KRDS Scientific Data Management Cost-Benefit Model Survey and Analysis Postprint

Authors: Wei Junchao, Li Sixue, Liu Panpan

Date: 2023-08-27T00:00:00+00:00

Abstract

[Purpose/Significance] To investigate and analyze the theory and practice of the KRDS model, providing guidance and reference for scientific data management initiatives in China. [Method/Process] Through literature and web-based research, this study summarizes the JISC-funded KRDS model and foreign university scientific data management practices based on this model, provides a detailed analysis of the application of the KRDS model in scientific data management cost analysis, and derives implications for scientific data management in China. [Results/Conclusion] The KRDS model is a universal, systematic framework. Conducting scientific data management cost analysis based on KRDS can comprehensively cover and predict various management stages. The KRDS model can provide a reference direction for scientific data management in China, standardize management processes, and detail management costs and benefits.

Full Text

Preamble

ChinaXiv Partner Journal, Vol. 62, No. 24, December 2018

Investigation and Analysis of the KRDS Scientific Data Management Cost-Benefit Model*

Wei Junchao, Li Sixue, Liu Panpan

Department of Library, Information and Archives, Shanghai University, Shanghai 200244

Abstract

[Purpose/Significance] This study investigates and analyzes the theory and practice of the KRDS model to provide guidance and reference for scientific

data management initiatives in China. [**Method/Process**] Through literature review and web-based research, this paper summarizes the JISC-funded KRDS model and its application in scientific data management practices at foreign universities, provides a detailed analysis of how the KRDS model is applied to scientific data management cost analysis, and derives implications for implementing scientific data management in China. [**Results/Conclusions**] The KRDS model is a universal, systematic framework. Using KRDS to analyze scientific data management costs enables comprehensive coverage and prediction of all management stages. The KRDS model can provide reference direction for China's scientific data management efforts, standardize management processes, and refine cost-benefit analysis.

Keywords: scientific data management; KRDS model; activity-based costing

Classification Number: G250

DOI: 10.13266/j.issn.0252-3116.2018.24.013

The fourth paradigm of scientific research, characterized by data-intensive computing, is emerging, making scientific data a crucial resource supporting scientific discovery. An increasing number of institutions and organizations have begun managing scientific data to effectively support knowledge discovery based on data. From an operational perspective, scientific data management comprises a series of activities centered on scientific data, including data organization, backup, archiving, sharing, publishing, and security management. These activities ensure data use and reuse, forming an important foundation for data-driven scientific discovery [1]. Research and analysis of the costs and potential benefits of these scientific data management activities can help us conduct scientific data management more effectively, ensure its sustainability, and provide reference for data management initiatives in China.

Foreign institutions have already conducted research and practice on scientific data management costs. A.S. Palaiologk et al. used activity-based costing to analyze the costs of scientific data management and long-term preservation at the Dutch Data Archiving and Networked Services (DANS) [2], focusing on dividing cost drivers into labor and non-labor costs through activity decomposition. The Consortium of European Social Science Data Archives (CESSDA) developed a cost-benefit advocacy toolkit for social science data management and long-term preservation [3], emphasizing social science data. Other mature digital resource preservation cost assessment models include the T-CMDP model proposed by the Dutch National Archives in 2005 [4], which primarily targets spreadsheets, emails, and similar data with cost accounting limited to present and future periods; NASA's improved NASA-CET model in 2008 [5], focusing on space and multidimensional data preservation; and the LIFE3 model proposed by the University of London and the British Library in 2010 [6], which is particularly suitable for electronic resources like books and newspapers, though its fixed cost values derived from empirical research on library resources lack generalizability.

Among these, the KRDS (Keeping Research Data Safe) model [7], funded by

the UK's Joint Information Systems Committee (JISC), supports cost-benefit accounting for multiple disciplinary data types, covers long time spans (past, present, and future), features relatively comprehensive cost driver classification, and addresses the gap in benefit analysis. Therefore, this paper selects the KRDS model as its research object.

Domestic research has primarily focused on cost-benefit analysis of long-term digital resource preservation. For instance, Su Xiaobo [8] and Xiao Ying [9] analyzed cost drivers for digital resource preservation; Yang Helin [10] and Zang Guoquan [11] examined cost-benefit relationships in library digital resource preservation; and Sun Chao [12], Xiao Qiuhui [13], and Li Haitao [14] compared and evaluated various foreign digital resource preservation cost models. However, these studies target broad digital resources including scientific literature, scientific data, audio, and video. Research specifically focusing on scientific data management through analysis of management activities for cost-benefit assessment remains scarce.

