A Review of International Data Curation Research and Practice Development

Authors: Zhenxin Wu, Chen Yao, Li Wenyan, Fu Honghu, Xu Liyuan, Wu Zhenxin

Date: 2016-01-25T00:00:00+00:00

Abstract

Through investigating the strategic plans of major international institutions, this study identifies the principal challenges of Data Curation in management, resource development, and technical infrastructure. In response to these challenges, it provides a comprehensive review of the evolution of international Data Curation research and practice across four dimensions: strategic planning, data evaluation and selection policies, key technologies, and auditing and certification. The analysis explores potential areas for library involvement in scientific data curation within the big data research environment, seeking development opportunities for libraries in Data Curation activities.

Full Text

Review on the International Development of Research and Practice of Data Curation

Wu Zhenxin[1], Chen Yao[12], Li Wenyan[12], Fu Honghu[1], Xu Liyuan[1]
([1] National Science Library, Chinese Academy of Sciences, Beijing 100190; [2] University of Chinese Academy of Sciences, Beijing 100190)

Keywords: scientific data, research data, curation, preservation, review, challenges, development opportunities

Abstract: This paper summarizes the current challenges of Data Curation in management, resource development, and technology infrastructure based on a review of major research institutions' strategic plans. It then comprehensively reviews the developments in Data Curation research and practice in four areas: strategic planning, data appraisal and selection policies, key technologies, and audit and certification. Finally, the paper identifies potential domains where libraries can participate in Data Curation activities and explores development opportunities for libraries in this field.

1 Introduction

The development of information technology has triggered an explosion in data and information volume, giving rise to a new scientific research paradigm—e-Science. Dr. Jim Gray termed this new data-driven research approach the "Fourth Paradigm" of scientific research, marking a shift from computation-centric to data-centric science where data has become the soul of research. Today, data is no longer merely an object for collection and storage; it has transformed into a fundamental national strategic resource that can be leveraged collaboratively to address problems across numerous domains. For instance, in the case of Malaysia Airlines Flight MH370, the Institute of Remote Sensing and Digital Earth of the Chinese Academy of Sciences used its archived remote sensing satellite data to compare with satellite imagery of the suspected crash site, successfully identifying several oil slick areas that proved crucial for locating the wreckage.

Data Curation is often translated as data stewardship or data preservation, with "Data" primarily referring to scientific research data. The field has many definitions, among which the one from the UK' s Digital Curation Centre (DCC) is particularly representative: "Digital curation involves maintaining, preserving, and adding value to digital data throughout its lifecycle. These activities enhance the long-term value of existing data; proactive management reduces threats to data reuse and mitigates risks from digital technology obsolescence. Furthermore, Data Curation activities enable data held in trusted repositories to be shared more widely with research institutions to support future research activities." Data Curation is a product of the e-Science environment, emerging from the need for research data sharing and large-scale scientific computing, and represents an essential management practice for addressing data management and preservation demands in the "big data" era. To meet new research service demands and establish effective models, functions, and mechanisms for serving digital research, the National Science Library of the Chinese Academy of Sciences has conducted a series of strategic investigations. This paper presents findings from the "New Technologies and Methods for Data Management and Infrastructure Development" component of that work.

2 Analysis of Challenges Facing Data Curation

As the scale and variety of research data continue to expand, traditional data preservation methods can no longer meet current needs. Although an increasing number of institutions are engaged in Data Curation to varying degrees, as an emerging research field, Data Curation still faces numerous problems and challenges.

The National Digital Stewardship Alliance (NDSA) summarized the issues and challenges in the data management field in its 2015 agenda as follows: (1) key

issues in building digital content collections, including global digital content problems, large-scale content selection methods, and challenges from special-format digital content; (2) enhanced research needs on cost and value due to insufficient resources supporting preservation activities; (3) lack of adequate digital management personnel; and (4) development of technical infrastructure, including the urgent need to coordinate distributed service ecosystems, develop file format action plans, and ensure content integrity [2].

