Scientific Data Management from an Archival Science Perspective: A Postprint Based on Relevant Outcomes of International Organizations

Authors: Wang Ning, Liu Yuenan

Date: 2023-04-01T16:02:47+00:00

Abstract

[Purpose/Significance] In the context of global e-science development, scientific data management practices are increasingly exhibiting a demand for interdisciplinary thinking and methods. The application of relevant theories and methods from archival science can enhance the quality and efficiency of scientific data preservation, sharing, and reuse. [Method/Process] Employing text analysis and comprehensive integration methods, this study conducts text coding and inductive analysis of archival theories and methods and related scientific data management work documented in the literature of four international organizations: OCLC, DCC, RDA, and ICA. [Results/Conclusion] From an archival science perspective, digital document continuity assurance, context information management, appraisal and disposition, and long-term preservation provide supportive functions for scientific data management. It is recommended to improve scientific data management effectiveness through pathways such as promoting interdisciplinary cooperative dialogue, establishing a cross-institutional continuity management institutional framework, and cultivating data curators with archival expertise.

Full Text

Scientific Data Management from the Perspective of Archival Science: A Study Based on the Relevant Achievements of International Organizations

Wang Ning[1],[2], **Liu Yuenan**[1],[2],[3] [1] Electronic Records Management Research Center, Renmin University of China, Beijing 100872 [2] School of Information Resource Management, Renmin University of China, Beijing 100872 [3] Key Laboratory of Data Engineering and Knowledge Engineering, Ministry of Education, Renmin University of China, Beijing 100872

Abstract

[**Purpose/Significance**] Against the backdrop of global e-science development, scientific data management practices increasingly demonstrate a need for interdisciplinary thinking and methods. Applying relevant theories and methods from archival science can enhance the quality and efficiency of scientific data preservation, sharing, and reuse. [**Method/Process**] Using text analysis and comprehensive integration methods, this study conducted textual coding and inductive analysis of archival theories and methods and their related scientific data management work as presented in the literature of four international organizations: OCLC, DCC, RDA, and ICA. [**Result/Conclusion**] The findings indicate that archival science perspectives on digital document continuity assurance, contextual information management, appraisal and disposal, and long-term preservation provide essential support for scientific data management. The study recommends improving scientific data management effectiveness through interdisciplinary collaborative dialogue, establishing cross-institutional continuity management frameworks, and cultivating data librarians with archival expertise.

Keywords: scientific data management; archival science; international organizations; interdisciplinary; data librarian
Classification Numbers: G203, G275
DOI: 10.13266/j.issn.0252-3116.2021.05.009

Introduction

With the continuous development of e-science, scientific data sharing and reuse have become common goals for the global scientific community. Scientific data, which records scientific activity processes, possesses dual attributes as both data and archives. Theoretically, archival science can provide theoretical and methodological support for scientific data management. Under the trend of integrating scientific data and related information resources, authoritative international organizations in both scientific data management and archival fields—such as the Online Computer Library Center (OCLC), the Digital Curation Centre (DCC), the Research Data Alliance (RDA), and the International Council on Archives (ICA)—have actively promoted the application of archival theories and methods in scientific data management. These organizations have conducted series research on the critical roles of archival theories and methods in scientific data appraisal, lifecycle management, contextual information management, and long-term preservation through establishing professional interest groups, conducting surveys, and publishing guideline tools, forming valuable reference achievements.

Meanwhile, scholars both domestically and internationally have proposed exploratory arguments on the value of archival theories and methods in scientific data management from different perspectives. These include collaboration be-

tween archivists and scientists to understand data management and preservation needs [1]; the crucial role of archival principles and skills such as provenance, appraisal, authenticity, metadata, risk management, and trust in scientific data management [2]; clarifying content in metadata that reflects scientific data quality (accuracy, reliability, authenticity, etc.) [3]; promoting archivists' participation in early stages of the scientific data lifecycle [4]; conducting value appraisal of scientific data [5]; and advocating for archival professionals to play an active role in scientific data management.

However, the above research achievements are relatively scattered, lack comprehensive integration, and are mostly products of Western institutional and management environments. In China, archivists' participation in scientific data management is generally low, and the application of archival theories and methods in scientific data management remains limited. To further enhance the integration of interdisciplinary methods in scientific data management, it is necessary to synthesize relevant research achievements from both scientific data management and archival science fields, analyze the supporting role of archival science in scientific data management, and thereby promote interdisciplinary research across library, information, and archival studies, providing inspiration for relevant practices.

---

1. Research Methods and Data Sources

This study primarily employs text analysis and comprehensive integration methods. First, through website investigation, we selected relevant research achievements published by authoritative international organizations including OCLC, DCC, RDA, and ICA as key analytical texts, supplemented by representative domestic and international scholarly literature (see Table 1). We coded the archival theories and methods involved and their supporting scientific data management work (see Table 2). Based on this, we comprehensively integrated the findings according to the problems solved by archival theories and methods, inductively analyzing archival science perspectives that can support scientific data management.

