ATLANTIS PRESS

Advances in Social Science, Education and Humanities Research (ASSEHR), volume 156

2nd International Seminar on Education Innovation and Economic Management (SEIEM 2017)

Research on the Standardized Process of Research Data Management in CHINA

Wenjing Chu*
Library
Jiangxi Normal University
Nanchang, Peoples R China, 330022
chuwenjing0202@126.com

Zhaohui Chu
Department of Mathematics and Physics, Hefei University
Hefei, Peoples R China, 230601
czh_2004@hfuu.edu.cn

Hongsheng Pang
Library, Shenzhen University
Shenzhen, Peoples R China, 518060
5849583@qq.com

Shuning Li
Library, Beijing Normal University
Beijing, Peoples R China
35748202@qq.com

***Abstract*—In view of intersected management and lack of macro management and coordination in research data management, this paper establishes a standardized and orderly management process for scientific data to make better use of the value of scientific data. This paper develops a corresponding standard management process based on the scientific life cycle model through a large number of foreign scientific data management practices, and combined with China's national conditions. The design of a standardized management process will ensure scientific data integrity, security, availability, long-term preservation and reuse. As a result, which saves the time of researcher, improves research efficiency, and accelerates the scientific process.**

Keywords—research data; research data management(RDM); process；library

Research data is the physical records and documents collected, studied and created. Its purpose is to support the results of the study. It is an important and expensive achievement of the academic study of all subjects, and also the most basic, most active science and technology resource in the information age with the widest influence. If the research institutes do not conserve the data in time, the research data may be lost forever and the research value will be greatly discounted. The current research ecosystem is very complex, focusing on research data management. Research Data Management (RDM) is an organization of data [1]. It disseminates the data throughout the research life cycle and saves the valuable data. From the 4th (2017) research data conference on August 3rd in CHINA, it is learned that the ministry of science and technology will gradually promote the research data management from four aspects in the future: policies and regulations formulation, research data center construction, research data exchange, data opening and sharing [2]. Research data management theory and practice start relatively late in CHINA. There is segmentation of trap and block in the management, lack of macro management and coordination of national level, and the relevant policy also hasn't formulated [3]. Therefore, it urgently needs to establish a normative and orderly research data management process to maximize the value of research data.

Humanities and Social Sciences Project of Jiangxi Higher Education Institutions(TQ162002); Social Science "13th Five-Year"(2017) Project of Jiangxi Province (17TQ02); Humanities and Social Sciences Project of Anhui Higher Education Institutions(SK2016A0766); Teaching Research Project of Anhui Higher Education Institutions((2015jyxm315)

I. RESEARCH DATA MANAGEMENT PRINCIPLES

A. Standardization

Research data management is a very complex and huge project which needs to be processed strictly in accordance with standardized procedures, in order to ensure data integrity and availability.

B. Fairness

Fairness needs to be reflected from 4 aspects of the research data, namely easy to find, easy to obtain, easy to operate and easy to reuse [4]: (1) Data can be found: the data are found mainly through standardized description, firstly, each (meta-) data will be assigned a unique lasting identifier to describe in detail, the description of metadata should be very clear and a searchable resource index needs to be established to search for (meta-) data. (2) Data can be obtained: (meta-) data need to be created with an open, free, universal standardized communication protocol to ensure that it can be retrieved; to ensure that the metadata can be obtained even if the data is not available. (3) Data is interoperable: (meta-) data should be described in standardized, accessible, shareable and common language; the vocabulary used in (meta-) data should follow the principle of fairness; the reference of other (meta-) data needs to be included within the (meta-) data. (4) Data is reusable: (meta-) data is the accurate and rich description with specific, accessible language which can meet standards of other relevant fields.

C. Constancy

The increase in connectivity has accelerated the progress of global research and it is estimated that there will be about twice as many scientific outputs each decade[4]. The increasing number of research activities led to the output of a series of

research data. Data management should also follow the principle of sustainability correspondingly to ensure the continuity of research data in the same subject area.

