to expand.

Data curation has become a crucial development strategy for libraries and archives, involving a series of policy questions in its implementation: What standards must data meet for curation? How should data be stored? What mechanisms govern data communication and exchange? This study investigates these questions through policy research to explore the policy elements of data curation and provide references for policy formulation and decision-making services in practical work.

2. Research on Scientific Data Management Standards Under Digital Curation

The processes of selection, storage, evaluation, analysis, reuse, and sharing involve not only engineering issues but also policy considerations. Research inquiries into the rights and interests associated with these services should address three fundamental questions: (1) What management procedures are required for scientific data generated during research activities? (2) What management approaches should apply to selected scientific data stored on data infrastructure? (3) What sharing mechanisms should academic communities follow when disseminating data submitted to and stored in data repositories?

Accordingly, this paper's research framework addresses both "research questions" and "observation questions," as shown in .

3. Scientific Data Selection Criteria

The International Council for Scientific and Technical Information (ICSTI) information lifecycle framework identifies selection as an indispensable component of the workflow. Data curation similarly depends on data selection for three primary reasons: (1) data backup and mirror site maintenance incur costs, making selection necessary to determine whether data should continue to be stored after the preservation period expires; (2) without selection, stored data may grow uncontrollably, become redundant, and make discovery, mining, and utilization difficult; and (3) valuable data may be lost when research project lifecycles end, but sound selection practices can preserve such data in a timely manner.

3.1 Compliance with Data Submission Workflow Requirements

Data curation is implemented in three primary scenarios: (1) data contributors generate, describe, and submit data themselves; (2) data curators collect, evaluate, select, store, and preserve data; and (3) data contributors submit data that curators then screen, review, manage, and provide access to. The third approach is most common, with both parties following an established workflow. Data contributors and curators must jointly adhere to two key principles: first, datasets should encompass all content requiring preservation, as the richness and completeness of preserved content significantly impacts data understanding, utilization, and curation; second, metadata should be as detailed and complete as possible to enable efficient use by others.

3.2 Priority Selection Principles

Under equivalent conditions, data curation aims not only to ensure long-term preservation but also to facilitate knowledge exchange. Prioritizing trustworthy, usable, and valuable data better serves this objective. Priority selection principles include: (1) one-time original data from transient or unique events such as weather observations, volcanic eruptions, or rainfall records; (2) non-reproducible data where the observed object may still exist, but the measured variables change over time, making the original experimental data irreproducible; (3) non-redundant data that eliminates duplicates or useless information in computer systems; and (4) data with scientific research, historical documentation, and socioeconomic value.

When implementing the fourth principle, data curators require operational definitions. Scholars reviewing reports from the National Research Council (NRC), DDC, and ANDS have identified three value dimensions: (1) scientific value, meaning data support scientific activities and enable verification of scientific results; (2) historical value, extending data use beyond researchers to social groups and individuals; and (3) social value, reflecting contemporary societal interests and contributing to future socioeconomic development through reuse.

3.3 Statement of Data Authenticity and Usability

Data contributors should submit a statement attesting to their data' s authenticity and usability, which encompasses four aspects: (1) data interpretability; (2) data verifiability and reusability, enabling effective validation of research results and confirmation of conclusions through data provenance; (3) absence of fabricated information, allowing the dataset to serve as scientific evidence supporting relevant conclusions; and (4) no deliberate screening or suppression of information. Such statements demonstrate contributors' accountability while helping curators quickly understand the data.

3.4 Non-Controversial Data Sources

Legality is the foremost principle of data curation, encompassing the basis, process, and content of data generation. Therefore, curated data must not be produced through illegal, unreasonable, or unethical means. Data curation should avoid legal, moral, and ethical controversies, including issues of data ownership, human subjects research, personal information, national security, confidentiality, and sharing prerequisites required by data providers. These principles require clarification of three points: (1) data generation must not violate academic ethics, scientific integrity, or existing laws and regulations; (2) scientific data from research activities should comply with legal, ethical, and social norms, with discipline-specific information management guidelines prioritizing such principles; and (3) data dissemination must consider the legitimate rights and interests of all stakeholders.

Industry standards provide important references for implementing these principles. In crystallography, for example, data users developing new products, conducting research, or applying for projects must communicate with data contributors to obtain formal or informal consent. In geobiology, scientific data from nationally funded instruments and projects generally require open sharing, while large institutes and commissioned companies often retain rights and propose embargo periods for open access.

4. Data Storage Standards

Data infrastructure encompasses: (1) large-scale instruments and their information platforms; (2) domain-specific data exchange networks; (3) data centers serving as big data resource bases or project repositories; (4) data banks providing storage repositories for contributors based on agreements; (5) data archives offering specific resources to interested end-user groups; and (6) libraries providing electronic document platforms, data repositories, and institutional repositories. Despite varying purposes, stakeholders (funding agencies, research managers, project leaders, data contributors, curators, and users) must reach consensus or adhere to common conventions when facing data curation challenges.