Among the four key international organizations surveyed, OCLC, founded in 1967, is an online computer library center jointly built by the global library community, creating WorldCat, the world's largest online public access catalog [6]. As one of the world's largest literature information service institutions, its online computer system products and services are widely used in libraries and research institutions worldwide. Regarding scientific data management issues, OCLC has conducted a series of archival science-related research activities, including establishing professional advisory groups, designing research learning agendas for archives and special collections, and publishing the report *The Archival Advantage: Integrating Archival Expertise into the Management of Born-Digital Library Materials*, focusing on data users' needs for data context and the evolv-

ing changes in academic records management.

DCC is an internationally recognized professional research institution in digital curation, dedicated to building data management capacity and skills, aiming to provide expert advice and practical assistance to institutions that store, manage, protect, and share digital research data [7]. Its designed data curation lifecycle model has extensive international influence. DCC emphasizes the important value of archival appraisal and other professional theories in scientific data management, developing guideline tools such as *How to Appraise and Select Research Data for Curation*, *Five Steps to Decide What Data to Keep*, and *Where to Keep Research Data*, providing practical guidance for scientific data management.

RDA is a community-driven international organization initiated in 2013 by the European Commission, the U.S. National Science Foundation, the U.S. National Institute of Standards and Technology, and the Australian Government Department of Innovation. It aims to build social and technical infrastructure to achieve the goal of global scientific data open sharing and reuse [8]. RDA established the Archives and Records Professionals for Research Data Interest Group (ARPRD) [9] to explore the intersection of information science and research data management represented by archives and records management, advocating for introducing archival skills and advantages in metadata, contextual information management, appraisal, and long-term preservation into scientific data management.

ICA is the most authoritative international organization in the archival field, committed to the effective management of records and archives and the protection and utilization of world archival heritage. Its Committee on University and Research Institution Archives Section for Scientific and Research Data specializes in research on scientific data and records management in universities. The committee published *Guidelines for the Management and Preservation of Scientific Records and Data*, proposing scientific data identification and management solutions based on research processes and appraisal standards and curation strategies for long-term preservation of scientific data.

Additionally, the author conducted field research at institutions including the Institute of High Energy Physics of the Chinese Academy of Sciences, the National Geological Archives, and the National Center for Bioinformation to understand their scientific data management status and actual archival department participation, examining and consolidating the objectivity of the integrated findings from an empirical perspective.

---

2. Research Findings

2.1 Digital Document Continuity Assurance and Scientific Data Management 2.1.1 Digital Document Continuity Assurance

Digital document continuity assurance can be understood as the concept of integrated records management in digital environments, adopting consistent methods throughout the entire records lifecycle from creation to destruction or permanent preservation as archives to reduce internal losses caused by inconsistent management at different lifecycle stages and achieve optimal overall benefits. The records lifecycle theory, conceived in the 1940s, initially revealed the connections between records and archives management activities at different stages [26]. With the popularization of electronic records, Australian archival scholars F. Upward and S. McKemmish proposed the records continuum theory in the 1990s, emphasizing the integrity and continuity of records and archives management activities. This theory has resonated widely in the global archival community, promoting the development of digital continuity policies and action plans, such as the Digital Continuity Project launched by the UK National Archives in 2007, the Digital Continuity Action Plan initiated by Archives New Zealand in 2009, and the *2020 Digital Continuity Policy* released by the National Archives of Australia in 2015, which emphasize building systematic frameworks for information management in continuous information movement [27].

In China, archival scholars have gradually realized that passively waiting for business output at the back end of the records lifecycle is not conducive to quality control and long-term maintenance of archives, proposing principles of front-end control and lifecycle management [28]. They advocate intervening in management at the records creation stage (or even earlier at the system design stage) through regulations, standards, systems, and technologies, and continuously controlling the entire process of records creation, capture, preservation, disposal, organization, and utilization to continuously ensure records authenticity, integrity, and usability.

2.1.2 Support for Scientific Data Management Work

In China's scientific data management practice, data management departments generally collect and preserve scientific data only after research activities have achieved stage or final results, providing sharing and utilization services without establishing a full-lifecycle management model for data from the source. Although departments such as the National Geological Archives have explicit quality requirements for data submission, they still face issues of diverse data formats in practice, posing significant challenges to data integration and long-term preservation. The *Measures for the Management of Scientific Data* issued by the General Office of the State Council in 2018 (hereinafter referred to as the *Measures*) [29] mainly stipulates scientific data submission systems and sharing and utilization work, without explicitly requiring data standardization, collection scope, and long-term preservation in the data generation stage, failing to reflect the concepts of full-lifecycle management and front-end control. Scientific data management lacking continuity assurance has potential problems and risks such as incomplete data collection, non-standard data quality, incomplete data associations, and insufficient data utilization, which are not conducive to effective value-added and development of scientific data resources.

