Postprint of Research Data Lifecycle Study and Data Repository Theoretical Framework

Authors: Yang Le, Yan Shilei, Li Hongbo

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

Abstract

[Purpose/Significance] To advance the development of domestic scientific research data management systems and clarify three key research areas along with foundational theoretical research on data repositories.

[Method/Process] This study employs empirical research methods to analyze and compare the current state of scientific research data management research both domestically and internationally, and conducts a detailed exploration of foundational research and technical demonstrations in this domain.

[Results/Conclusion] The study explicitly identifies the three foundational research components for constructing a scientific research data management system: the research activity cycle and data lifecycle, theoretical framework and process research for data repositories, and technical demonstrations for data repositories. Simultaneously, it proposes system construction plans and mechanism construction plans for the management system.

Full Text

Preamble

ChinaXiv Partner Journal, Vol. 63, No. 1, January 2019

Research on Research Data Lifecycle and Theoretical Framework of Data Repositories

Yang Le[1], Yan Shilei[1], Li Hongbo[2]

[1] Wenzhou-Kean University, Wenzhou 325060
[2] Wenzhou Medical University, Wenzhou 325035

Abstract

[**Purpose/Significance**] This paper aims to promote the systematic development of research data management systems in China and clarify three key research areas and foundational theoretical studies for data repositories. [**Method/Process**] Using empirical research methods, this study analyzes and compares the current state of research data management domestically and internationally, with detailed discussion of fundamental research and technical demonstration in the field. [**Result/Conclusion**] The paper explicitly identifies three core components of foundational research for constructing a research data management system: research activity cycles and data lifecycles, theoretical frameworks and workflow studies for data repositories, and technical demonstration of data repositories. It also proposes systematic and policy development strategies for management system construction.

Keywords: research data management; data repository; research data lifecycle
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2019.01.013

Introduction

Research data management has developed rapidly in foreign academic circles and university libraries in recent years. The 2015 Horizon Report, jointly released by the New Media Consortium and EDUCAUSE Learning Initiative, identified research data management as a key future trend for university libraries [1-2]. In 2016, the Association of College & Research Libraries (ACRL) released its top ten research trends, ranking "research data services" first [2-3]. Against the backdrop of collaborative research, interdisciplinary studies, and rapid information technology development, issues such as how to preserve, discover, cite, reuse, and maintain large volumes of research data long-term have attracted significant attention from the academic community [4].

Against this international backdrop and with China's increasingly active academic research leading global developments in many areas of basic and applied research, the General Office of the State Council issued the "Measures for the Management of Scientific Data" on March 17, 2018, formally requiring that research activities funded by government budgets must properly collect, produce, process, organize, share, and manage scientific data [5]. However, domestic university research data management services remain in the discussion and demonstration stage. No effective localized model for the data lifecycle has been established, and foundational research and development of research data repositories have not yet begun. How to construct a service system for research data management from the foundational demonstration stage, establish a localized data lifecycle model, research workflows for various research activities, build a complete framework for research data management, and explore the development of a comprehensive data repository system have become issues that universities and research institutions must address. This concerns not only the

strategic planning of universities and research institutions themselves but also the orderly development of domestic scientific research.

2 Research Status

2.1 Foreign Research Status

Research data management services in U.S. universities and research institutions maintain high standards in funding, professional qualifications, position structures, and technical requirements. University library data management teams provide embedded consulting services starting from the initial project application phase [6]. Using research data management plan platforms (such as the DMPTool developed by the California Digital Library) and leveraging expertise from information professionals in areas like data formats, archiving, curation, and metadata configuration, they guide applicants and researchers in writing data management plans that meet funding agency requirements. Complementing these consulting services are training programs where research data librarians conduct workshops, lectures, and online learning sessions on data literacy and research data lifecycles for researchers, faculty, and graduate students [7-8].

K. Akers et al. from the University of Michigan [9] systematically documented the timelines for research data services, management, evaluation, traditional institutional repository construction, and data repository development at eight leading U.S. university libraries, as shown in Figure 1 [Figure 1: see original paper]. The figure reveals that Cornell University Library began research data management services as early as 1984, while Emory University started in 1996. The National Institutes of Health (NIH) began requiring data management plans for projects exceeding $500,000 on February 26, 2003. Following this, research data management and data repositories developed rapidly. On January 18, 2011, the National Science Foundation (NSF) mandated that all grant proposals must include data management plans and implementation strategies, prompting research-intensive university libraries to popularize data management practices.