The UK Data Archive identified challenges in its 2010–2015 strategic plan as: (1) establishing and issuing storage certifications; (2) ensuring multi-source funding, synchronizing with user expectations and technical requirements, and promoting collaborative development in planning; (3) establishing more effective management structures and internal record management systems; (4) developing effective tools for data (collection) selection, acquisition, ingestion, and preservation to improve data quality and package effectiveness, and developing self-archiving capabilities; and (5) developing new data access models, distribution and visualization tools, reconstructing data registration and licensing systems, and integrating relevant data services [3].

The DCC summarized challenges for Data Curation over the coming decades as: (1) development of data management software; (2) consistency in reviewing commitments in data management plans; (3) impact of limited-term data preservation strategies (management evaluation); (4) identifying data resources that should be preserved; (5) data intellectual property rights; and (6) understanding true semantic long-term preservation [4].

These perspectives reveal that challenges and issues in Data Curation for the foreseeable future concentrate in several areas: management aspects primarily involve strategic planning, cost research, personnel and training, intellectual property rights, and audit and certification; resource development focuses on large-scale data selection and preservation of special-format resources; and technical infrastructure development centers on data organization, format management, data quality assurance (integrity protection), preservation systems (tools), and architectural development.

3 Developments in Data Curation Research and Practice

In recent years, numerous institutions and projects have conducted extensive and in-depth research and practice in Data Curation. Based on the aforementioned challenges, this paper summarizes and analyzes relevant research and practical activities undertaken by various institutions and projects to address these challenges. Due to the authors' research scope limitations, this paper does not cover education and training or intellectual property rights aspects.

3.1 Strategy and Planning

The formulation of strategies and plans is a critical first step in implementing Data Curation. These strategic plans include policy planning, sustainable

development strategies, and collaboration strategies. Currently, international research on global policy planning and collaboration strategies for Data Curation is relatively mature, yielding strategic frameworks, solutions, and tools with practical reference value. However, research on sustainable development strategies remains in its initial stages, with only a few research outcomes on cost studies that are insufficient to support preservation practice.

3.1.1 Data Curation Policy Planning In policy planning, the DCC provides extensive reference materials and action guidelines, offering a framework for developing research data management policies [5] that includes five steps: (1) listing existing management frameworks; (2) creating a management content table; (3) obtaining management support; (4) consultation, drafting, and revision; and (5) approval and implementation.

The MaRDI-Gross project also provides solutions for developing digital management plans (DMPs) in "big science" contexts [6], proposing a practical process framework for DMPs from six aspects: (1) establishing preservation objectives; (2) data publication plans; (3) data validation; (4) preservation of software and services; (5) costs and cost models; and (6) modeling data loss.

Currently, several mature Data Curation planning tools are available, including DMPonline developed by the DCC, DMPtool developed by UC3, CARDIO developed by IDMP, Plato developed by SCAPE, and OpenDOAR.

3.1.2 Collaboration Strategy Planning The exponential growth of data volume and increasing complexity of data types pose increasingly severe challenges to Data Curation. To address these issues and mitigate preservation risks, cross-domain collaborative action plans are urgently needed. The Digital Curation Unit (DCU) helps solve Data Curation problems by promoting interdisciplinary collaborative research planning and action plans, proposing a six-aspect action plan [7]: (1) using a lifecycle approach to manage curated information objects that includes dynamic interaction with designated communities; (2) adopting an event-centric approach to fully represent data "activity events" ; (3) broadly defining Data Curation practitioners to include those involved in the public dissemination and utilization of information objects; (4) identifying a fundamental interdisciplinary scope that enables Data Curation to fully meet differentiated disciplinary needs; (5) treating explanatory content related to information objects as community digital memory and conducting simulated archiving; and (6) advocating institution-oriented approaches to curation.

With the development of collaboration policies, a series of effective collaborative practices have positively impacted various aspects of digital curation, such as promoting collaboration in open-source software development, sharing personnel and resource information, participating in standards and practice development, coordinating digital curation responsibilities, and developing collaborative selection decisions and digital collection policies. Notable examples include the

International Internet Preservation Consortium (IIPC), whose members collaborate to develop a series of open-source tools and support sustainable shared maintenance models.