Establishing digital continuity management thinking and implementing full-lifecycle continuity management can fundamentally improve scientific data quality. We can learn from DCC's data curation lifecycle model and ICA's full-lifecycle data identification and management solutions that embody continuous management thinking, while encouraging archivists to participate in early stages of research activities to conduct continuous management of scientific data and improve data management quality from the front end [4].

DCC's data curation lifecycle model is a typical representative embodying continuous management thinking. According to this model, ideal data curation activities should cover the entire data lifecycle, including initial conceptual design, data creation and receipt, appraisal and selection, data acquisition, preservation and storage, access and utilization, transformation and migration, community observation and participation, and data description [16]. Among these, conceptual design involves envisioning and planning data creation activities, including setting capture methods and data storage scope; creation and receipt involve creating metadata and receiving data from data creators, other repositories, or data centers [16]. These two activities fully embody front-end control thinking for data management, embedding data management solutions and collection scope at the data formation stage, clarifying metadata requirements, and comprehensively ensuring continuity throughout the entire data lifecycle.

ICA also embodies scientific data continuity management thinking by proposing a full-process data identification and preservation management solution based on the scientific activity lifecycle. This solution summarizes the entire process of a general research project as an eight-stage cyclical process (see Figure 1 [Figure 1: see original paper]): scientific question formulation, planning, raw data collection, analysis, evaluation and auditing, results reporting, financial reporting, and generating new research [20]. Throughout the research activity process, archiving is regarded as a core activity, with data collection and preservation required at every stage except research question formulation, thereby forming a complete scientific data flow. This solution advocates conducting data management activities centered on archiving, embodying concepts of real-time capture, synchronous control, and integrated management, which help ensure the continuity of scientific data collection, preservation, and management.

2.2 Contextual Information Management and Scientific Data Management 2.2.1 New Provenance Concept and Contextual Information Management

The principle of provenance is a globally recognized archival arrangement theory and one of the pillar theories of archival science. The principle emphasizes respecting provenance, respecting the integrity of fonds, and respecting original organic connections within fonds [26]. Organizing archival information by provenance rather than by subject has become a unique method in the archival field. In the electronic records era, the principle of provenance has been challenged by new technological environments, leading to its "rediscovery" and the emer-

gence of the "new provenance concept." From the new provenance perspective, scholars have reinterpreted the concept of "provenance," breaking through the inherent understanding of "records creator provenance" or "institutional provenance" and considering the comprehensive background information of record formation—who created the file, in what functional activity, for what purpose, and using what structural form—as provenance information [26].

Contextual information is not only the basis for archival organization but also an important reference for archival appraisal. That is, archival appraisal should not only judge the value of individual records but also assess the value of a complete set of records generated from the same business activity based on contextual associations among records. From the perspective of record composition, context, content, and structure are the three elements of a record. The archival community consistently holds that context is key to making records and archives evidence of business activities and is the focus of maintaining electronic records authenticity, integrity, and comprehensibility, particularly emphasizing attention to "macro-level connections" such as functions, plans, activities, and business that demonstrate the context of record formation. Therefore, context is a core concept in archival theory, and contextual information management is one of the core skills of archival work. Archivists need to simultaneously capture contextual information such as record creation processes, usage permissions, custody conditions, and intended purposes when capturing digital files themselves.

2.2.2 Support for Scientific Data Management Work

In scientific data arrangement, classification by discipline or subject is considered common practice. However, in three cases investigated, two followed the principle of provenance. The National Geological Archives organizes all materials formed in a survey activity into a file, a practice originating from requirements for the "completeness" arrangement of scientific and technical archives. The National Center for Bioinformation organizes bioscience data according to a "Project-Sample-Experiment-Run" structure, which is highly consistent with archival methods that arrange records according to business provenance context. Although scientists and data managers at this institution are not familiar with archival theories and methods, their practical information organization approach precisely demonstrates the vitality of contextual association.

Regarding data traceability, with the continuous development of big data technology, the trend of integrating scientific data from different fields, sources, and types for comprehensive analysis to solve scientific problems has become increasingly prominent. The authenticity and credibility of these data directly affect the accuracy of analysis results, making the contextual traceability of scientific data information increasingly important. OCLC's research on data reuse satisfaction and data users' concerns about data context found that from the perspective of scientific data reuse needs, preserving contextual information about data generation is as important as preserving data content, and several data quality attributes—completeness, accessibility, operability, and credibility—have signif-

icant positive correlations with data reuse satisfaction [11-12]. Therefore, scientific data preservation should not only preserve result data and process data themselves but also preserve background information such as data software information, data providers, research project information, processing activities, and sharing and utilization. Otherwise, future researchers may not find complete materials to support new research, and records reflecting relevant research history may be incomplete [13]. OCLC points out that archivists have advantages in contextual information management, and understanding context at every stage of the records lifecycle—from preliminary investigation to data processing and metadata creation—is equally important in research fields [10]. Before scientific discovery and publication, the absence of archivists' participation may lead to loss of important information about data sources, context, and projects [4].

