Academic Libraries in Research Data Management Service: Perceptions and Practices

Qi Zhou

Library of Jinan University, Guangzhou, China
Email: tlibcf_zhou@jnu.edu.cn

Abstract

Purpose: Explore the process of conducting research data management services and provide effective recommendations for academic libraries to conduct data management services. Design/methodology/approach: On the premise of summarizing and analyzing the connotation of research data management, this paper discusses the characteristics of university research data management. By sorting out many cases of research data management, we summed up several major elements of research data management service practice in universities: policy formulation, infrastructure, service content, funding model, and staffing. Findings/value: Systematically sort out the basic elements and development status of university research data management service practice and provide reference for the university to carry out research data management service.

Subject Areas

Information Science

Keywords

Research Data Management, Research Data, Academic Libraries, Practice and Exploration

1. Introduction

In the context of E-Science, as scientific research has become more collaborative, data-intensive, and computational, the demand for research data management and services among academic researchers has been increasing [1]. The Association of College and Research Libraries Research Planning and Review Commit-

tee released the top of the ten mainstream research trends at the end of 2016 as "research data service" [2]. In recent years, the value of research data has been increasingly valued. The scientific community has placed greater emphasis on promoting the open sharing of data. C. P. Clement [3] and others pointed out: "More and more people realize that the basic materials that constitute knowledge should receive the same attention as research articles that synthesize and interpret these original materials". This emphasis is also reflected in many other areas. National Institutes of Health, the National Science Foundation, UK AHRC (Arts and Humanities Research Council), MRC (Medical Research Council), and many other research funding agencies have issued explicit open data and data sharing policies. Private institutions such as Gates, Ford, and the Sloan Foundation have also followed suit and both have proposed data opening and data sharing requirements. PLOS ONE, Nature and other academic journals have also begun to use data sharing as one of the conditions for publication. Driven by this environment, many universities began to launch research data management services to better meet the requirements of data opening. More than half of the academic libraries in the United States have conducted data support related services. With the help of institutions such as JISC (Joint Information Systems Committee of UK) and DCC (the Digital Curation Centre), more than 30 universities in the United Kingdom carry out data management related projects. With the promotion of ADNS (Australian National Data Service), Australia also has more than 30 universities that conduct research data management services. In the field of research, China has conducted many researches on foreign research data management service models in various fields, mainly including analysis of research data management services based on network research or more in-depth concrete service case analysis, and data management policy analysis, research data management system platform analysis, and data literacy education research.

Research data management has many benefits for scientific research, such as facilitating research and verification, promoting scientific communication and understanding [4], and facilitating long-term data storage. What really promotes the development of research data management is the transformation of the scientific data paradigm and the booming development of the open access movement, and universities have thus become the main positions for the development of data management services. Researching data management services is a systematic and long-term project with many factors.

2. Research Data Management Concepts and Features

Data management first appeared in the computer field. With the formation of the e-Science environment and the large-scale use of computers, data management has gradually penetrated into the field of graphic arts. Data Management, Data Curation, and Digital Curation are currently used more often. The "data" in data management refers to "research data". In foreign literature, research data

is mostly expressed in Research Data and Scientific Data, but there is no difference between them. They refer to the original, basic data in scientific research activities. The research data defined by Johns Hopkins University is a record of the results to be used to reconstruct and evaluate reports or otherwise published, such as laboratory notes, original experimental results, and instrument output values. The definition of research data given by Yale University is that information collected, observed, or created for analysis purposes to produce the original study. This includes observation variables such as survey data on rainfall, wind speed, and water quality, seismic simulation data, laboratory data, and derived and compiled data for text mining or testing algorithms. Research data can take any digital file format such as video, text, photos, numbers, etc.