On February 22, 2013, the Obama administration signed a science and technology policy memorandum requiring all federally funded research to publicly release raw research data unconditionally within specified timeframes through open access institutional or data repositories. Following this policy implementation, all federal funding agencies successively issued specific guidelines for research data management, and U.S. university and research institution libraries comprehensively launched related initiatives. Research-capable academic libraries took the lead in providing research data management services and consultation, initially storing and publishing research data on existing institutional repository platforms while developing research data management systems through interlibrary cooperation.

Regarding data repository systems, three main approaches exist: First, distinctive open-source systems like Harvard University's Dataverse and the University

of North Carolina's iRODS; second, regionally unique and personalized data repositories such as Purdue University's PURR, the University of California system's DataONE (current version Dash), and the University of Michigan's specialized ICPSR for social sciences and humanities [10]; third, many universities use traditional institutional repository systems like DSpace and Fedora for research data storage and publication [11]. Additionally, the next-generation institutional repository system Samvera has emerged, integrating traditional and research data functions with Fedora as the backend and specialized frontend applications like Hydra, Avalon, Hyku, and Hyrax to support various data formats—though this development has not yet been covered in existing journal literature.

2.2 Domestic Research Status

On March 17, 2018, the General Office of the State Council issued the "Measures for the Management of Scientific Data," formally requiring proper collection, production, processing, organization, sharing, and management of scientific data for research activities funded by government budgets [5]. Currently, domestic funding agencies at various levels have not yet established requirements for data management plans or implementation. Most domestic university libraries and research institutions have limited involvement in research data management or consulting services.

Existing domestic journal literature discusses foreign university data management concepts, models, and systems [12], focusing on implications for domestic university development [13-15], but with limited progress in practice and implementation. University management and academic teams provide weak support for research data management, with low requirements for funding, professional qualifications, position structures, and technical specifications [2, 12].

Research on research data lifecycles represents a key focus in research data management. Only by establishing data lifecycle models localized for China's research environment—or models tailored to individual universities' disciplinary strengths—can research data management services and system construction be effectively targeted [3]. Data repository structure, workflow establishment, and function settings must align with the research data lifecycle. Domestic research on research data lifecycles is scarce, with existing literature merely comparing foreign results and discussing implications for domestic library and information science construction [16-17], without presenting a mature, domestically applicable data lifecycle and architecture model.

In terms of data repository systems, compared to the comprehensive integration of traditional institutional repositories and data repositories in the U.S. [18-20], domestic universities show unbalanced development [21]. Many universities have not yet established traditional institutional repositories or data repositories, with only a handful having preliminary data repositories. For instance, Peking University's Open Research Data Platform, Fudan University's Social

Science Data Platform, and Hong Kong University of Science and Technology's DataSpace all use the U.S. Dataverse open-source software. Technologically, aside from existing traditional institutional repositories like CSpace from the Lanzhou Branch of the National Science Library, Chinese Academy of Sciences, and CIRP from CNKI, China lacks localized data repository systems.

3 Problem Statement

U.S. repository systems each have functional and technical strengths but also limitations. Dataverse is a complete Java application with frontend and backend, but its frontend is not truly decoupled from backend logic, its REST API design is suboptimal, customization is limited, and it provides only basic data visualization. Its advantages lie in highly personalized metadata configuration and comprehensive version archiving for research data at different stages, making it particularly suitable for regional data and capable of integrating with existing traditional repository software. iRODS, using C/C++, targets large-scale distributed data with fast big data processing, commonly used in meteorological modeling and biological nucleic acid sequence analysis, but suffers from an unfriendly user interface since its frontend applications are provided by certain Microservice plugins, making customization and deep configuration difficult.