In digital environments, relevant contextual information is typically manifested through metadata, which is the foundation of standardized data management and an important component of data management plans. In digital archival resources management, as long as we always grasp the metadata of their formation, management, and utilization and associate it with archival resource content, we can understand their origin and development and effectively maintain the historical connections of archival resources [24]. Metadata is also a basic tool for scientific data management. However, scientific data generators—researchers—often lack the understanding that metadata "drives all steps in the data management lifecycle," resulting in insufficient supply of context-describing metadata that requires data curators to supplement and improve [2]. Moreover, contextual metadata is often generated continuously during the data management process and is difficult to supplement afterwards. ICA proposes that archival metadata statements should be designed into project planning stages, capturing workflow data as much as possible during data preservation and management processes to ensure timely and complete preservation of scientific data and their contextual information and avoid loss of valuable contextual information [20].

2.3 Archival Appraisal and Disposal and Scientific Data Management

2.3.1 Archival Appraisal and Disposal

"Appraisal is the most sublime function and the core of contemporary archival practice" [30]. In archival science, appraisal, also called value appraisal, refers to the work of determining whether original business information (i.e., records) still has preservation value after business activities conclude. Appraisal relates to the selection of archival management objects and is the most core and critical archival management activity, including assessing the value of record information and determining its value in business, institutional, legal, financial, and historical aspects and its potential future use value, thereby determining whether it falls within the scope of archiving and establishing its retention period. Archival science has developed appraisal methods with strong theoretical foundations during its long-term development, including the theory of old archives, functional appraisal, direct appraisal, and use demand forecasting methods. In the digital

era, functional appraisal has gained widespread recognition in global archival theoretical research and practice fields and has been reflected in archival appraisal policies in multiple countries including China, the United States, Australia, and Canada. Based on appraisal results, archival departments carefully design retention schedules to support the division of retention periods and disposal work, including transferring archives with long-term preservation value to archives for long-term or permanent preservation and destroying records whose retention period has expired [22].

2.3.2 Support for Scientific Data Management Work

With the rapid generation of scientific data in various research activities, the cost and benefit issues of massive scientific data storage have emerged: On one hand, despite decreasing costs of data storage media, expenses for data backup, metadata maintenance, format management, quality testing, and other data maintenance have increased exponentially. Scientific data should only be preserved continuously when its value exceeds its management costs, yet not all scientific data possesses such potential value. On the other hand, preserving all data poses enormous challenges to data retrieval and utilization; the more content preserved, the higher the signal-to-noise ratio for retrieval, and the lower the efficiency for data users to accurately obtain target data. Therefore, conducting scientific data appraisal is essential. Data managers at the Institute of High Energy Physics of the Chinese Academy of Sciences mentioned that in data management practice, although large amounts of scientific data may have preservation value, only data of significant value can be selected for preservation due to funding constraints, and the preservation duration will also depend on funding support. DCC also points out that "the scale of scientific data storage is very large, and sufficient metadata must be preserved to ensure data remains traceable, understandable, and usable over time. Considering future costs of long-term preservation and management, data creators and managers cannot avoid making appraisal decisions" [14]. However, current practice still focuses primarily on recent scientific data submission and sharing, lacking effective experience in long-term appraisal and disposal issues, and the responsible entities for appraisal are not clearly defined [5].

Archival appraisal and disposal methods can help support scientific data managers in effectively selecting data and determining its preservation value and retention period. International organization achievements have reached consensus on the value of archival appraisal and disposal methods in scientific data selection and cost-benefit trade-offs, all considering usability, reuse value, and data quality as important reference standards for scientific data appraisal. They have also explored appraisal subjects and technical operation approaches, forming valuable reference achievements, including DCC's comprehensive value assessment method and five-step implementation strategy, ICA's three appraisal standards and records and data retention and disposal plans, and OCLC's multi-stage appraisal approach.

DCC explicitly includes "appraisal and selection" as one of eight activities in the

curation lifecycle, requiring data managers to "appraise and select data for long-term curation and preservation" [16]. It recommends that "data librarians and archivists in research institutions be primarily responsible for developing selection and appraisal policies, referencing opinions from stakeholders such as data creators, data reusers, and research communities" [14]. Research institutions' appraisal policies need to specify seven criteria for evaluating dataset value: relevance to the institution's mission; scientific, cultural, or historical value of data; data uniqueness; data quality; data irreproducibility; economic costs; and completeness of description, and determine data retention periods and destruction times [14]. Additionally, DCC proposes conducting data appraisal through five specific steps: considering potential data reuse needs, examining various data indicators (ensuring legal and policy requirements are met), determining data with long-term preservation value, weighing economic costs, and developing preservation or disposal actions [15].

ICA advocates that data appraisal should follow three basic standards to ensure data credibility, validity, and quality: "1) data must be necessary for verifying results; 2) data must ensure feasible access after preservation; 3) data must have the possibility of reuse and creating new research" [20]. Based on appraisal and corresponding to scientific data generation at each research activity stage, ICA developed a records and data retention and disposal plan for each stage except research question formulation (example see Table 3), specifying in detail the types of records and data to be collected, carrier formats, retention and disposal requirements, and access restrictions, providing directly referable norms for scientific data archiving scope and retention periods [20].