Regarding the definition of research data management, A. Whyte and J. Tedds considered it to be a "data organization process throughout the entire research life cycle" [5]. A. M. Cox and S. Pinfield believe that research data management includes "a series of activities and processes in the research life cycle, including data construction and generation, storage, security, preservation, sharing, and reuse, as well as technical, moral, legal, and regulatory issues" [6]. We believe that research data management refers to embedding in the scientific research environment, collecting, sorting, excavating, classifying, and storing research data, and then sharing processed high-value data with scientific researchers to provide personalized information throughout the entire data life cycle and consulting services. Research data management is not a pure data management behavior, but a complete data processing process management and monitoring from data description, data storage, data long-term preservation to data publishing, sharing, reuse.

University research data management has the following characteristics.

2.1. Embedded Education

The diversity of university research data management is reflected in many aspects. First, the diversity of data formats. M. Henty [7], A.L. Whitmire *et al.* [8] conducted a survey of university science management practices and found that university-generated data cover almost all types, including spreadsheets, database data, text, SPSS/XML, and experimental data, digital images, document reports, etc., and different data formats, their processing methods, storage scale are very different. Second, the breadth of the discipline. Compared with the monographs of other research institutions, universities, as comprehensive research institutions, involve many disciplines and their respective fields have different characteristics. Research data management services need to be targeted. Third, the richness of research groups. The research users in the university include teachers, undergraduates, graduate students, etc. The career development stages of different groups, research data management awareness, and skill levels are different, and the demand for research data management is also different.

2.2. Cross-Sectoral

Cross-sectoral research data management encompasses a range of activities and services involving many areas of expertise and work processes, and no single department can independently conduct research on data management services. Compared to other types of institutions, universities set up many large departments, each of which can play its own advantages and integrate resources to form a unified and complete service system. For example, libraries are good at consulting and training services, and technical departments are good at technical services.

2.3. Low Value Density

As a primary scientific research institution that carries out scientific activities, the university has teachers' teaching and research tasks, students have learning tasks, and scientific research activities have become supporting tasks, leading to uneven levels of scientific research, low attention to data, and disorganized and irregular data management. The overall data value density is low. Therefore, to carry out university research data management services, in addition to enhancing the research data management awareness and level of university scientific research users, we must also introduce an effective research data value evaluation system to eliminate a large number of low-level research data and identify truly valuable ones. It makes sense to reduce the load on the management system and reduce the budget.

3. Composition of University Research Data Management Service System

Driven by a series of policy and research funding agencies, many universities have begun to conduct comprehensive research data management services or put them on the formal agenda. Among these universities, some use their own practical experience to conduct service exploration alone; others establish pilot projects with funding from related agencies. At present, countries that have developed better are the United States, Britain, and Australia. There are also some universities in China such as Peking University, Fudan University, Wuhan University and Xiamen University has made initial attempts.

3.1. Service Policy Formulation

Formulating research data management policies is the first step for academic institutions such as universities to conduct research data management, to standardize research data management services and to define the direction of future development. At present, most of the universities involved in the study of data management are at the stage of policy formulation. A survey conducted in mid-2012 showed that 1/3 of universities in Australia participated in the formulation of research data management policies, with the highest degree of activity in the world. It also ranks among the top in the world in researching the history

of data management policy and the maturity of content, followed by the United Kingdom (17.3%) and Ireland (12.5%).

Policy formulation firstly refers to the policies of internationally renowned data service organizations such as DCC and ANDC. It thoroughly understands the needs of various stakeholders for research data management and fully understands the research data management requirements of various research funding agencies and then issues policies. The content of the policy mostly covers the definition and objectives of research data management, data access, acquisition and preservation, data management plans (DMP), data sharing, data usage specifications, and data ethics. In general, Australia's university's data management policies are the most comprehensive and mature, with the most revisions. British universities' data policies are the shortest, mostly clarifying the direction and objectives of policy formulation, and lacking detailed explanations and specificities; American University's data policy places great emphasis on data access, data storage, and other main services.