4.1 Prioritizing Compliance with Funding and Institutional Data Policies

Funding agencies' policies, terms, and management regulations stipulate how project leaders should handle data. As infrastructure supporting data curation, repositories must comply with these policies to request, invite, and accept data that meets selection standards for curator review and user access. Based on preliminary research, funding and research institutions require data from funded projects to be preserved and openly shared in compliance with institutional regulations. These policies typically specify: (1) disciplinary scope; (2) submission timeframe, generally within 6-12 months after project completion; (3)

minimum preservation period, with research data preserved for at least 3 years for verification and 10 years for utilization; (4) open access date, implemented within 12 months of formal research publication; and (5) storage location, in institutional repositories or third-party data centers with verified preservation and dissemination capabilities that protect stakeholder rights.

4.2 Ensuring Data Integrity

Data integrity is a crucial quality metric. Unlike data contributors' statements of authenticity and completeness (see Section 3.3), storage standards for data infrastructure emphasize specific measures to strictly safeguard integrity. For example, the Inter-university Consortium for Political and Social Research (ICPSR) developed a social science data repository based on the OAIS model that emphasizes contextual information, preservation description information, and user access permissions to ensure integrity.

Basic integrity requirements include: (1) preventing tampering—if data require modification, three options exist: contributors should document changes when updating data for version clarity; curators making minor corrections for format standardization or preservation must inform contributors or follow established protocols; and third-party objections regarding storage, dissemination, or content should be addressed by contributors, curators, and research managers; (2) preventing unintentional modifications—curators should only revise data to correct errors from initial submission flaws, format migration losses, or other damage; and (3) secure backup—data should be replicated or migrated according to fair use principles to prevent loss from natural disasters, accidents, or organizational changes.

4.3 Meeting General Technical Standards

Data curation relies on appropriate technical standards for three reasons: (1) unified standards improve efficiency and reduce costs during data migration; (2) similar tools, methods, and professional skills during backup reduce data noise; and (3) standards prevent dependence on few vendors or skill groups during conversion, increasing implementation options.

General technical standards typically cover file formats, reference standards like OAIS, persistent unique identifiers, and standards supporting remote access, storage, and verification. Characteristics include machine readability, human recognition, easy access, format convertibility, and openness compatible with these conditions—essential for periodic migration and adaptation to multiple technical strategies cost-effectively.

4.4 Ensuring Long-Term Sustainable Development

A core objective of data curation is maintaining high-quality open sharing and preventing important data loss. Open access enables full utilization, provenance

tracking, and recognition of research contributions, requiring long-term sustainable mechanisms to ensure these records persist.

Long-term sustainability addresses three concerns: (1) avoiding one-time projects—valuable scientific data often disappears when projects end, making sustainable curation mechanisms essential; (2) adopting non-commercial business models—curation has cost-benefit considerations, and high standards requiring objective, complete, and consistent data 不受利益因素干扰 (free from interest-based interference) prioritize non-commercial approaches; and (3) preparing contingency measures for force majeure incidents—standard operating procedures are needed to address losses from natural disasters or human-caused accidents.

5. Communication and Dissemination Mechanisms

A key goal of data curation is providing open access to scientific data. While data infrastructure establishes internal management mechanisms for submission, organization, processing, storage, and sharing, actual data sharing faces obstacles such as inadequate notification to contributors about data usage and insufficient protection of contributor rights. From a policy planning perspective, complete data curation requires communication mechanisms alongside selection standards and storage specifications.

5.1 Compliance with National Laws and Industry Best Practices

When considering dissemination mechanisms, national laws take precedence, followed by industry best practices. Specific considerations include: (1) compliance with copyright and related rights (moral rights deserved by contributors and allocation of property rights), intellectual property conversions (IPCs) (requirements by funders and managers for free access to commercially valuable data), institutional and/or repository policies; and (2) scientific record management standards such as Australia' s *Code for the Responsible Conduct of Research* and the UK Research Councils' policies on good research conduct, which specify data storage, duration, methods, and sharing.

5.3 Disclaimer for Dissemination Activities

The primary ethical issue in data sharing involves sensitive personal or organizational data and security controls. Disclaimers inform users that data selected, stored, and openly accessed by curators meets certain quality standards. While contributors declare non-controversial data sources, curators have an obligation to inform users about legitimate sources and proper usage. For example, human subjects data require informed consent documentation; sensitive or political data need ethics committee approval; and third-party contract data requires authorization letters. Disclaimers not only fulfill notification obligations but also foster good academic exchange, promote self-regulatory ethical practices, and enhance the open access cycle.