In the data-intensive research paradigm, scientific data management has become a routine activity for most research institutions. Analyzing its cost-effectiveness can help identify key activities and promote efficient management. This paper compares foreign scientific data management cost-benefit practices and models, selecting the JISC-funded KRDS evaluation model as its focus. The KRDS model supports multi-type scientific data management cost-benefit analysis, features comprehensive cost drivers, long time spans, strong applicability, and can provide clear guidance for domestic applications. Through literature and web-based research of KRDS project reports—including Keeping Research Data Safe (KRDS1) [15], Keeping Research Data Safe 2 (KRDS2) [16], the I2S2/KRDS Benefits Analysis Tools Project [17], the KRDS2 project website [18], and application reports from universities such as Cambridge [19], King's College London [15], and Southampton [20]—combined with relevant research papers, this paper provides a multi-dimensional analysis of the KRDS model from the perspectives of components, cost drivers, cost accounting frameworks, benefit analysis frameworks, and practical applications. The goal is to understand the KRDS scientific data management cost-benefit model, identify key activities constituting management costs, summarize cost drivers, and outline application procedures to provide reference for cost-benefit analysis of scientific data management in China.

1. Overview of the KRDS Model

1.1 Development of the KRDS Model

The KRDS model is a project outcome funded by JISC for evaluating the costs and benefits of scientific data management and preservation. It primarily helps institutions identify and determine management and preservation costs while raising awareness of associated benefits. Through investigation of long-term costs and benefits in Higher Education Institutions (HEIs), the project devel-

oped relevant theories, tools, and methods to guide the UK Higher Education Funding Council for England (HEFCE) and institutions, ensuring sustainable scientific data management.

The KRDS model has undergone three development phases: Phase 1 (KRDS1) completed in 2008 established the main cost analysis model and explored major factors constituting scientific data management costs; Phase 2 (KRDS2) completed in 2009 modified and optimized the model, identifying and analyzing benefit factors related to long-term data preservation; Phase 3 focused on promoting and applying the cost-benefit model and tools, transitioning research results into practice [7].

1.2 Components of the KRDS Model

The KRDS model consists of three main components: an activity model, cost drivers, and a cost accounting framework (see Figure 1 [Figure 1: see original paper]). The activity model identifies scientific data management activities with cost implications and arranges them into a hierarchical structure of activities and sub-activities. KRDS cost drivers are key variables (e.g., salary levels or inflation rates) that affect preservation activity costs, divided into two categories: economic adjustments and service adjustments. The KRDS cost accounting framework links costs (staff, equipment, etc.) with activity duration (1 year, 2 years, etc.) to form a comprehensive cost accounting model similar to the Transparent Approach to Costing (TRAC), which is widely used across 165 UK higher education institutions to calculate funding for teaching, research, and other major activities [21].

The KRDS activity model helps identify resource-consuming activities; cost drivers help institutions identify economic variables affecting activity costs and factors that must be considered in management and service processes, such as data formats, which ultimately require negotiation between data management institutions and data submitters; the cost accounting framework connects the activity model and cost drivers, facilitating learning, reference, and cost accounting. Thus, data repositories can use the KRDS model to identify resource-consuming management activities, determine cost drivers, and account for and analyze costs throughout the entire scientific data management process according to local conditions.

2. KRDS Activity Model and Cost Drivers

2.1 Activity Model

Scientific data management comprises activities conducted by researchers that consume resources and generate costs. Identifying these behaviors and activities is the first step in clarifying management costs. The KRDS activity model serves as a tool for identifying scientific data management activities, consisting of a hierarchical structure including two activity phases (pre-archival and archival)

plus supporting services and property management. The pre-archival phase involves all activities related to data creation and management before archiving. The archival phase includes activities for archiving research data into repository storage operated by universities or other institutions. Both phases relate to scientific data lifecycle costs. Supporting services comprise activities that assist pre-archival or archival activities, typically involving financial, IT, and other public service infrastructure. Property management involves administration of buildings and other infrastructure.