3.2. Service Content

3.2.1. Research Data Management Plan

Before each scientific research project begins, a data management plan should be prepared to reduce risks and rationally allocate resources. The research data management plan originated in the United States. In 2010, the National Science Foundation (NSF) began to require all fund applicants to submit data management plan (DMP) for the project in the application materials in order to achieve the sharing and dissemination of fund research results [9]. The introduction of this policy was a big challenge for most of the researchers at the time, so many people began to consult librarians and seek help. The American library community follows the trend, deepens research in this area, actively conducts data management plan guidance, compiles research data management guidelines and templates, and solves practical difficulties for researchers. At present, in the United States, in addition to the National Science Foundation, many fund research institutions such as the National Institutes of Health (NIH) and the National Aeronautics and Space Administration (NASA) have made clear requirements for research data management, especially DMP. The American university library also takes DMP guidance as the focus of data management services. The services provided include DMP inventory templates, naming rules, data types and format descriptions, data ownership, data ethics, copyright license, data storage and sharing requirements, etc. British and American libraries have also developed DMP's online production tools, such as DMPonline and DMPTooI, and there are so many users.

3.2.2. Metadata Service

Metadata services are mainly to help researchers establish metadata that meets a certain standard, enhance operability among data sets, increase the probability of data being discovered, and provide more comprehensive and detailed data

descriptions. In the establishment of metadata standards, it can synthesize metadata standards such as EML, ISO 19115, Dryad, TEODOOR, PANCAEA, etc. for various subject areas, as well as popular metadata standards such as DataCite, DC to form multi-level, personalized metadata. For example, the CRC/TR32 project summarized four sets of internationally popular metadata standards and formed a three-tiered elemental system based on general information, project-specific information, and data type-specific information, both retaining its uniqueness and maintain interoperability with other metadata standards. In addition to manual information processing, there are also some metadata services that provide automatic information extraction. For example, University of Bristol uses content analysis tool Apache Tika to extract content types, sizes, checksums, and some embedded information [10]. There are two main options for metadata services. One is to establish specialized training courses for specific metadata standards, although the process is cumbersome but flexible; the second is through the system gradually guides or metadata tools (such as Morpho) to automatically create metadata records for users that apply to specific projects or specific disciplines.

3.2.3. Research Data Storage

The long-term preservation of valuable research data is an essential part of scientific research activities and can be used to accumulate basic data for future research. Research data storage service refers to assisting scientific research personnel in submitting, storing, backing up, and updating research data. It is a core function of researching data management services. Libraries have rich experience in resource conservation. They can expand the self-built institutional knowledge base or collaborate with relevant institutions and scientific research institutes in schools to build a storage platform to provide advanced technical service platforms and a good hardware environment for research data storage.

Many university storage platforms are doing well, such as the DSpace of MIT Library, DataSpace of Princeton University, DataStar of Cornell University, E-Data of Purdue University, HMDC of Harvard University and Massachusetts. The platforms can provide scientific research workers with storage and sharing services for research data, and some organizations also call it data curation services. Oxford University has established a data management two-tier storage system that not only satisfies the needs of researchers for local data management, but also provides convenience for scientific institutions to manage and maintain data. The data storage facilities of Australian universities are generally constructed and managed by the university. Most libraries are only responsible for providing storage guidance services for users, mainly instructing scientific research workers to apply for storage space, selecting storage methods suitable for their own needs, and how to conduct data version control and ensure data security. In 2014, Fudan University and the Dataverse Network of Harvard University established the Fudan University Social Science Data Platform. This is the first university social science data platform in China and can provide research

data storage, publication and exchange for universities, research institutes and government, sharing and online analysis and other functions. The "University Scientific Data Sharing Platform", one of the CALIS Phase III projects undertaken by Wuhan University Library, is also in the trial run phase. It is based on the open source software DSpace and has basically established data submission, organization, preservation, sharing, use and other specifications.