5.4 Documentation for Data Reuse

Ultimately, all eleven policy elements across selection, storage, and communication aim to ensure good data management and dissemination. The final checkpoint is documentation explaining how to understand data structure and field meanings for reproducing results, replicating products, or recreating research. Such documentation includes data dictionaries, contextual information about data creation environments (project nature, collection and processing methods), and recommended data access and citation practices.

Licenses are specifications by copyright holders on content usage, with open access licensing agreements being crucial for protecting contributor rights and granting user rights in digital, networked environments. The widely used CC-BY license requires users to attribute contributors, while CC0 licenses apply to objective facts like government statistics where attribution is unnecessary. Since data collection, processing, selection, and submission involve intellectual labor, datasets typically use CC-BY or stricter licenses, while metadata describing these datasets should use CC0.

6. Conclusion

6.1 Summary of Policy Elements for Data Management Standards in Curation

This paper operationalizes the three research questions into observable issues and policy elements, identifying three management priorities: data selection criteria, data storage standards, and communication mechanisms, as summarized in .

6.2 Practical Significance

In the e-Science environment, discussions of research and publication workflows are extensive, and while data curation has been addressed in important studies, most focus on technology, systems, and education rather than policy—particularly data-specific policy, which remains in early exploration domestically. Data curation is a systematic project involving data objects and their integrity, technical measures, legal and organizational factors, and other elements like policy standards, open specifications, and metadata. This paper supplements previous research, and its research framework (Table 1) and policy framework (Table 2) encompass the basic rights of all stakeholders, providing references for policy formulation.

6.3 Research Limitations

The questions listed in Table 1 require further contextualization based on actual conditions, focusing on analyzing current practical problems in data curation. The operational data curation plan outlined in Table 2 must be implemented in coordination with data management policies of funding agencies, research

institutions, and information service organizations, and combined with Data Management Plans (DMPs) for optimal effectiveness. When generalizing these findings, case studies should be incorporated, and actual needs of funding agencies, research managers, project leaders, and researchers must be considered when formulating management standards.

6.4 Future Research

Future work should conduct field research on research teams and analyze policy elements of data management plans and rights management for data and derived data, in addition to data selection. Open access licensing agreements should be coordinated with research management departments to develop complete workflow solutions based on this study.

Acknowledgments

We thank Ms. Wang Hui and Ms. Ouyang Zhengzheng from the National Science Library, Chinese Academy of Sciences, for their guidance from the subject librarian perspective and for answering questions and providing suggestions. We also thank the National Science Library for organizing the compilation of *How to Appraise & Select Research Data for Curation*, which informed this study. Compilation team: Mu Huige, Wang Lu. We appreciate Wu Rong' s work on standardizing the collected references.

References

[1] Liu Jingjing, Ku Liping. The Policy Research and Analysis of Data Journals: Taking Scientific Data as an Example[J]. Chinese Journal of Scientific and Technical Periodicals, 2015, 26(4): 331-339.

[2] Rombouts J, Princic A. Building a 'Data Repository' for Heterogenous Technical Research Communities Through Collaborations[EB/OL]. [2015-06-14]. http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1014&context=iatul2010.

[3] Shreeves S L, Cragin M H. Introduction: Institutional Repositories: Current State and Future[J]. Library Trends, 2008, 57(2): 89-97.

[4] Zhang Zhixiong, Wu Zhenxin, Liu Jianhua, et al. Analysis of the Difference Between Digital Curation and Digital Preservation[J]. New Technology of Library and Information Service, 2014(1): 4-13.

[5] McLure M, Level A V, Cranston C L, et al. Data Curation: A Study of Researcher Practices and Needs[J]. Portal: Libraries and the Academy, 2014, 14(2): 139-164.

[6] Wright S J, Kozlowski W A, Dietrich D, et al. Using Data Curation Profiles to Design the Datastar Dataset Registry[J]. D-Lib Magazine, 2013, 19(7-8). http://www.dlib.org/dlib/july13/wright/07wright.html.

[7] Song Xiufen, Deng Zhonghua. The Knowledge and Skills Needed for Data Curation and Education Research[J]. Research on Library Science, 2014(21): 5-11.

[8] Cao Xia. Research on Development of Studying on Data Curation in China[J]. Library and Information Service, 2014, 58(18): 144-148.

[9] Gao Hongwen, Chen Qingwen. Foreign Data Curation Research Review[J]. Research on Library Science, 2013(10): 2-4, 27.

[10] Wang Fang, Shen Jinhua. Advances in Data Curation Abroad: Research and Practice[J]. Journal of Library Science in China, 2014, 40(4): 116-128.

[11] Hodge G M. Best Practices for Digital Archiving: An Information Life Cycle Approach[J]. D-Lib Magazine, 2000, 6(1). http://www.dlib.org/dlib/january00/01hodge.html.