OCLC proposes that appraisal can be conducted in one or multiple stages. Given that electronic records appraisal adds technical appraisal of usability status to value appraisal of content usefulness, OCLC suggests conducting value appraisal when storage institutions hand over materials with depositors or after materials are collected and preserved, and conducting technical appraisal before collection of materials containing original digital information—i.e., using appropriate digital tools to inspect whether content is damaged or tampered with [10].

2.4 Digital Archives Long-Term Preservation and Scientific Data Management 2.4.1 Digital Archives Long-Term Preservation

As an important task in information management, long-term preservation of digital information has attracted joint attention from multiple disciplines including library science, archival science, and data science, with multi-department collaboration in practice. Archival departments, based on their responsibility for long-term custody of social memory assets, are committed to ensuring the long-term availability and trustworthiness of records and archives information. After years of exploration, the international archival field has accumulated rich experience in long-term preservation of digital archives information and formed unique technical routes, such as the UK National Archives' digital format reg-

istry system PRONOM project [31], the Swiss Federal Archives' SIARD solution for long-term preservation of relational databases based on XML [32], and the metadata encapsulation scheme (VEO) adopted by Victoria, Australia [33], all of which have generated extensive international influence. Digital archives have been widely recognized internationally as important types of digital repositories, providing storage and access platforms for datasets and conducting standardized data quality control and complete lifecycle management [25].

2.4.2 Support for Scientific Data Management Work

Scientists increasingly recognize that they lack the skills and expertise needed to meet data preservation requirements and are seeking help from "data archivists," because collection, organization, and long-term preservation of archival resources are the professional mission of archival workers [1]. Many foreign research funding agencies and research management institutions have listed "data archiving and long-term preservation" as important components of data management plans [34], while China's *Measures* only provide principled requirements for scientific data preservation [29]. According to the author's investigation, although actual institutions have considered long-term preservation, they mainly adopt basic strategies such as backup, lacking application of core strategies such as migration, emulation, and preservation metadata. At the 2019 National Academic Symposium on Long-Term Preservation of Digital Resources, experts pointed out that methods and practical achievements formed by archival workers in archival management practice can provide certain references for scientific data long-term preservation.

DCC provides a guide to optional data storage solutions, pointing out that there are hundreds of repositories for data storage with different advantages and disadvantages, and that factors considered for selecting repositories for open access differ from those for long-term preservation. For open access and data sharing, options include discipline-specific data repositories, scientific data centers, general data repositories, institutional data repositories, journal supplementary material services, and websites. For long-term preservation, the focus is on cost, security, and availability of long-term preservation, recommending comprehensive consideration of options such as institutional data archives, secure centers, cloud storage, and third-party data archiving services [17].

From a professional archival perspective, the ICA guidelines propose basic standards and strategies for long-term preservation and curation of scientific data and records, emphasizing that scientific data and records management should be integrated with research activity processes to prevent losses and risks that cannot be remedied later or would be too costly [20]. RDA's APARD group also pays close attention to long-term preservation of scientific data. The theme of its 9th Plenary Meeting was "Focus on Digital Preservation" [19], and at its 11th Plenary Meeting, it proposed drafting a brief guide on digital preservation, collecting APARD members' and other groups' views on special challenges scientific data face in long-term preservation compared to other digital assets, and discussing potential updates or revisions based on the National Digital Steward-

ship Alliance's (NDSA) "Levels of Digital Preservation" document [18].

In foreign scientific data management practice, there are already cases of research institutions collaborating with archives or building digital archives for data long-term preservation, providing practical experience for archival institutions' participation in scientific data long-term preservation. For example, the University of Illinois at Urbana-Champaign Library's Research Data Service (RDS) collaborates with the university archives, committing to preserve and promote access to datasets for at least 5 years after RDS publishes data. After RDS receives research data for five years, based on archival appraisal theory, decisions are made on whether to continue retention, increase resources, or destroy [35]. The U.S. National Science Foundation-funded National Center for Atmospheric Research (NCAR) established a research data archive to support long-term preservation of irreplaceable scientific data and heterogeneous archived data spanning over 40 years, continuously updating IT technology to enhance data discovery and access capabilities and providing data management support for NCAR researchers [36].

3. Discussion

Based on international organizations' research on applying archival theories and methods to scientific data management work, and considering the current situation of low participation by Chinese archival institutions and practitioners in scientific data management and the relative absence of archival perspectives in scientific data management, this study recommends promoting interdisciplinary, cross-domain, and cross-institutional collaborative exchanges in scientific data management, fully leveraging archival science advantages, and cultivating data librarians with archival expertise to participate in developing scientific data appraisal plans, long-term preservation norms, metadata schemes, and continuity management systems, thereby promoting quality and efficiency improvements in scientific data management and sharing services.