In addition to building and promoting an internal storage system that meets the needs of users, it is also necessary to properly understand external warehousing suitable for scientific research personnel. Current establishments such as re3data.org, Open DOAR, ROAR, OAD, and other data warehousing registration systems provide scientific research personnel with an understanding of the outside of the organization. Data Warehousing provides convenience. Research data managers should actively understand these data warehouses and make appropriate recommendations based on the subject areas, confidentiality, content type, and data size of the research data.

Research data storage services are divided into transitional storage and long-term storage services according to storage. Managers need to set the retention period for the data of scientific research personnel by referring to the length of the research project cycle and the value of data reuse. This will not only ease the conflict between the growth of data resources and space, but also improve the overall data quality.

3.2.4. Research Data Mining and Sharing

Research data is stored for its mining and sharing services. Analyze, collate and evaluate saved data, tap the correlation between these data and other related information, carry out secondary development, realize data value-added, and allow more scientific researchers to use these valuable values through certain sharing channels. That is the ultimate goal of research data management.

Data sharing services include data publishing, discovery, retrieval, download, and other services. Data sharing involves the rights and interests of researchers, and creators can set their own data sharing permissions, different data and different groups of people, such as users outside the organization and outside the organization, users inside and outside the project, confidential data and non-confidential data, will have very different sharing rights. Before the data is shared, a series of issues such as metadata compilation, data navigation, and copyright protection matching with these data also require the help of the library.

Research data managers need to establish a stable publishing process for scientific users to help them obtain DOI numbers for specification reference of data. Harvard University analyzes the stored research data, generates SPSS and STATA analysis data tables, and then shares it with scientific researchers through the search platform, which can be directly used as reference data. The German National Library of Science and Technology (TIB) allocates a digital object identifier for each stored data, enabling effective linking of research data

and library documents, which greatly facilitates information acquisition by scientific researchers. In 2014, the National Natural Science Foundation of China and the Chinese Academy of Sciences began requesting fund-funded research results to be stored in the knowledge base and to achieve open access. The Cross-border Integrated Search Service provided by the Chinese Academy of Sciences' Documentation and Information Center can now be used for non-documentary data resources search.

Data discovery services refer to increasing the visibility of data to help researchers discover and locate the required data and promote the sharing and utilization of data. Data discovery services are generally collected by specific organizations for data storage of various types of data warehousing, such as RDA, RDRDS, DCI, DataCite, etc. They are not data storage sites themselves, but metadata that includes data in major data storage. After the user retrieves the metadata and obtains the records of related data, he can further mine and reference the data. Therefore, to improve the utilization rate of internal data, the management department must cooperate with these organizations to increase the visibility of data.

3.2.5. Research Data Reference

Reference consulting is a kind of personalized deep service and it is also one of the core work contents of subject librarians. Librarians can be embedded in the work of scientific research teams in order to develop reference services for research data. Based on a large number of library data sources, they can actively understand the scientific research workers by focusing on data classification, uploading, storage, maintenance, and sharing. The data management needs, to answer the various problems they encountered in the data management process, provide them with targeted services throughout the life cycle of the research data.

Some universities in Australia have set up research data librarians. They provide data reference for scientific research workers in the form of FAQs, E-mails, and online consultations on the basis of academic services. Librarians from the University of Maryland Library participated in the development and design of school research projects. Based on the data management consulting issues they had completed, the project team provided research recommendations for sustainable development and reduced the duplication of research work. There are also universities that provide data management appointment consulting services for researchers. Researchers can make appointments online or by telephone. After a meeting, the librarians will analyze and discuss with them about their data management issues, and further propose research or management suggestions.