Among traditional institutional repository software, Fedora is more suitable for data repositories. While its distributed big data processing capabilities approach those of iRODS, its most prominent advantages include advanced REST API web services, high-precision version control, high-speed caching, support for multiple data storage technologies, pluggable user authentication systems, and highly extensible architecture. Its user interface and technical support for research data integration far exceed those of the aforementioned systems. DSpace, another widely used system, has an outdated frontend framework (Cocoon), though the developing DSpace 7 will adopt the new Angular 2/4 to create single-page applications with a redesigned REST API. DSpace's limitations include lack of support for some data format versions and composite data types, with limited visualization methods, making it more suitable for published literature.

In summary, although foreign data repositories have relatively complete systems, they have not effectively solved functional aggregation, only partially integrating certain system functions and exhibiting characteristics of multiple coexisting but incompatible functional systems [22-23]. According to a 2016 survey by U.S. scholar R. Uzwyshyn, while 74% of U.S. research universities provided research data system platforms, only 13% used specialized data repository platforms, with others using traditional institutional repositories or websites instead [24]. Chinese universities remain in the foundational demonstration stage for research data management and data repositories, with no established basic data lifecycle model and only a few data repositories using existing U.S. technology that has technical shortcomings and insufficient language modules, particularly in metadata support for discovery functions in foreign academic

retrieval systems. This phenomenon is not limited to China; in a survey of data repositories in China, Japan, and Korea, Korean scholars S. Kim and W. Lee found that while data repositories from these three countries account for 42.2% of Asian academia, the vast majority are traditional database frameworks and institutional repository frameworks (like DSpace) [25].

Domestic library and information science research must conduct deeper studies on existing international data repository architectures and models, combine these with localized needs of Chinese university researchers for research data management, establish localized data lifecycle models, and build a more comprehensive data management model and data repository system that is both localized (e.g., Chinese metadata sets) and internationalized (e.g., configuring Schema.org or OAI-PMH metadata protocols for search engine retrieval).

4 Research Content

To achieve breakthroughs in this field, domestic library and information science researchers must address three main aspects (three stages): refined research activity cycles and research data lifecycle models; theoretical frameworks and workflows for data repositories; and technical demonstration for data repositories.

4.1 Research Activity Cycle and Data Lifecycle Model

Research on research data lifecycles has a long history in U.S. academia. Most Association of Research Libraries (ARL) member libraries have established generic lifecycle models based on their institutions' disciplinary profiles to guide campus research data management services and data repository development. As shown in Figure 2 [Figure 2: see original paper], a generic data lifecycle model includes four modules: long-term data curation ensures permanent preservation and format updates; data access permissions allow researchers to set appropriate sharing levels based on research activity cycles and workflows, facilitating data sharing among team members and cross-team collaborators; data security requires data librarians to anonymize sensitive information about human subjects according to funding agency and institutional requirements; and data publication involves not only meeting funding agencies' open access requirements but also ensuring integrity-preservation curation metadata strategies and discovery metadata strategies that support external sharing.

Research on discipline-specific research activity workflows and localization of research data lifecycles is essential for constructing applicable domestic models. These studies require surveys and interviews with researchers to obtain firsthand feedback and information about experimental data formats across disciplines.

4.2 Data Repository Theoretical Framework and Process

The second research component involves comprehensively studying how existing international data repositories and related systems process research data, deconstructing the overall framework, and constructing a structural framework and model for data repository systems based on localized research cycles and research data lifecycle models.

According to existing international literature and empirical cases, major research data management systems have three functional modules and workflows: data management planning, research data storage and publication, and long-term curation and integration (see Figure 3 [Figure 3: see original paper]). This system covers the entire research data lifecycle, yet no single data repository system currently effectively aggregates all three functional modules; most operate as independent systems. For example, the California Digital Library's DMPTool handles data management plans, Dash handles data storage and publication/citation, and Chronopolis handles long-term preservation; some repositories partially integrate storage, citation, publication, and long-term curation, such as Purdue University's PURR system, which lacks only the data management plan component.

The data management plan platform allows data management staff to embed into early-stage research cycle planning. After registration, researchers can retrieve appropriate templates based on disciplinary needs and, with assistance from data management staff, draft and submit data management plans for review, ensuring preliminary data management concepts and layouts at the research activity's initial stage. These discipline-specific templates must be built upon localized research data lifecycle models.