Based on research and analysis of LIFE, NASA-CET, OAIS, and TRAC models, the KRDS activity model evolved from KRDS1 to KRDS2 through continuous modification and expansion. The detailed breakdown is shown in Figure 2 [Figure 2: see original paper] [23].

The KRDS activity model divides scientific data preservation management activities into two phases—pre-archival and archival—plus supporting services and property management. The pre-archival phase primarily focuses on creating scientific data and converting it into archival format, considering factors such as data format and metadata that affect data generation and acquisition. This requires specifying relevant data preservation and sharing plans, generating descriptive information and user documentation, negotiating formats and logical structures with data creators, and providing archival training and support for data submitters and creators.

The archival phase is the main stage for long-term scientific data management. Initial stages involve developing data selection policies and negotiating submission agreements with data creators. Subsequently, data is transferred to repositories or preservation institutions, and unselected data is securely destroyed. Data management institutions provide appropriate storage capacity and equipment to receive data and convert formats to archival requirements. During preservation, management metadata, descriptive metadata, and user documentation are generated, along with mechanisms for updating archival content and semantic links explaining original data. Data integrity must be maintained during archival management, with timely additions, modifications, and deletions. When users request access and queries, result sets and reports must be generated in real-time, along with relevant training and support. Throughout data creation, management, and sharing, interaction with data consumers and producers must be maintained to track changing requirements and technologies, update preservation techniques and strategies, and develop new preservation technologies, tools, and standards requiring institutional collaboration.

Supporting services require administrative staff to provide daily support and control, establish and maintain data preservation standards and policies, provide software interfaces for data platforms, distributed applications, continuous utilities, daily office supplies, and staff training and development. Property management and service fees involve building rental, space management, and maintenance, which in the KRDS model are classified as cost factors separate from other public services based on function (e.g., laboratory/non-laboratory)

and calculated at variable rates.

2.2 Cost Drivers

To conduct cost accounting, preservation institutions must determine cost drivers after identifying management activities. These factors and their adjustments cause costs to emerge and change through resource and asset consumption. Cost drivers include various variables that affect scientific data management costs. The KRDS model lists these variables [15] to help institutions identify cost-influencing variables in management activities and account for resource consumption, costs, and changes through specific variable adjustments.

(1) General Factors. Before identifying specific cost variables for each activity phase, KRDS first identifies factors that holistically affect management costs, termed general factors. These are listed in Table 1 [15]:

Table 1. General Factors - Dataset level and preservation objectives - Controlling future costs (timing of actions) - Cost dependencies, linkages, and "ripple effects" - Sensitivity to workload and process scheduling - Development of preservation technologies and availability of commercial off-the-shelf (COTS) or mature open-source software/community standards and best practices ("first-mover innovation")

These help identify cost variables in subsequent specific activity phases. For example, most datasets in higher education institutions are used only by project teams, sometimes by very few external users. Preservation objectives ensure research data remains securely stored with sufficient descriptive information for data recovery. Timing of actions, such as generating descriptive metadata and user documentation during pre-archival rather than archival ingestion, is particularly important for cost savings. Cost dependencies exist in any scientific data preservation cost model. Human resources cannot quickly adapt to changes in total storage volume or short-term workload fluctuations, especially when preservation institutions have little control over when data arrives or processing speed. Technology development and COTS availability or mature open-source software applications significantly impact costs across lifecycle stages, typically suitable for external funding and collaboration.

(2) Economic and Service Adjustments. KRDS divides specific cost drivers into economic adjustments and service adjustments. Economic adjustment factors primarily refer to economic variables affecting scientific data management costs—when they change, resource and asset cost accounting changes accordingly, such as inflation rates and investment returns. Management institutions must negotiate these variables with data submitters for subsequent cost accounting.