Meanwhile, relevant collaborative organizations continue to emerge, such as the global CLOCKSS network, which uses distributed, geographically diverse preservation models to ensure the complete preservation of common digital assets within organizations; Data-PASS, a voluntary alliance of institutions aimed at archiving, cataloging, and preserving data used in social science research; MetaArchive, a digital preservation network created by numerous memory institutions that is also a secure and cost-effective repository; and the Digital Preservation Network (DPN), which prevents catastrophic loss due to technological, organizational, or natural disasters by preserving dataset copies across different nodes. These organizations and their demonstrated multi-institutional management approaches have significantly increased in usage and social recognition.

3.1.3 Sustainable Development Planning Completing digital management tasks requires appropriate resource support, but insufficient resources exist to enable repositories to preserve all data. How to effectively budget, manage, and allocate curation costs and how to obtain required resources have become important issues for sustainable development. However, due to the inherent complexity of Data Curation and its involvement of multiple stakeholders, digital management cost estimation remains complicated and ambiguous, with almost no models supporting comparative or longitudinal cost estimation data.

The 4C project (Collaboration to Clarify the Costs of Curation), funded by the EU, primarily addresses preservation cost issues. It analyzed ten existing cost models and tools, evaluating each one. By examining previous digital preservation cost modeling work, the project proposed best practice recommendations for establishing sustainable digital preservation and access. Currently, 4C provides cost model tools and frameworks that attempt to address benefits, risks, value, quality, and sustainability, and has preliminarily developed an economic sustainability reference model and a curation cost exchange platform tool –CCEx.

The POWRR project represents another important initiative researching long-term preservation of digital objects with limited resources. It aims to help small and medium-sized institutions that struggle to implement digital curation due to resource constraints. The project is evaluating tools and services that can achieve long-term digital preservation in such institutions to provide effective solutions.

These project outcomes will help clarify costs and support decision-making and strategic planning, which in turn can promote long-term management of digital preservation and the development of sustainable infrastructure.

3.2 Data Appraisal and Selection Policies

The characteristics of digital data make its collection exceptionally complex and consequently complicate its preservation. Data scale continues to expand, and data granularity and interconnectivity become increasingly intricate. While traditional resource appraisal and selection are typically based on institutional priorities, capabilities, and guiding policies, digital data has unique characteristics that render corresponding appraisal and selection policies more complex.

The NDSA proposed a series of recommended practices for data appraisal and selection, covering data relevance, documentation, funding, research and application needs, availability, risk, and ease of use, which help institutions launch digital management plans involving the entire information lifecycle.

The DCC proposed a framework for selecting and appraising data for curation [8] that uses a weak analytical framework to assist in determining which data requires curation, considering factors including: (1) data whose future reuse value is difficult to assess; (2) pre-disciplinary data; (3) quality of data and related documentation; (4) irreplaceable observational data (as opposed to experimental data); (5) cost of regenerating experimental data; and (6) estimated cost of preserving specific datasets.

In 2012, the Natural Environment Research Council (NERC) published the NERC Data Value Checklist to help research communities select data requiring preservation.

Research practice demonstrates that vast amounts of digital data permeating daily life, culture, and academia remain inaccessible to libraries or archives. Therefore, such born-digital materials should be prioritized in selection policies, with active acquisition of special born-digital materials (such as web archives, digital records, and hard drives from document and manuscript archives). Additionally, selection of digital materials is often related to institutional strength and mission.

3.3 Key Technologies

3.3.1 Development of Metadata Standards and Specifications Metadata has always been an important focus area in Data Curation. Many renowned institutions and projects have developed their own metadata standards or recommended specifications.

The four levels defined in the NDSA' s "Levels of Digital Preservation" include different types of metadata in the Data Curation process: record-keeping, administrative, descriptive, structural, technical metadata, and preservation metadata.

Information released by the DCC on disciplinary metadata standards (concepts, user communities, and usage methods) has attracted significant attention from the Research Data Management (RDM) community, leading to the creation of

a disciplinary metadata webpage [9] to help users determine which metadata standards meet their needs.

Wayne State University proposed a contextual metadata framework for digital preservation of cultural heritage objects, consisting of eight contextual dimensions that identify the types of information needing capture. This framework ensures sufficient contextual information is recorded in a metadata scheme, greatly facilitating future search, examination, utilization, management, and preservation activities.