For data repository and preservation/integration platforms, data repository functions must satisfy corresponding data lifecycle requirements, allowing researchers and data management staff to collaboratively complete data submission, metadata configuration, system optimization, permission setting, cross-disciplinary integration, and final data publication and citation. All these functional developments must be based on prior research data lifecycle model studies. Without precise lifecycle models as theoretical support, data repository workflow and function construction would lack direction.

4.3 Data Repository Technical Demonstration

Technically implementing a comprehensive research data management process requires adherence to three principles: fully recognize and utilize existing domestic and international resources to select the most suitable software platform; ensure functional decoupling in software design so modules cooperate actively without mutual interference, where customization or updates to any module minimally impact others; and employ contemporary software technologies while striving for innovation in application.

Based on existing platforms, researchers must thoroughly study current data repository technologies, such as Fedora and iRODS for distributed big data processing, Dataverse with its consistent frontend-backend logic, or DSpace 7 using Angular 2/4 for single-page applications with redesigned REST APIs. Technically explore the most reasonable data storage architecture for the three functional modules and optimize technical implementation of research data management business logic. Finally, construct a complete research data management system model based on sufficient empirical data, theory, and technology.

For the critical repository backend platform, design must integrate advanced concepts such as linked data and semantic web-oriented architectures. The backend must support highly customized metadata formats and handle any file format, enabling development of discipline-specific research data management systems and providing flexibility for localization. The customized backend platform will integrate mainstream user authentication systems (e.g., LDAP), provide multiple storage options, establish metadata-based fast full-text search services in both Chinese and English, and connect to major world research databases and dataset discovery systems (e.g., Google Dataset Search) to facilitate discovery, publication, and utilization of research data.

The final technical solution involves rationalizing user platforms and optimizing frontend-backend metadata strategies. User platforms should serve three groups: external research data users, research data management staff, and researchers. The platform will provide fast, convenient single-page applications and mobile apps for researchers, information users, and general users. This frontend interface will communicate only through well-designed REST API web services, achieving complete frontend-backend separation.

5 Construction Plan

Through initial-stage surveys and interviews with researchers across disciplines on research data types, formats, application scopes, usage methods, data lifecycles, and research methodologies, library and information science researchers can begin detailed planning for data repository construction. To effectively promote system development, the construction plan can be divided into two components: system construction and mechanism construction.

5.1 System Construction Plan

As illustrated in Figure 3, the first workflow component is data management planning, which should enable researchers from different disciplines to retrieve appropriate plan templates for developing data management plans. Data management staff must review and provide feedback to help researchers improve their plans for more reasonable research data management. In the second workflow component, library data management staff must develop a standardized metadata scheme covering all disciplines at the database level while maintaining extensibility for specific disciplinary requirements. Researchers submit research

data and perform initial metadata configuration, after which data management staff conduct standardization to ensure effective search engine retrieval and facilitate data reuse. In the third workflow component, data management staff perform long-term curation to ensure research data integrity while enabling cross-disciplinary integration and interoperability, providing precise personalized services to interdisciplinary researchers and helping them retrieve, use, and cite existing research data across disciplines.

5.2 Mechanism Construction Plan

A comprehensive research data management paradigm (including data repositories) requires complete supporting mechanisms to better serve academic research within and across institutions. Specific mechanisms must follow several principles: all research data submitted to repositories must have approved data management plans, with collaborative implementation by data management staff and researchers; researchers can change sharing permissions during and after projects to ensure orderly progress, and may request data deletion anytime without violating funding agency or institutional regulations; data management staff standardize and modify metadata during and after projects to optimize retrieval and discovery mechanisms; and researchers should include data management costs in funding applications based on institutional circumstances, ensuring sustained financial support, maintenance, and upgrades for research data management systems.

This approach fundamentally solves the aggregation problems of foreign research data management systems, creating a compatible management system covering data management planning, service provision, storage, publication, and long-term curation. It better responds to the State Council's call for scientific data management and helps researchers plan, manage, reuse, and publish research data more systematically and reasonably.