3.1 Conducting Interdisciplinary Collaborative Dialogue in Scientific Data Management International organizations have made corresponding progress in interdisciplinary dialogue on scientific data management. For example, OCLC specifically designed a research learning agenda for archives and special collections within research library systems, establishing advisory groups composed of archives and special collections directors to understand different management issues and knowledge needs across departments and professional fields within the entire research management ecosystem, promoting publicity and development of archives and special collections resources, with expert members of advisory groups providing regular consultation and opinions throughout the research process [37]. RDA also proposes that archivists, records management professionals, and librarians have long worked together to acquire, appraise, catalog, manage, preserve, and provide access to digital and analog re-

search materials, and that these professionals possess skills and expertise that can make significant contributions to best practice development, with collaboration better serving the goals of good scientific data management and sharing [38]. In RDA alliance practice, ARPRD collaborates with the Libraries for Research Data Interest Group, holding joint meetings at the 11th Plenary to discuss cooperative projects and topics between the two groups, including research data appraisal, digital preservation, and metadata, hoping to promote development in research data management-related fields through group cooperation. The two professional groups are also committed to cooperating in developing scientific data management infrastructure and best practices to ensure datasets remain accessible and usable for five, twenty, fifty, one hundred years or longer [39].

Although China advocated integrated management and development of libraries, information, and archives since the 1980s, significant barriers still exist in disciplinary research and practice. Besides lacking cross-domain institutional cooperation, China has not yet established comprehensive research association organizations integrating information disciplines such as library science, information science, archival science, and data science. Future development could consider strengthening cooperation in this comprehensive disciplinary field. Drawing inspiration from international organizations' cooperative research and interest group mechanisms, it is recommended that China's Library Society, Archives Society, and Scientific and Technical Information Society fully utilize existing cooperation platforms or establish relevant research interest groups within the China Committee of the International Council for Science Committee on Data to strengthen exploration and innovation on issues of common concern across disciplines, such as long-term preservation, data appraisal, metadata, and data repositories, enhance understanding of other disciplines' strengths, and cooperate to promote development and improvement of relevant scientific data management infrastructure to serve the full lifecycle management of scientific data.

3.2 Establishing a Cross-Institutional Scientific Data Continuity Management Framework In archival science, continuity management thinking emphasizes not only full-lifecycle continuity of information objects but also management continuity from records creation units to archives, forming a complete cross-institutional institutional framework that can also provide a full-lifecycle continuity management perspective for scientific data management, promoting multi-stakeholder cooperation in scientific data management and strengthening management coherence.

From a resource perspective, integrated management and services of scientific data and research archives need to be promoted. Scientific data and research archives overlap in objects, but their management in China has long been in a "fragmented" state, with neither archival management links incorporated into scientific data management processes nor research archives sharing and utilization incorporated into scientific data sharing and utilization. Scientific data management and archives management in research institutes are generally un-

dertaken by different functional management departments, with significant division differences and different business focuses, without forming a complete chain from scientific data generation to archival preservation. As the international scientific community increasingly emphasizes integrated sharing and utilization of research results, scientific process data, and research management archives, there is also an objective demand for integrated services of research archives stored in archives and scientific data stored in scientific data centers. Therefore, it is necessary to establish collaborative work systems between scientific data management institutions and archival departments, with archives and scientific data centers negotiating and researching data submission formats, data submission norms, data management plans, and long-term preservation plans [25], promoting integrated development of both works. Simultaneously, continuous advancement of research archives' datafication and resource integration services should break down "information silos" and enhance the overall level of scientific and technical information resource management and services. Currently, only a few institutions in China, such as the National Geological Archives, have integrated management functions for scientific and technical information resources, undertaking archives functions while conducting scientific data management. Such institutions' collaborative development models should be strongly supported and promoted to facilitate coordinated development of scientific data and research archives.

From a management perspective, scientific data management is not merely the task of research institutions; its generation to preservation may require cross-institutional implementation, also needing to build a cross-institutional management framework. The foreign scientific data management field has already regarded archives as an important type of data repository. With continuous construction of digital archives in China, archival institutions also possess certain capabilities for long-term preservation of digital information resources and have established relatively mature long-term preservation technical strategies. They can serve as participants in scientific data repositories, jointly undertaking scientific data preservation and management work with scientific data centers, especially for scientific data with important social, historical, and cultural value, which can be transferred to archives for preservation. On this basis, archival workers will have opportunities to become repository managers, data librarians, or data scientists, thereby applying their skills and expertise to participate in scientific data management work [22].