Service adjustment factors refer to considerations that must be addressed when providing services or making demands at each activity phase, such as data vol-

ume, storage format, and user numbers. Scientific data management institutions must establish relevant standards and expectations before archival activities, enabling data submitters to follow standard regulations and procedures for cost accounting and ensuring subsequent preservation services. Table 2 lists these economic and service adjustment cost drivers [15]:

Table 2. Variables Affecting Scientific Data Management Costs - *Economic Adjustments:* Inflation/deflation; depreciation; financing and investment returns - *Service Adjustments:* - **Acquisition, processing, and ingestion:** Number of users; storage quantity, method, and frequency; number, complexity, and type of file formats; data volume; metadata, documentation, ethics, and intellectual property; level of processing, verification, and calibration; cost of de-accession - **Archival storage, preservation planning, data management:** Retention period; management and updates; number of versions and copies; storage media (capacity, cost); archival media monitoring - **Access:** Number of users and user communities; standard or custom interfaces; level of user support; access control; number and volume of accesses; access/distribution methods; service response time; processed products

(1) Economic Adjustments. Economic adjustment factors require negotiation among management institutions, data submitters, and funding agencies. Inflation rates typically apply to costs such as staff; deflation rates typically apply to equipment cost accounting; depreciation is usually calculated based on time elapsed or asset activity/usage levels; financing and investment returns include financing costs and minimum retained earnings. Their changes affect resource and asset cost accounting, making identification of economic variables particularly important.

(2) Service Adjustments. Across all activities, special attention must be paid to labor costs, which should include salaries, insurance, and pensions—representing the primary cost in scientific data management activities. In case studies, 70% or more of activity costs relate to labor, historically the main component of management costs. Due to the importance of staff costs, automation level becomes a significant variable affecting total costs, with impact depending on achievable economies of scale. To model inflation/deflation costs and adjustments, activity duration must be recorded. Additionally, different activity phases emphasize different cost types: the startup phase emphasizes fixed costs for installing system infrastructure, while subsequent operational phases emphasize variable costs for ongoing capacity.

During acquisition, processing, and ingestion, proprietary archival file formats significantly impact costs—non-proprietary formats simplify acquisition and migration procedures, reducing management risks and costs. For many new fields and applications where only proprietary formats are available, a crucial factor is supporting data export/import for these formats. The timing of descriptive metadata, ethics, and intellectual property activities is critical—conducting these during pre-archival substantially reduces costs; adjusting them during archival phases increases expenses and may significantly decrease data value.

Notably, most scientific data management costs occur during acquisition and ingestion rather than long-term archival storage and preservation.

In archival storage, preservation planning, and data management, longer retention periods require more protection measures to ensure data integrity and accessibility, resulting in higher total costs. Service levels also affect costs, such as throughput, error rates, and hardware replacement frequency. Similarly, media migration frequency and sample collection frequency involve personnel and equipment utilization, generating costs. During user access, costs are potentially the most variable phase. Depending on access level and method, costs are elastic—for example, web access versus staff-handled requests generate very different costs. Meeting user demands for high-speed access and specific data products also incurs expenses.

3. KRDS Cost Accounting Framework

The cost accounting framework is a simplified, universal framework for accounting scientific data management costs, linking elements from the KRDS activity model with cost drivers to facilitate cost accounting for scientific data management institutions. The framework covers pre-archival, archival, and support services, plus cost categories based on TRAC, with additional archival fees and outsourcing costs required by the KRDS model. Typically, the activity model helps identify resources or assets needed for management activities; economic adjustment cost drivers help clarify cost variations from economic variables or operational conditions that must be considered; service adjustment factors help identify and adjust specific resource variables involved in activities. Using this framework (see Table 3), institutions can clearly identify cost-related factors and variables for each activity, enabling smooth cost accounting.

Table 3. Cost Accounting Framework [15] - Duration (1 year, 2 years, etc.) - Outsourcing/archival fees - Cost categories based on TRAC classification: staff, equipment, travel, consumables, property costs, and indirect costs

In the complete TRAC classification, staff costs are divided into directly incurred or directly allocated costs. Note that the KRDS model stipulates that economic adjustment cost drivers must be incorporated into accounting considerations after negotiation between data management institutions and funding agencies. The cost accounting framework is a simplified model requiring more detailed division based on local conditions in practice.

4. KRDS Benefit Analysis Framework

Cost analysis alone is insufficient to assess the economic viability of specific scientific data management activities—it should be accompanied by benefit analysis, i.e., anticipating value generated from investment while maintaining long-term data existence and accessibility. Unfortunately, measuring benefits is challenging, especially when they are not easily quantifiable. Analyzing future economic

viability requires balancing costs and benefits. As a first step, developing important dimensions to clarify potential benefits is crucial.

The KRDS model proposes a benefit analysis framework [16] derived from case studies of UK data archives and Oxford University, analyzing potential benefits across three dimensions (see Table 4). While quantifying many benefits is difficult, using the KRDS benefit analysis framework enables data management institutions to better understand benefits from scientific data management.