Research Data @ Essex established a three-tier metadata model based on the IDMB project' s metadata model.

In April 2013, the UK published a metadata application profile and guidelines for its repositories (RIOXX).

The US sound recording metadata scheme development project created a standard method for collecting and managing metadata for recorded music and developed a tool (Content Creator Data Tool, CCD) to help data producers and owners collect data.

3.3.2 File Format Identification, Selection, and Conversion The stability of digital file formats and the risk of format obsolescence pose major challenges for digital management institutions, particularly in big data research environments where selecting an appropriate data format for curation is a challenging and forward-looking task. Faced with accumulating large digital collections, practical strategies and means for monitoring and mining information from heterogeneous native digital documents under institutional management are particularly important.

To prevent file format obsolescence, the European Fusion Development Agreement (EFDA) proposed clear solutions in its Data Curation practice [10], stating that repositories should preserve core information about all used file formats and record which data uses these formats, with this core information updated regularly. When selecting a format for Data Curation, considering only the format' s current state is insufficient; its long-term characteristics and future development potential must also be evaluated.

The US National Archives and Records Administration' s "Format Action Plan for Publicly Released Materials" promotes practical development by encouraging digital content producers to select a more precise set of digital formats, particularly for departments that can achieve centralized control to some extent, such as federal, state, local, and regional governments.

The "Geospatial Archiving and Preservation Partnership (GeoMAPP)" project supported by NDIIPP provides a geospatial data file format reference guide offering quick reference for common geospatial raster and vector dataset types and serves as a utility for rapidly identifying common geospatial file format types used by state governments.

The NDSA recently published a research report on the PDF/A format standard, analyzing the characteristics of PDF/A, once considered one of the gold standard formats for long-term preservation, and its impact on long-term preservation.

The Library of Congress released recommended format specifications for long-term preservation, and the Florida Digital Archive (FDA) also published its own format selection scope. Archivematica has taken a crucial practical step by transforming format strategies and action plans into actions directly implemented and managed by tools and software on its software platform.

Relevant available tools include the UK National Archives' file format management tool system PRONOM and the Global Digital Format Registry (GDFR). Open-source tools for format identification, validation, and feature extraction include JHOVE (LGPL), DROID, Fuzzy Logic for document format damage analysis, and related PDF validation tools and methods.

3.3.3 Verification of Data Fixity and Integrity One of the most important tasks in Data Curation is ensuring data fixity and integrity, with data validation playing a crucial role in ensuring data trustworthiness. The common method for verifying data fixity and integrity is checking fixity information, which can detect whether data has been corrupted, monitor hardware degradation, meet trustworthiness requirements (such as ISO 16363/TRAC, NDSA Levels of Digital Preservation), support documentation provenance and chain of custody, and help diagnose potential system or human errors that may occur during the Data Curation management cycle.

Fixity checks are generally divided into two categories: (1) statistical fixity checks, which verify fixity by counting document numbers and file sizes; and (2) content fixity checks, which typically use algorithms to compare and calculate document content to determine whether it has changed.

Stanford University' s LOCKSS system uses an Opinion Polls mechanism that employs multiple nodes preserving the same content for regular content comparison and monitoring.

The Fedora Repository uses MD5 to verify digital object fixity, generating and storing MD5 checksums for each datastream segment and version of archived objects to facilitate fixity verification.

The DAITSS system regularly calculates checksums for all document copies using MD5 and SHA1 algorithms.

UC3's Merritt repository provides multiple types of interfaces in a micro-services manner and supports various common digest types, allowing fixity verification to be implemented at any time through configuration services.

Commonly used tools and algorithms for generating and verifying fixity information include Expected File Size, Expected File Count, CRC, MD5, SHA1, and SHA25. Currently, tools specifically developed for long-term preservation

include the open-source ACE (Auditing Control Environment) tool developed by the University of Maryland' s ADAPT project and the ontology tool vplan under development for dataset validation.