3.3 Cultivating Data Librarians with Archival Expertise With the rapid development of e-science and open science and increasing scientific data management needs, a new position type has emerged in scientific data management institutions such as research institutions, research funding agencies, academic libraries, and information centers: data librarians who implement scientific data management, conduct data curation, and serve data open utilization. Although as an emerging profession, the academic community has not yet provided a consistent definition, high demand has already emerged in the scien-

tific data management practice field. Among 64 recruitment postings collected by the International Association for Social Science Information Services and Technology in 2017, 41 positions were related to data librarians [40]. Gu Liping et al. propose that "data librarians in the open science environment should be data management professionals who apply library work principles, possess scientific data management knowledge and skills, understand open science operation mechanisms, and have background knowledge in specific research fields" [40]. This definition first emphasizes the importance of library work principles in scientific data management but does not explicitly state the necessity of archival theories and methods. The author believes that archival knowledge and skills should automatically be included within "scientific data management knowledge and skills," meaning fully understanding digital document continuity management thinking, understanding contextual information management needs, being familiar with data appraisal principles, and mastering data long-term preservation skills.

Therefore, it is necessary to strengthen the cultivation of data librarians with archival expertise. First, training and guidance on archival science-related knowledge and skills can be organized for existing data librarians in research institutions, research funding agencies, academic libraries, and information centers to broaden their perspectives on applying archival thinking to scientific data management. Successful cases include data curation and training programs offered by the U.S. National Archives and Records Administration and specialized training services for social science scholars provided by the UK Data Archive [25]. Additionally, degree programs in library science, archival science, data science, and information resource management at higher education institutions and research institutes can provide required and elective courses on data curation and archival science for students interested in scientific data management work to perfect their knowledge structure and cultivate comprehensive data management literacy. Many internationally renowned information schools, such as those at UCLA, Indiana University, Simmons College, University of Maryland, and University of Michigan, have established courses on digital records and information management, digital preservation, data curation, metadata, and trusted digital repositories in their library and information science or archival science master's programs, focusing on comprehensively cultivating students' digital information curation skills. Finally, when employing data librarians, relevant institutions should include graduates with archival science education backgrounds or personnel with archival work experience in their recruitment scope to enrich the talent structure of scientific data management. Only by incorporating archival knowledge and skills into the training and employment demand framework for data librarians can archival expertise truly be leveraged to better serve and improve scientific data management work.

References

[1] DHARMA A, ANNZ, MORGAN D, et al. The application of archival concepts to a data-intensive environment: working with scientists to understand data management and preservation needs[J]. Archival Science, 2011, 11(3/4): 329-348.

[2] ALEX H. How has your science data grown? Digital curation and the human factor: a critical literature review[J]. Archival Science, 2015, 15(2): 101-139.

[3] TRACEY P, BARBARA L, FRASER T, et al. Today's data are part of tomorrow's research: archival issues in the sciences[J]. Archivaria, 2007, 64(Fall): 123-179.

[4] JULLIAN C, CHRISTINE L, MATTHEW S. Moving archival practices upstream: an exploration of the life cycle of ecological sensing data in collaborative field research[J]. The International Journal of Digital Curation, 2008, 1(3): 114-126.

[5] Deng Jun, Song Wenfeng. Research progress on scientific data value appraisal[J]. Information Science, 2012(6): 943-946, 958.

[6] WorldCat.org: The world's largest library catalog[EB/OL]. [2020-06-24]. https://www.worldcat.org.

[7] About DDC[EB/OL]. [2020-05-21]. https://www.dcc.ac.uk/about.

[8] About RDA[EB-OL]. [2020-05-22]. https://www.rd-alliance.org/about-rda.

[9] IG Archives and records professionals for research data[EB/OL]. [2020-05-25]. https://www.rd-alliance.org/ig-archives-and-records-professionals-research-data.html.

[10] The archival advantage: integrating archival expertise into management of born-digital library materials[EB/OL]. [2020-06-18]. https://www.oclc.org/research/publications/2015/oclcres archival-advantage-2015.html.

[11] Social scientists' satisfaction with data reuse[EB/OL]. [2020-06-15]. https://www.oclc.org/research/publications/2015/social-scientists-satisfaction-with-data-reuse.html.

[12] Context from the data reuser's point of view[EB/OL]. [2020-06-16]. https://www.oclc.org/research/publications/2019/context-from-data-reuser-point-of-view.html.

[13] The evolving scholarly record[EB/OL]. [2020-06-16]. https://www.oclc.org/research/publications/2014/ocl evolving-scholarly-record-2014-overview.html.

[14] How to appraise and select research data for curation[EB/OL]. [2020-06-12]. https://www.dcc.ac.uk/guidance/how-guides/appraise-select-data.

[15] Five steps to decide what data to keep[EB/OL]. [2020-06-20]. https://www.dcc.ac.uk/guidance/how-guides/five-steps-decide-what-data-keep.

[16] Curation lifecycle model[EB/OL]. [2020-06-17]. https://www.dcc.ac.uk/guidance/curation-lifecycle-model.

[17] Where to keep research data[EB/OL]. [2020-06-17]. https://www.dcc.ac.uk/guidance/how-guides/where-keep-research-data.

[18] Joint meeting: IG Archives and Records Professionals for Research Data, and IG Libraries for Research Data Interest Groups Plenary 11[EB/OL]. [2020-06-18]. https://www.rd-alliance.org/archives-and-records-professionals-research-data-ig-rda-plenary-11-berlin.