D. Coordination

Research data management involves many departments such as research institutes, funding agencies, IT service agencies, publishing houses, data service centers and libraries. Collaboration principle among different departments should be followed to enhance communication and exchange, save management costs and increase efficiency.

E. Policy

Research data management programs must comply with the research data management policies of funders or research institutes in order to sustain their support and funding, and to ensure the continuity of research data[5].

II. RESEARCH DATA MANAGEMENT STANDARDIZED PROCESS

Research data have the characteristics of diversified production mode and management mode, wide range of application, various application modes, numerous subjects fields, widely distributed, diverse data types and uneven quality status, and it can be seen that research data management is a huge and complex project which needs to be implemented with standardized process. The life cycle of research data in the research process mainly includes procedures of data collection, processing, preservation, dissemination, retrieval, access and utilization, disappear or no longer use of the data. Correspondingly, the standardized process of research data management can be shown in Figure 1, the research team, data information service center and funding agency or research institutes are the responsible units in different stages of research data management, whose coordination are basically all needed in every stage.

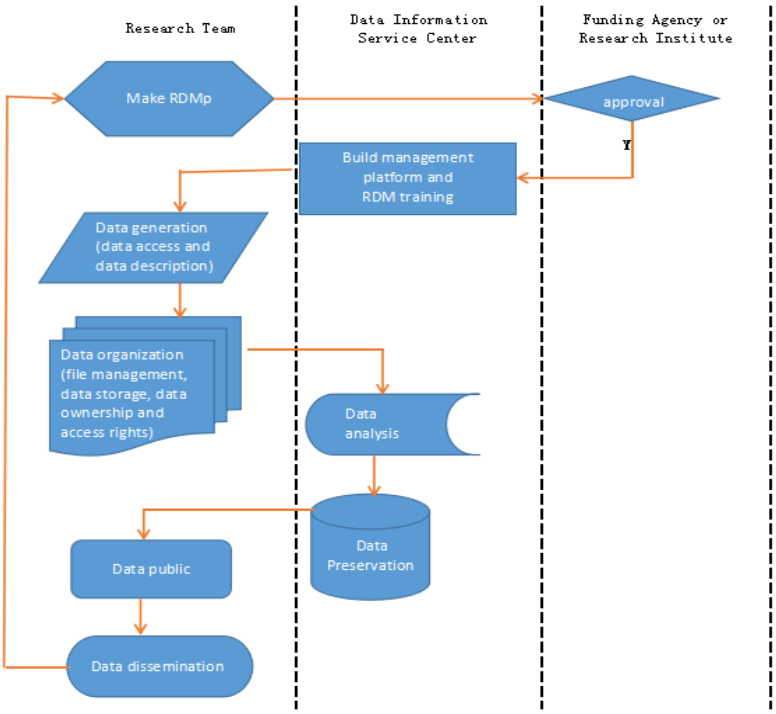

Fig. 1. flow chart of research data management

A. Establish research data management plan

When the project is declared, the scientific research team shall, according to the content of the declared project, formulate a detailed research data management plan which can be completed with the help of personnel of the data information service center. The research data management plan (RDMp) is an official document that outlines the researcher's management of the entire research data lifecycle during and after the research process. Data management programs help to design, implement, and track how research data is collected, organized, used and managed for the highest quality and long-term sustainability. A good plan will save data management costs while increasing the value of the data. When accepting project applications, many projects funding organizations abroad require an attached research data management plan of unified format and content just like our project declaration.

The research data management plan generally includes the following: (1) what types of data will be created. (2) What criteria the data will follow. (3) How to obtain and protect the ownership and intellectual property of the data. (4) How to describe the data. (5) The perimeter of data sharing and reuse. (6) Data management tools. (7) Data management hardware facilities requirements. (8) The person in charge of data management during the project [xii]. (9) Data management cost planning.