3.3.4 Data Unique Identifiers and Registration How to uniquely identify massive amounts of data is a key issue facing Data Curation institutions. Curators choose to adopt universal identifier systems to maintain consistency with traditional resources, including ARK (Archival Resource Key), DOI (Digital Object Identifier), Handle (Handle System Identifier), URN (Uniform Resource Name), PURL (Persistent Uniform Resource Locator), and URI (Uniform Resource Identifier).

Specialized research data registration services have also emerged. The Australian National Data Service' s (ANDS) Cite My Data service helps research institutions automatically assign DOIs to cited research datasets. Other systems providing identifier services for data include DataCite developed by the British Library, EZID developed by UC3, and WebCite.

3.3.5 Preservation Technology Strategies Over years of preservation research and practice, diverse and more practically applicable technology strategies have gradually emerged. The authors have previously provided detailed introductions and reviews [11]; this paper only supplements subsequent developments.

Bit preservation is generally considered the simplest and most understandable preservation method and is widely adopted. Format conversion and migration are also effective technical strategies currently used by many projects. Emulation is considered the most important future measure to ensure data usability, but due to high investment requirements, technical difficulty, and usage barriers, only a few projects are currently conducting related research.

The KEEP project supported by the EU' s Seventh Framework Programme proposed an "emulation as a service" approach, releasing an Emulation Framework that allows users to access old computing files and programs through emulation, currently applied to CD data and web information emulation services.

The SCAPE project has conducted extensive research based on evidence foundations of format migration, format risks, and repository performance.

3.3.6 Large-Scale Data Preservation Systems and Infrastructure The rapid growth of massive data, the speed (frequency) of data object (collection) updates, and the diversity (heterogeneity) of data objects pose enormous challenges to large-scale data preservation systems and infrastructure.

The SCAPE project primarily addresses intensive computing and preservation platform scalability issues, conducting research through three sub-projects: large-scale digital archiving, scientific datasets, and web archiving, mainly processing scientific data and scientific workflows. In addressing big data

challenges, SCAPE has achieved initial results, providing practice-based solutions and constructing a data-centric distributed SCAPE long-term preservation platform that can provide infrastructure for large-scale data execution processes.

UC3' s Merritt system for big data storage adopts a "micro-services" development model, enabling system scale and functionality to expand and update in a modular micro-services pattern. The small and independent characteristics of micro-services make them easier to develop, deploy, maintain, and upgrade, endowing Merritt with ideal features for big data preservation systems, such as high service availability, reliability, efficiency, adaptability, and sustainability.

Stanford University' s LOCKSS system employs a typical distributed storage approach, providing libraries with an open-source distributed storage system for locally collecting and managing electronic resources. LOCKSS achieves reliable preservation of large amounts of digital resources through multi-institutional participation and multi-copy storage mechanisms.

Chronopolis, a collaboration between SDSC, the University of California San Diego Library, the National Center for Atmospheric Research (NCAR), and the University of Maryland, provides the largest collaborative preservation environment in the US, using grid technology to offer monitoring, maintenance, and archival management of massive data across multiple sites and platforms.

Archive-It is a non-profit project—the Internet Archive' s web archiving service —that helps institutions capture, build, and preserve digital content collections.

Portico is a digital archive supported by the world' s largest digital archiving community, providing a sustainable business model to help libraries, publishers, and funders collaboratively preserve electronic journals, e-books, and other electronic scholarly content.

The DuraCloud service solves digital content secure storage infrastructure problems for libraries and research institutions in a cost-effective agency manner by utilizing numerous cloud storage providers (both commercial and non-profit).

3.3.7 Summary The above demonstrates that key technology development has always been an important research and development theme in advancing Data Curation. Through years of effort, Data Curation has achieved fruitful results in key technology research and practice.