Table 4. KRDS Benefit Analysis Framework - *Direct benefits:* For current researchers and students—no data loss in short-term circulation; short-term reuse of selected data; secure storage for data-intensive research; availability of journal paper data. For future researchers and students—value increases over time as collection volume and critical mass grow. - *Indirect benefits (avoided costs):* No data re-creation; no lost future research opportunities; lower future preservation costs; data repurposing for new users; repurposing methods. - *Wider benefits:* New research opportunities; academic exchange/data access; data repurposing and reuse; improved research efficiency; incentivizing new networks/collaborations; knowledge transfer to industry; technology base/skill composition; increased productivity/economic growth; research/research integrity verification.

This benefit framework aims to motivate management institutions, communities, and society to better understand benefits of long-term preservation and management, clarifying the benefit side of the cost-benefit equation. The three benefit categories—direct, indirect, and wider—help institutions think about benefits related to long-term preservation to better assess their relative weight against management costs. While many benefits are difficult to quantify, even qualitative expression can raise awareness among funders and decision-makers.

5. Applications of the KRDS Model

Many foreign institutions and universities have implemented scientific data management services, particularly UK universities that have conducted cost-benefit analyses based on the KRDS model, such as Cambridge, King's College London, and Southampton. Table 5 summarizes the steps these institutions follow when using the KRDS model for cost analysis.

Table 5. KRDS Model Application Cases - *Cost analysis cases:* King's College London (university data repository; national data center; arts and humanities data) — 1) Analyze general cost-impact issues using KRDS cost driver general factors list; 2) Allocate costs for each activity phase using KRDS cost accounting framework and TRAC elements; 3) Conduct specific cost analysis and accounting. - *Benefit analysis cases:* Southampton University (national data center; chemistry data) — 1) Divide management activities into three periods; 2) Conduct cost analysis for original and migrated data; 3) Perform benefit analysis using KRDS benefit analysis framework.

Based on research conditions, representativeness, and data availability from various universities, this paper selects three typical cases—King's College London, Cambridge, and Southampton—for detailed analysis of KRDS model applications.

5.1 King's College London KRDS Cost Model Application

This case study is based on 11 years of experience ingesting and managing complex scientific datasets and e-Research Center practices at King's College London (KCL). KCL's scientific data management is part of a larger project integrating a Virtual Research Environment (VRE) to support e-Research practices. The specific cost analysis steps are:

(1) Analyzing General Management Issues. Before cost accounting, KCL identified general cost-impact factors using the KRDS cost driver general factors list. Three key scope areas were determined: content and quality of KCL research data; integration of legacy systems with VRE; and VRE user requirements. Additional considerations included: the college's responsibility for scientific data beyond supporting its own research needs; whether data could be open access; responsibilities to the broader research community; how to meet additional costs related to open access; how to reasonably allocate responsibilities for scientific data from cross-institutional collaborative projects given increasing international cooperation in interdisciplinary research; how to meet KCL's scientific data management and preservation expenses; and whether TRAC could provide viable solutions.

(2) Applying the KRDS Cost Accounting Framework. After determining general management issues, KCL's scientific data management institution allocated costs using the KRDS cost accounting framework and TRAC elements. Cost elements from KRDS activity model phases were divided into three categories: directly incurred, directly allocated, and indirect. The institution provided audit trails for direct data management expenses, with directly allocated costs based on Full Economic Cost (FEC) [15] of management equipment. Following TRAC guidelines, all costs were directly allocated rather than indirect (see Table 6).

(3) Cost Analysis. KCL's cost analysis focused on labor and hardware costs. Key staff (costs) included an archivist (support activities, salary £45,000) and a half-time systems administrator (installing and managing software/hardware). Hardware costs were based on 2005 purchases including 15TB storage, a tape library, and a distributed server for end-user access [15]. Specific practices must also consider equipment storage capacity, organizational IT environment, and funder requirement changes, including dataset size, complexity, and type changes, with continuous equipment maintenance and updates throughout.