3.2.6. Research Data Literacy Cultivation

The formation of an intensive data environment has caused the scientific data generated by universities to increase in terms of "quantity," "category," and "speed." Researchers face a range of data management issues, such as data man-

agement planning, data citations, data publishing and ethical use of data, etc. Scientific data literacy, as a key concept of data management, has become one of the necessary capabilities for academic researchers to research and communicate. Scientific data literacy is similar to information literacy, including data awareness, data management knowledge, and data management skills. At the same time, scientific data literacy is cyclical, emphasizing the activities of collecting, processing, evaluating, managing, and utilizing scientific data, and paying attention to the various skills required to manage data in the basic flow of scientific research. In addition, scientific data literacy emphasizes the ability to analyze data, present data, and use data management tools.

For academic libraries, cultivating the data literacy of teachers and students is an urgent and important task, as well as new opportunities and challenges. Some academic libraries in Europe and the United States have already carried out appropriate literacy education activities to increase scientific personnel's data awareness, data collection and analysis skills, and promote scientific data management and sharing. The library should design data literacy education courses based on the needs of users, proceed from internal teaching and research work flow, take libraries as the leading factor, adopt multi-party cooperation, and develop a management mechanism of "inter-institutional" collaborative development. Libraries should also actively develop specialized, personalized and embedded data literacy education.

3.3. Technology Infrastructure

Technical infrastructure refers to the storage system used to manage and store research data, including technology platforms and data warehousing. Infrastructure is the precondition for effective management of research data, especially for organizations that are academic-based and highly data-intensive. Without a platform dedicated to storing data, effective data organization, preservation, and security cannot be achieved. From the point of view of the type of infrastructure, it can be divided into general-purpose and customized data management platforms. The general-purpose management platform has the advantages of cost and efficiency, but to deeply meet the needs of different disciplines and different fields, the customized data management platform has more development potential. The transformation of the general-purpose platform for customization usually adopts a gradual and open-source development method. At present, as the demand deepens, the research data management systems of some universities have integrated various functional collaboration platforms or personalized discipline management platforms, forming a hierarchical or collaborative system structure, such as the United States Data Management Alliance DATA-PASS platform group.

The construction of infrastructure requires a lot of capital investment. At present, there are three kinds of programs for building research data management infrastructure:

1) Conduct research data management services using the infrastructure, technologies, and management methods of the institutional repository. This is a reasonable choice for universities with smaller data sizes to avoid duplication of work and save a lot of economic costs.

2) Establish a special data repository (DR). The establishment of a special DR can be done through independent design, independent research and development, or using existing mature software such as Alfresco, CKAN, DSpace, EPrints. These software form a micro-service architecture through a modular approach that allows users to develop and utilize new features to meet specific needs.

3) Utilize cloud storage technologies such as Google Drive, Microsoft One-Drive, ownCloud, etc. The advantages of cloud storage systems are that they are secure, replicable, and cost-effective, and that they can be protected from physical damage and incidents. They can also be downloaded anywhere, facilitating remote operations and cooperation. The disadvantage is that the bottleneck of uploading or downloading may occur due to the speed of the network, and the demand for scientific research tasks with large data volumes cannot be met, while the independent storage system in the organization has the relative advantage in network speed.

3.4. Funding Mode

Capital investment is also a prerequisite for the development of research data management services. The focus of discussions on research data management has also changed from how to develop services to how to maintain services. M. Lewis has pointed out that establishing a sustainable business model to maintain the operation of research data management is a key circle to meet the scientific research of the 21st century.

At present, research data management projects mainly have the following major financial sources. In the initial stage of the project, some institutions such as Jisc, DCC, and ANDS that are dedicated to promoting research data management will provide financial support. Most university research data management projects obtain funding sources in this way. However, once the project is over, the capital chain is broken and the maintenance of services has become a difficult problem for many universities. At this time, in addition to obtaining limited funding from the university's own finance department, there are also some universities that use the method of charging service fees to researchers to reduce the economic burden. Researchers usually use project funding from research funding agencies to pay this fee. There are various ways to collect fees, such as annual payment, one-time payment of all possible fees, etc.