In metadata standard development, many projects have proposed and defined metadata frameworks and specifications that meet the special needs of data curation based on existing standards and specifications. This integrated and convergent approach better ensures rapid satisfaction of preservation practice requirements while guaranteeing metadata standard usability. Format management, as a crucial preservation task, has seen multiple institutions release recommended format collections suitable for preservation for different data types, alongside many open-source format validation tools. Through mechanisms like

format registries, the field has relatively maturely addressed format obsolescence and conversion issues. Data integrity verification, as an effective means to ensure long-term data authenticity and usability, is addressed in the Data Curation field by adopting existing mature technical methods and developing comprehensive mechanisms tailored to practical needs. Preservation technology strategies have received relatively little investment and research in recent years, with only a few projects conducting in-depth research on emulation technology and minimal other research. To cope with continuously expanding data scale, many institutions have explored and developed numerous systems and infrastructures suitable for large-scale data preservation with flexible and scalable characteristics, attempting to solve fundamental digital storage problems from various perspectives and levels.

3.4 Developments in Audit and Certification

After recent vigorous development, research and practice in Data Curation audit and certification have made certain progress, with many trustworthy content management processes receiving recognition and standardization, and several international standards have been formed.

The RLG' s 2007 publication "Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)" became an ISO international standard (ISO 16363) in 2009. The German nestor' s "Catalogue of Criteria for Trustworthy Digital Repositories" became a German national standard in 2011.

The Netherlands' DANS project launched a digital seal of approval granting service, providing 16 guidelines for repositories to conduct self-assessment. Based on the above three standards, the EU proposed a three-level certification framework including basic certification (self-assessment according to DSA), extended certification (organized external audit according to ISO 16363 or DIN 31644 with public self-assessment), and formal certification (full certification according to ISO 16363 or DIN 31644).

The DCC developed a "Digital Repository Audit Method Based On Risk Assessment" (DRAMBORA) based on TRAC and nestor metrics while introducing risk management concepts.

The National and State Libraries Australasia (NSLA) proposed a five-level preservation capability maturity model (initial, repeatable, defined, managed, optimizing) based on Carnegie Mellon University' s Capability Maturity Model (CMM) to assess member libraries' long-term preservation activities.

To assist institutions conducting long-term preservation in selecting preservation solutions, Tessella company proposed a Digital Preservation Maturity Model for identifying characteristics of different types of long-term preservation solutions.

The "Levels of Digital Preservation" released by NDSA is a hierarchical technical practice guide designed to provide clear technical baseline descriptions for preserving digital content while allowing institutions to assess preservation levels

for their specific curated resources.

Despite these research and practical achievements, much work remains to be done. Currently, no certification process has gained widespread recognition in the preservation community. Research on the reliability of centralized and distributed preservation networks has just begun, and developing a comprehensive, robust trust framework for preservation networks remains a major challenge.

Data has revolutionized scientific research paradigms, and research data curation presents both opportunities and challenges for libraries to develop new services. Libraries can not only actively participate in the e-Science environment but also leverage their strengths to provide crucial support for research data curation. Winston Tabb, Dean of Libraries at Johns Hopkins University, stated: "In the e-Science environment, libraries are part of a distributed network, data can become collection resources, data centers become new types of library stacks, and librarians are data scientists who can provide data services" [12]. Libraries should respond to demands, seize opportunities, and create new models, functions, and mechanisms that effectively serve digital research.

Libraries can research and explore solutions for research data preservation management in big data research environments based on the research data lifecycle. Key research areas include:

- **Research Data Curation Planning:** Every research institution needs to develop its own Data Curation policy based on actual needs to clarify its responsibilities in research Data Curation and use the policy as an implementation framework to guide specific research Data Curation actions, including data selection policies.
- **Collaboration Models and Sharing Mechanisms:** Data Curation actions should follow the research data lifecycle, closely integrate with research activities, and seamlessly embed into research workflows to effectively support and promote research output, innovation, and sharing. Therefore, it is necessary to build long-term collaboration and sharing mechanisms that seamlessly embed into research workflows and work closely with research teams. How to achieve effective collaboration and sharing while respecting intellectual property rights and complying with policies and regulations involves multiple issues including policies, regulations, and technologies. Relevant policy incentives, copyright protection, and privacy protection for research data are important issues that must be considered in collaboration and sharing mechanisms.
- **Service Content and Mechanisms:** Research the curation services needed at each stage of the research data lifecycle, analyze how to seamlessly embed into research workflows, and provide diversified curation services more effectively to maximize the scientific, economic, and social value of scientific data. This involves in-depth exploration of dynamic scientific data service mechanisms and models where libraries embed into research workflows.