(4) Predicting Data Management Costs. Predictions are based on the concept of cost "peaks"—as archives expand, more equipment and storage capacity are needed, increasing costs over time. Similarly, as collections grow

annually, additional staff are required. Labor costs are determined based on collection volume, with full-time 8-level staff processing 30 collections annually (10 "simple" collections with standard formats/metadata for images/text and 20 complex collections with images, video, and interlinked documents). Staff spend 20% of their time on general tasks [15] such as reviewing and updating license agreements and standards.

5.2 Cambridge University KRDS Cost Model Application

Cambridge's case study focuses on DSpace@Cambridge and the Department of Chemistry's Unilever Centre for Molecular Science Informatics. DSpace@Cambridge began as a collaborative project funded by Cambridge-MIT from 2003-2006 to establish the DSpace software platform as Cambridge's institutional repository [19], accepting various digital content formats, primarily images and data from traditional research publications, now including chemistry informatics, archaeology, and anthropology field data. The main cost accounting steps are:

(1) Analyzing General Management Issues. Cambridge identified general cost-impact factors using the KRDS cost driver general factors list. Important considerations for determining medium- and long-term costs included: selection and/or evaluation processes; creating adequate metadata (resource-intensive); supported formats/versions; preservation plans for different formats; authenticity and availability requirements; sustainability; and hiring digital preservation experts to coordinate preservation activities.

(2) Applying the KRDS Activity Model. Cambridge identified specific personnel for each activity, focusing cost analysis on labor costs (see Table 7).

(3) Cost Analysis. Cambridge's cost data [19] covered: DSpace@Cambridge staff (three 8-level full-time staff; one 6-level full-time staff); digital and digital preservation experts (8-level full-time). Specific salary accounting methods were not specified. Hardware investment included approximately 150TB of mirrored storage costing £176,293.82.

5.3 Southampton University Cost-Benefit Model Application

This benefit analysis case study is based on longitudinal cost information from Southampton University's Chemistry Department, covering the National Crystallography Service from 1970-2009, during which experimental instruments, computing capabilities, and storage media (paper, digital video discs, etc.) changed completely. Management activities were divided into three periods based on technological transitions: 1970-1990, 1990-2000, and 2000-present [20].

(1) Cost Analysis. Southampton's analysis focused on preservation and migration costs. Experimental results are crystal structures—processed raw data into result data. Notably, generating a crystal structure currently costs £328, but recreating a structure from 1970-1990 would cost approximately sixty times

more. Managing raw data accounts for about 70% of total data management costs [20], making raw data management the critical component.

(2) Benefit Analysis. The case study emphasizes benefits aligned with the KRDS benefit analysis framework (see Table 8). Specific benefits are detailed for each activity in the KRDS activity model, determining benefit types, stakeholders, realization timelines, and impact weights to form a complete value chain.

6. Implications for Scientific Data Management Cost-Benefit Analysis in China

Although some Chinese universities have begun emphasizing scientific data management—such as Fudan University's establishment of a Humanities and Social Sciences Data Center in 2011 for centralized construction and unified management [24], and its 2013 cooperation with Harvard's Dataverse Network [25] to build a social sciences data platform, and Peking University's 2014 development of an open research data platform also based on Dataverse [26]—these platforms are currently free services without clear data submission protocols or detailed cost accounting models and regulations. Research and practice on scientific data management cost-benefit analysis remain incomplete. Based on the model research and case analysis above, China should consider the following when conducting scientific data management cost-benefit analysis:

6.1 Clarify Scientific Data Management Processes

According to current domestic data platform status, scientific data management should first clarify management processes, establish data management plans in advance, specify data submission formats and protocols, and identify factors holistically affecting costs. Data collection and access levels and preservation objectives must be determined, with particular attention to workload and process scheduling to ensure staff can adapt to storage volume changes and workload fluctuations. Research shows that collaborative establishment of scientific data management centers across multiple institutions, joint software development, technology tracking, management plan formulation, and shared protocols can significantly reduce management costs and improve data sharing efficiency. While Fudan and Peking Universities have surveyed and collaborated with multiple departments and research institutions, their scope remains limited to their own campuses with insufficient scale effects. Without clear data submission protocols and management processes, cost and benefit analysis becomes difficult.