[12] Whyte A, Wilson A. How to Appraise & Select Research Data for Curation[EB/OL]. [2015-06-14]. http://www.dcc.ac.uk/sites/default/files/documents/How%20to%20Appraise%20a

[13] Higgins S. Draft DCC Curation Lifecycle Model[J]. International Journal of Digital Curation, 2008, 2(2): 82-87.

[14] Constantopoulos P, Dallas C, Androutsopoulos I, et al. DCC&U: An Extended Digital Curation Lifecycle Model[J]. International Journal of Digital Curation, 2009, 4(1): 34-45.

[15] Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation' s Scientific Information Resources[M]. Washington: National Academies Press, 1995.

[16] Carlson J, Fosmire M, Miller C C, et al. Determining Data Information Literacy Needs: A Study of Students and Research Faculty[J]. Portal: Libraries and the Academy, 2011, 11(2): 629-657.

[17] Cragin M H, Palmer C L, Carlson J R, et al. Data Sharing, Small Science and Institutional Repositories[J]. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 2010, 368(1926): 4023-4038.

[18] The Latest "Earth Biology Frontier" Project[EB/OL]. [2015-05-12]. http://blog.sciencenet.cn/blog-528739-807103.html.

[19] Beagrie N, Greenstein D. A Strategic Policy Framework for Creating and Preserving Digital Collections: A Report to the Digital Archiving Working Group[R]. London: British Library Research and Innovation Centre, 1998:17-18.

[20] Zhang Yao, Ku Liping, Yang Yunxiu, et al. Research Data Policies of Research Funding Agencies: Case Study of British and American Research Councils[J]. Library and Information Service, 2015,59(6): 53-59.

[21] Yang Yunxiu, Ku Liping, Zhang Yao, et al. Research on Data Policies of Research and Education Institutions: Case Study of British Universities[J]. Library and Information Service, 2015, 59(5): 53-59.

[22] Hedstrom M. Research Issues in Migration and Long-term Preservation[J]. Archives and Museum Informatics, 1997, 11(3): 287-292.

[23] Mellor P, Wheatley P, Sergeant D. Migration on Request, a Practical Technique for Preservation[A]. // Research and Advanced Technology for Digital Libraries[M]. UK: Springer Berlin Heidelberg, 2002: 516-526.

[24] Barateiro J, Antunes G, Freitas F, et al. Designing Digital Preservation Solutions: A Risk Management-Based Approach[J]. International Journal of Digital Curation, 2010, 5(1): 179-194.

[25] Hedstrom M, Lampe C. Emulation vs. Migration: Do Users Care?[J]. RLG DigiNews, 2001, 5(6): 5-11.

[26] Knight G, Hedges M. Modelling OAIS Compliance for Disaggregated Preservation Services[J]. International Journal of Digital Curation, 2008, 2(1): 62-72.

[27] McMeekin S M. With a Little Help from OAIS: Starting down the Digital Curation Path[J]. Journal of the Society of Archivists, 2011, 32(2): 241-253.

[28] Besek J M. Copyright Issues Relevant to the Creation of a Digital Archive[J]. Microform & Imaging Review, 2003, 32(3): 86-97.

[29] Bearman D. Intellectual Property Conservancies[J]. D-Lib Magazine, 2000, 6(12). http://www.dlib.org/dlib/december00/bearman/12bearman.html.

[30] Australian Code for the Responsible Conduct of Research[EB/OL]. [2015-06-08]. http://www.nhmrc.gov.au/publications/synopses/r39syn.htm.

[31] RCUK Policy and Code of Conduct on the Governance of Good Research Conduct[EB/OL]. [2015-06-12]. http://www.rcuk.ac.uk/review/grc/default.htm.

[32] Kirschenbaum M G, Ovenden R, Redwine G, et al. Digital Forensics and Born-Digital Content in Cultural Heritage Collections[M]. Washington: Council on Library and Information Resources, 2010:49-50.

[33] Xie Jing, Chen Ya. Research on the Curation of Scientific Dada Home and Abroad[J]. Journal of Academic Library and Information Science, 2014, 32(4): 114-119.

[34] Shen Tingting, Lu Zhiguo. Science Data Curation Methods at Different Stages of Scientific Research Projects[J]. Library Development, 2013(3): 49-51.

[35] Yu Haiyan, Wei Junchao. Investigation and Analysis on University Data Curation Projects Abroad[J]. Library and Information Service, 2014, 58(22): 38-47.

[36] Granger S. Emulation as a Digital Preservation Strategy[J]. D-Lib Magazine, 2000, 6(10). http://www.dlib.org/dlib/october00/granger/10granger.html.

Author Contributions:

Zhang Mengxia: Policy detail analysis, expert interviews, manuscript drafting
Ku Liping: Research design, information provision, policy element analysis, final manuscript revision

Received: June 26, 2015
Revised: November 24, 2015

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