References

[1] NEW MEDIA CONSORTIUM. EDUCAUSE. Horizon report[EB/OL].[2018-05-21]. http://cdn.nmc.org/media/2015-nmc-horizon-report-library-EN.pdf.
[2] Guo Hua. Research on research data management services in university libraries[J]. Library Science Journal, 2017(9): 81-84. DOI: 10.14037/j.cnki.tsgxk.2017.09.019.
[3] Chen Yuanyuan, Ke Ping. Research on university library research data service models[J/OL]. Information Theory and Practice.[2018-05-21]. http://kns.cnki.net/kcms/detail/11.1762.G3.20171204.1444.008.html.
[4] Association of College & Research Libraries. 2016 ACRL academic library trends and statistics for Carnegie Classifications: associates of arts colleges baccalaureate colleges master's college and institutions doctorate granting institutions[M]. Chicago: ACRL, 2016.
[5] General Office of the State Council. Measures for the management of scientific data[EB/OL].[2018-05-21]. http://www.gov.cn/zhengce/content/2018-

04/02/content_{5279272}.htm.
[6] Zhang Shasha, Huang Guobin, Di Hongyang. Research on research data management services in U.S. university libraries[J]. Library Journal, 2016(7): 59-66. DOI: 10.13663/j.cnki.lj.2016.07.008.
[7] Wu Xinnian. Research data management services in academic libraries[J]. Information and Documentation Services, 2014(5): 74-78.
[8] Xing Wenming, Wu Fangzhi, Si Li. Survey and analysis of research data management and sharing services in university libraries[J]. Library Tribune, 2013(6): 19-25.
[9] AKERS K, SFERDEAN FC, NICHOLLS N, et al. Building support for research data management: biographies of eight research universities[J]. International journal of digital curation, 2014, 9(2): 171-191.
[10] Wang Mingming, Wang Juanle, Zhao Qiang, et al. Construction experience and implications of the ICPSR scientific data center[J]. China Science and Technology Resources Review, 2017(6): 100-107.
[11] AMORIM R, CASTRO J, DASILVA J, et al. A comparison of research data management platforms: architecture, flexible metadata and interoperability[J]. Universal access in the information society, 2017, 16(4): 851-862.
[12] Chen He. Discussion on the construction of scientific data repositories in domestic universities[J]. Digital Library Forum, 2017(12): 45-51.
[13] Chen Yuanyuan, Ke Ping. Review of research on research data services in university libraries[J]. Library Work and Study, 2017(12): 67-72.
[14] Shao Yuan, Teng Wenjing. Construction of an aggregated subject service system for research data management in university libraries[J]. Library Science Journal, 2017(2): 74-77.
[15] Wang Pu. Research on sustainable development of research data management information infrastructure[J]. Library Development, 2016(8): 44-48.
[16] Nie Fengying. Design of a research service model based on research team lifecycle[J]. Library, 2016(9): 87-91.
[17] Li Hang. Construction of an academic library research data management system based on data lifecycle models[J]. Library Science Journal, 2016, (12): 19-33.
[18] Yang Zhiwei, Wei Junchao. Research on institutional repository construction based on Data Curation: a case study of Johns Hopkins University's Data Conservancy Project[J]. Library Science Research, 2016(7): 55-61.
[19] TRIPATHI SM, SHUKLA A, SONKER S. Research data management practices in university libraries: a study[J]. Journal of library & information technology, 2017, 37(6): 417-424.
[20] ROOS K, MIAS E, VAN ROOYEN J. Suggesting an institutional data repository for the University of Cape Town[R/OL].[2018-05-21]. DOI: 10.5281/zenodo.263823.
[21] Liu Xia, Rao Yan. Preliminary exploration of scientific data management and services in university libraries: a case study of Wuhan University Library[J]. Library and Information Service, 2013, 57(6): 33-38.
[22] PATEL D. Research data management: a conceptual framework[J]. Library review, 2016, 65(4/5): 226-241.

[23] MAKANI J. Knowledge management, research data management, and university scholarship: towards an integrated institutional research data management support-system framework[J]. VINE, 2015, 45(3): 344-359.
[24] UZWYSHYN R. Research data repositories: the what, when, why, and how[EB/OL].[2018-05-21]. https://digital.library.txstate.edu/handle/10877/7597.
[25] KIM S, LEE W. Global data repository status and analysis: based on Korea, China and Japan[J]. Library hitech, 2014, 32(4): 706-722.

Author Contributions

Yang Le: Research framework guidance, paper revision.
Yan Shilei: Preliminary online research, literature review and summary.
Li Hongbo: Topic selection, paper writing and revision.

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