- **Infrastructure and Key Technologies:** Comprehensively analyze important plans, progress, solutions, technical frameworks, and related technical methods in international Research Data Curation Infrastructure (RDCI). Particularly study the strategies and business models for literature and information institutions to engage in RDCI construction, providing useful references for building research data support and service environments that integrate into the research lifecycle. Conduct in-depth research on key technical methods for Data Curation, analyze relevant standards, technical strategies, and tool systems, and construct a research Data Curation technical framework for big data research environments.
- **Literacy Education Research:** Systematically analyze the roles and responsibilities of various stakeholders in research Data Curation and services (creators, experts, managers, data librarians), construct the knowledge and competency structures required for various roles to participate in research data management and services, and provide theoretical foundations and teaching material frameworks for relevant personnel training and continuing education.
- **Sustainable Development Research:** Conduct detailed research on cost and benefit models covering the research Data Curation lifecycle, analyze the needs, costs borne, and benefits obtainable by different stakeholders, and provide specific cost-benefit analyses to establish and maintain major investments for research Data Curation activities. Based on this, conduct research on sustainable economic models to form a self-sustaining research Data Curation ecosystem.

References

[1] DCC. What is digital curation? [EB/OL]. [2014-12-2]. http://www.dcc.ac.uk/digital-curation/what-digital-curation.
[2] NDSA. 2015 National Agenda for Digital Stewardship [EB/OL]. [2014-12-2]. http://www.digitalpreservation.gov:8081/ndsa/documents/2015NationalAgenda.pdf.
[3] UK Data Archive. UK Data Archive Strategic Plan, 2010-2015 [EB/OL]. [2014-12-2]. http://www.data-archive.ac.uk/media/196518/ukda-strategicplan20102015full.pdf.
[4] Research Data Management: Practical Strategies for Information Professionals [M]. Purdue University Press, 2014: 399-406.
[5] DCC. Five Steps to Developing a Research Data Management Policy [EB/OL]. [2014-12-2]. http://www.dcc.ac.uk/sites/default/files/documents/publications/DCC-FiveStepsToDevelopingAnRDMpolicy.pdf.
[6] DMP Planning for Big Science Projects [R/OL]. [2014-12-2]. http://arxiv.org/pdf/1208.3754v1.pdf.
[7] DCU. Key challenges and strategies [EB/OL]. [2014-12-2]. http://www.dcu.gr/index.php?p=dcu&lang=en&
[8] Whyte A, Wilson A. How to Appraise & Select Research Data for Curation [M]. Digital Curation Centre, 2010.
[9] DCC. Disciplinary Metadata [EB/OL]. [2014-12-2]. http://www.dcc.ac.uk/resources/metadata-standards.

[10] Layne R, Capel A, Cook N, et al. Long term preservation of scientific data: Lessons from jet and other domains [J]. Fusion Engineering and Design, 2012, 87(12): 2209-2212.
[11] Wu Zhenxin, Zhang Zhixiong, Guo Jiayi. Analysis of Long-term Preservation Technology Strategies for Digital Information Resources [J]. New Technology of Library and Information Service, 2006 (4): 8-13.
[12] Reilly S, Schallier W, Schrimpf S, et al. Report on integration of data and publications [J].

Author Introductions

Wu Zhenxin, female, born 1968, researcher and master' s supervisor at the National Science Library, Chinese Academy of Sciences.
Chen Yao, male, born 1991, master' s student at the National Science Library, Chinese Academy of Sciences and University of Chinese Academy of Sciences.
Li Wenyan, female, born 1989, master's student at the National Science Library, Chinese Academy of Sciences and University of Chinese Academy of Sciences.
Fu Honghu, female, born 1976, librarian at the National Science Library, Chinese Academy of Sciences.
Xu Liyuan, female, born 1986, librarian at the National Science Library, Chinese Academy of Sciences.

Address: Information System Department, National Science Library, Chinese Academy of Sciences, No. 33 Beisihuan West Road, Zhongguancun, Beijing
Postal Code: 100190
Phone: 15600602409
Email: chenyao@mail.las.ac.cn

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