6.2 Apply Mature Cost Analysis Models

Currently, China's scientific data centers have not developed standardized cost analysis models nor effectively adapted foreign mature cost-benefit analysis models. Various data management institutions face continuous problems during cost analysis, lacking clear activity phase divisions, activity identification, and

proper cost driver identification. Therefore, scientific data management institutions or data centers should combine their actual conditions to construct, adapt, and apply mature foreign cost analysis models like KRDS, identifying suitable activities including various management activities in pre-archival and archival phases, and determining cost drivers affecting each activity, involving variables such as data format, volume, and user numbers. Subsequently, linking activities with drivers forms a standardized localized cost analysis model, improving current cost analysis practices.

6.3 Conduct Cost Accounting Collaboratively

Currently, Chinese university scientific data platforms are in free service phases without clear charging protocols for data submission, download, and sharing. Platforms lack published cost accounting steps and methods, with financial accounting separated from platform services, easily causing cost omissions or duplicate accounting. Platform cost accounting should be conducted collaboratively by technical personnel and financial staff. The process must first clarify cost categories for each activity. Foreign case studies show that scientific data management institutions typically divide costs in two ways: according to TRAC methods (directly incurred or directly allocated) or by resource type (labor and asset capital), requiring institutions to determine required staff types and levels for each activity phase. Note that economic adjustment variables must be considered during cost accounting as they significantly impact overall costs.

6.4 Emphasize Scientific Data Management Talent Development and Automation

Labor costs account for the majority of scientific data management expenses, reaching up to 90% in some fields, making management talent a critical factor. While foreign countries have implemented many training programs, China has yet to offer library and information science courses on scientific data management, making talent cultivation urgent. From a data preservation activity perspective, labor costs for data ingestion and pre-ingestion management account for approximately 55% of total costs. However, highly automated operations can significantly reduce these labor costs. Scientific data management institutions should track technological developments in real-time, updating software and hardware to ensure efficient data ingestion and management, thereby reducing management costs.

References

[1] Managing and sharing data [EB/OL]. [2018-04-13]. http://www.data-archive.ac.uk/media/2894/managingsharing.pdf.

[2] PALAIOLOGK A S, ECONOMIDES A A, TJALSMA H D, et al. An activity-based costing model for long-term preservation and dissemination of digital

research data: the case of DANS [J]. International journal on digital libraries, 2012, 12(4): 195-214.

[3] CONSORTIUM OF EUROPEAN SOCIAL SCIENCE DATA ARCHIVES. CESSDA SaW [EB/OL]. [2018-04-13]. https://www.cessda.eu/Projects/All-projects/CESSDA-SaW.

[4] KEJSER U B, NIELSEN A B, THIRIFAYS A. Cost model for digital preservation: cost of digital migration [J]. The international journal of digital curation, 2011, 6(1): 225-267.

[5] NASA. CET V2 p4 tech description 080915 [EB/OL]. [2018-04-14]. http://opensource.gsfc.nasa.gov/projects/CET/index.

[6] LIFE. An introduction to the third phase of the LIFE project [EB/OL]. [2018-04-14]. http://www.life.ac.uk/3/docs/life3_{report}.pdf.

[7] BEAGRIE C. Keeping research data safe: cost-benefit studies, tools, and methodologies focusing on long-lived data [EB/OL]. [2018-04-15]. https://www.beagrie.com/krds.php.

[8] Su Xiaobo, Chang E. Analysis of cost influencing factors in long-term preservation of digital resources [J]. Library and Information, 2011(1): 20-24, 44.

[9] Xiao Ying. Research on digital information preservation costs [J]. Library Journal, 2004(11): 14-17.

[10] Yang Helin. Data curation: a new exploration in American university libraries [J]. Journal of Academic Libraries, 2011(2): 18-21, 41.

[11] Zang Guoquan, Li Sisi. Research on investment decision-making for digital preservation projects—based on analysis of uncertainty in project investment costs [J]. Library Theory and Practice, 2014(8): 36-41.

[12] Sun Chao, Wu Zhenxin. Analysis and evaluation of foreign digital resources long-term preservation maturity models [J]. Library and Information Service, 2017, 61(1): 32-39.

[13] Xiao Qiuhui, Xu Xiaotong, Bian Junyan. Comparison and evaluation of digital preservation cost models LIFE and CDL-TCP [J]. Library and Information Service, 2017, 61(18): 12-18.

[14] Li Haitao. A digital preservation cost model for public libraries from an accounting cost perspective [J]. Library Tribune, 2017, 37(2): 75-82.