[19] RDA P9 Archives & Records IG: notes & slides[EB/OL]. [2020-05-30]. https://www.rd-alliance.org/group/archives-and-records-professionals-research-data-ig/post/rda-p9-archives-records-ig-notes.

[20] New handbook: management and preservation of scientific records and data[EB/OL]. [2020-05-22]. https://www.ica.org/en/new-handbook-management-and-preservation-scientific-records-and-data.

[21] GRANT R. Record keeping and research data management: a review of perspectives[J]. Records Management Journal, 2017, 27(2): 159-174.

[22] CHILDS, MCLEOD J, LOMASE, et al. Opening research data: issues and opportunities[J]. Records Management Journal, 2014, 24(2): 142-162.

[23] BORGERUD C, BORGLUND E. Open research data, an archival challenge?[J]. Archival Science, 2020: 1-24.

[24] Mao Tianyu. Analysis of the application and enlightenment of archival theory in digital curation research[J]. Archives Science Bulletin, 2016(1): 34-38.

[25] Yan Peng. Analysis of archival departments' participation in scientific data management from a stakeholder perspective[J]. Archives World, 2019(3): 23.

[26] Feng Huiling. Introduction to Archival Science[M]. Beijing: Renmin University of China Press, 2006.

[27] Zhou Wenhong, Zhang Ning. Global digital continuity action panorama and enlightenment—based on discussion of national policies in the UK, New Zealand, Australia, and the USA[J]. Information Theory and Practice, 2017, 40(3): 138-142, 137.

[28] Li Ming. Records management: the core of front-end control and lifecycle management[J]. Zhejiang Archives, 2005(11): 7-8.

[29] Notice of the General Office of the State Council on issuing the Measures for the Management of Scientific Data[EB/OL]. [2020-06-12]. http://www.gov.cn/zhengce/content/2018-04/02/content_{5279272}.htm.

[30] CAROL C. Archival appraisal: a status report[J]. Archivaria, 2005, 59: 83-107.

[31] Zhang Ning, Yang Jingjing. Comparative study of foreign typical digital format registry systems—taking PRONOM, GDFR, and UDFR as examples[J]. Beijing Archives, 2015(9): 17-20.

[32] Qian Yi, Liu Lichao. Research on technical paths for database electronic records archiving and long-term preservation[J]. Archives Science Study, 2017(4): 67-72.

[33] Liu Yuenan. Rethinking electronic records metadata encapsulation strategies—research triggered by changes in VERS standards[J]. Archives Science Study, 2019(4): 116-123.

[34] Liu Feng, Zhang Xiaolin. Research on data management plan composition specifications and their operable data curation models[J]. New Technology of Library and Information Service, 2016(1): 11-16.

[35] Processing digital research data[EB/OL]. [2020-11-19]. https://saaers.wordpress.com/2016/05/11/processi digital-research-data/.

[36] Research data archive, National Center for Atmospheric[EB/OL]. [2020-06-15]. https://rda.ucar.edu/#!about.

[37] Research and learning agenda for archives, special, and distinctive collections in research[EB/OL]. [2020-06-15]. https://www.oclc.org/research/publications/2017/oclcresearch-research-and-learning-agenda.html.

[38] RDA and librarianship, archival science and information science | RDA[EB/OL]. [2020-06-22]. https://www.rd-alliance.org/rd-and-disciplines/rd-and-librarianship-archival-science-and-information-science.

[39] 23 Things: libraries for research[EB/OL]. [2020-06-22]. https://rd-alliance.org/group/libraries-research-data-ig/outcomes/23-things-libraries-research-data-supporting-output.

[40] Gu Liping, Zhang Xiaoyue. Analysis of data librarians' practice in the open science environment[J]. Library and Information Knowledge, 2020(2): 60-74, 112.

Author Contributions

Wang Ning: Proposed research questions, developed research framework, wrote the paper.
Liu Yuenan: Developed research framework, revised research framework, wrote and revised the paper.

Abstract

[**Purpose/significance**] In the context of global e-science development, scientific data management practices have increasingly shown a desire for interdisciplinary thinking and methods. The use of relevant theories and methods in the field of archives can help improve the quality and efficiency of scientific data preservation, sharing, and reuse. [**Method/process**] By use of text coding analysis and comprehensive integration method, the archival theories and methods and their involved scientific data management work were extracted and inducted from the research achievements of four international organizations including OCLC, DCC, RDA and ICA, as well as other related literature. [**Result/conclusion**] It is found that the methods of archival science include appraising and disposal, digital continuity, context management, long-term preservation are necessary to carry out scientific data management. It is recommended to improve the effectiveness of scientific data management by conducting interdisciplinary cooperation dialogues, establishing cross-agency continuity management regulation framework, and cultivating data librarians with archival expertise.

Keywords: scientific data management; archival science; international organizations; interdisciplinary; data librarian

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