3.5. Service Staff Configuration

The university subject librarian facing research data management should be a comprehensive high-quality talent. In addition to the basic skills of general subject librarians, such as basic knowledge of discipline services, enthusiastic service

attitude, and good communication and coordination skills, etc. They must also need some qualities, just as high data literacy and data management service skills, a sense of advanced data management services, the ability to integrate data and collection resources, data analysis and data mining capabilities, skilled application of data platform capabilities, and data management consulting and guidance ability, in addition to a certain degree of scientific research. The improvement of these capabilities requires the joint efforts of universities and subject librarians. The university should adopt a combination of active introduction and comprehensive training, actively explore qualification certification of librarians, and strengthen management mechanisms and incentive mechanisms to innovate through multiple trainings, exchanges, and more. The way of learning provides the conditions for librarians' growth and success.

Research data management is about to open up a new field of academic library knowledge services. How effective the service is, the key lies in the ability level of subject librarians. In order to truly participate in the research data management work and realize the important transformation from document service to knowledge service, academic libraries must pay attention to and strengthen the construction of librarians' research-oriented data management service capabilities.

4. Conclusions

The increasingly prominent value of research data, the transformation of scientific research paradigms and the development of open access movements, and the increasing demand for open data and data management, provide opportunities for libraries and other traditional institutions to innovate services and integrate into the new scientific research revolution. Major universities in the world have participated in this trend.

However, as a whole, the university research data management service is still in its infancy. It needs to go through such links as policy formulation, infrastructure construction, service content design, service team formation, service user mining, and service fund raising. These links together constitute the practice of research data management. Universities and their libraries need to have a deep understanding of the operational processes, best practices, and influencing factors of each link, and in combination with their own development, establish a continuous and effective research data management service model to promote the further development of the open access movement.

Project

References

[1]   Tenopir, C., Birch, B. and Allard, S. (2012) Academic Libraries and Research Data

Services. Current Practices and Plans for the Future. Association of College and Research Libraries, Chicago, IL.

[2] ACRL Research Planning and Review Committee. (2016) 2016 Top Trends in Academic Libraries. *College & Research Libraries*, **77**, 274-281.

[3] Clement, G. and Schiff, L. (2015) Mapping the Landscape of Research Data: How JLSC Contributors View this Rapidly Emerging Terrain. *Journal of Librarianship and Scholarly Communication*, **3**, No. 2.

[4] Sayogo, D. and Pardo, T. (2013) Exploring the Determinants of Scientific Data Sharing: Understanding the Motivation to Publish Research Data. *Government Information Quarterly*, **30**, S19-S31. https://doi.org/10.1016/j.giq.2012.06.011

[5] Whyte, A. and Tedds, J. (2016) Making the Case for Research Data Management. http://www.dcc.ac.uk/webfm_send/487

[6] Cox, A.M. and Pinfield, S. (2014) Research Data Management and Libraries: Current Activities and Future Priorities. *Journal of Librarianship and Information Science*, **46**, 299-316. https://doi.org/10.1177/0961000613492542

[7] Henty, M. (2008) Dreaming of Data: the Library's Role in Supporting E-Research and Data Management. http://apsr.anu.edu.au/presentations/henty_alia_08.pdf

[8] Whitmire, A.L., Boock, M. and Sutton, S.C. (2015) Variability in Academic Research Data Management Practices—Implications for Data Services Development from a Faculty Survey. *Program: Electronic Library and Information Systems,* **49**, 382-407. https://doi.org/10.1108/PROG-02-2015-0017

[9] NSF. Dissemination and Sharing of Research Results. http://www.nsf.gov/bfa/dias/policy/dmp.jsp

[10] Hiom, D., Fripp, D., Gray, S., Snow, K. and Steer, D. (2015) Research Data Management at the University of Bristol: Charting a Course from Project to Service. *Program: Electronic Library and Information Systems*, **49**, 475-493. https://doi.org/10.1108/PROG-02-2015-0019