[15] Keeping research data safe 1 [EB/OL]. [2018-04-23]. https://www.webarchive.org.uk/wayback/archive/201

[16] Keeping research data safe 2 [EB/OL]. [2018-04-28]. https://www.webarchive.org.uk/wayback/archive/201

[17] I2S2 idealised scientific research activity lifecycle model [EB/OL]. [2018-04-30]. https://www.beagrie.com/krds-i2s2/.

[18] Keeping research data safe 2: a JISC-funded project [EB/OL]. [2018-05-02]. https://www.beagrie.com/jisc/.

[19] UNIVERSITY OF CAMBRIDGE. Research data management [EB/OL]. [2018-05-01]. http://www.data.cam.ac.uk/repository.

[20] Southampton data survey: our experience and lessons learned [EB/OL]. [2018-05-01]. http://www.disc-uk.org/docs/SouthamptonDAF.pdf.

[21] HIGHER EDUCATION FUNDING COUNCIL FOR ENGLAND. Transparent approach to costing: an overview of TRAC [EB/OL]. [2018-04-15]. http://www.jcpsg.ac.uk/guidance/downloads/Overview.pdf.

[22] KRDS user guide [EB/OL]. [2018-04-17]. https://www.beagrie.com/static/resource/KeepingResearchData

[23] Detailed version of KRDS2 activity model [EB/OL]. [2018-04-20]. https://www.beagrie.com/KRDS2_{{{Activity}}*{{Model}}}*{detailed}.doc.

[24] Liu Xia, Rao Yan. Preliminary exploration of scientific data management and services in university libraries—a case study of Wuhan University Library [J]. Library and Information Service, 2013, 57(6): 33-37.

[25] 复旦大学社会科学数据平台 [EB/OL]. [2018-05-05]. https://dvn.fudan.edu.cn/home/index.jsp.

[26] 北京大学开放研究数据平台 [EB/OL]. [2018-05-05]. http://opendata.pku.edu.cn/.

[27] UNIVERSITY OF ESSEX. UK data archive [EB/OL]. [2018-04-14]. http://www.data-archive.ac.uk/.

[28] KRDS benefits framework worksheet [EB/OL]. [2018-05-02]. https://www.beagrie.com/KRDS_{{{Benefits

[29] Benefits impact worksheet [EB/OL]. [2018-05-03]. https://www.beagrie.com/Benefits%20Impact%20Works

[30] Value chain and benefits impact worksheet [EB/OL]. [2018-05-03]. https://www.beagrie.com/KRDS_{ValueChainImpactTool}%20Worksheet_{{{v1excel}}}_{{July2011}}}.xls.

[31] Wei Junchao, Yu Haiyan. Investigation and analysis of institutional data asset auditing from the perspective of the "Data Asset Framework (DAF)" [J]. Library and Information Service, 2016, 60(8): 59-67, 92.

[32] WHEATLEY P, AYRIS P, DAVIES R, et al. LIFE: costing the digital preservation lifecycle [J]. New Technology of Library and Information Service, 2008(1): 69-74.

[33] SMITH M K, BARTON M, BASS M, et al. DSpace: an open source dynamic digital repository [J]. D-lib magazine, 2003, 9(1): 10-17.

Author Contributions

Wei Junchao: Responsible for overall structure and final editing; Li Sixue: Responsible for research and paper writing; Liu Panpan: Content revision.

Investigation and Analysis on KRDS Model of Cost and Benefits for Research Data Management

Wei Junchao, Li Sixue, Liu Panpan

Department of Library, Information and Archives, Shanghai University, Shanghai 200244

Abstract: [Purpose/significance] With the emergence of the fourth paradigm of data-intensive scientific discovery, research data management has become an important prerequisite for such discovery. By surveying the Keeping Research Data Safe (KRDS) model, this paper aims to provide reference for research data management in China. [Method/process] This paper analyzed the JISC-funded Keeping Research Data Safe (KRDS) model and KRDS-based practice of overseas universities through literature research and web-based survey, then analyzed the characteristics, steps, experiences and lessons learned of research data management cost analysis based on KRDS. Finally, it proposed specific measures for research data management cost analysis in China. [Result/conclusion] The KRDS model is a universal and systematic framework. It can fully involve and predict the key costs of each activity. KRDS could provide direction, standardize the management process, and refine management costs and benefits in China.

Keywords: research data management; KRDS model; activity-based costing

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