B. Set up research data management platform and RDM training before project implementation

After the approval of the scientific data management plan, the technical staff of the data information service center will set up a data management platform according to the needs of scientific researchers. A practical platform should include basic functions of data submission, publication, browsing, inquiry, downloading, data management and user management.

Before the implementation of the project, staff of the center conducts research data management (RDM) training for the research team to understand the data management specification in the field, data management practices and relevant standards in this project area.

C. Data generation

1) Data access

Before obtaining data, researchers and data managers need to (1) investigate and understand some sample data in the field of this project. (2) Understand the data description object and knowledge in the field. (3) Grasp the dynamic and the role of data in management process. (4) Master the international, national and field standard. (5) Master the use of some tools to collect data. (6) Grasp the research data management platform used in this project. Many foreign research-intensive organizations have mature management system platforms in different fields. After mastering above information, it needs to be figured out that what data worth sharing and conserving need to be obtained, and data types, which includes experimental data, observation data, telephone survey data, questionnaire data, structured interview or semi-structured interview data.

2) Data description

a) *Dataset description*

Data description is an important part of data management, which facilitates scientific researchers to understand the data in the future. Meanwhile, by providing sufficient data-related information and standardized description, background of data acquisition could be known, and the visibility of data would be improved and more citations could be accessed. The following basic elements should be considered when describing the data, such as General Overview (title, creator, identifier, date, method, processing, source), Content Description (subjects, place, variable list, code list), Technical Description (file inventory, file formats, file structure, version, necessary software), Access (rights, access information) [6].

b) Metadata description:

Metadata are data about the data, mainly describing information about the attributes of the data, which are used to support functions such as indicating storage locations, history data, and resource lookups, file records, etc. An accurate description of metadata is the key in ensuring that resources can be continued and be accessed in the future. Metadata are as important as scientific data itself, which provides descriptive meaning to raw research data. Many disciplines have their own standard metadata model with fixed elements and structures. Curtin University has viewed common metadata examples [6].

D. Data organization

1) File management

Data management all begins with a file, first of all, data file naming rules need to be normalized: (1) all the files in the entire project need to be named based on the same rules, so as to ensure the consistency of the entire project files; (2) In addition, the file name needs to be concisely and accurately described, without randomly giving a different name to the file, and the date of creation (YYYY-MM-DD or YYYYMMDD) and the version number should be contained in the name; (3) Avoid the use of special characters, such as &%@, etc.; (4) The file name needs to be unique.

2) Data storage

Data storage is very important because the digital storage media format is easy to be invalidated, and all file formats and physical storage media will eventually become obsolete, so the following aspects should be considered in data storage.

- The storage location of data. When choosing a storage location, the first step is to estimate the size of the project's scientific data, and for text data, the storage space is small, and compression needs to be considered for photos and video data. The second step is to consider the data access rights, such as generally who is provided with the data and the right of remote access.
- Data storage format. The format of the data file will be influenced by the methods used to analyze the data, the hardware used to obtain the data, the software and the criteria for different disciplines. The principle of data file storage format is to select the standard, common and more durable format for storage, to ensure the normal preservation and access, and to avoid the difficulties and costs of future data format migration.
- Data storage security. The following measures can be taken in data storage: first, cloud storage. The second is to create a redundant copy. If the research project data is not eligible for cloud storage, then regular backups could be done to produce several redundant copies and store them in different physical locations. But the regular backup needs to be ensured and to be done when there is a big change in the data, if the data can't be automatically backed up, then someone needs to be arranged to ensure the regular backup of the data.

3) Data ownership and access right

Work in this phase includes following aspects: (1) Determine the owner of the research data. The problem seems to be simple, while it is actually quite complicated to be defined. Ownership mainly involves the principal investigators, initiate organizations, funding institutes and anyone involved in the project. (2) Access right. By setting access rights, the confidentiality of research data and original material can be protected.

E. Data analysis

Data manager needs to first choose which kind of analysis software to be used based on the data at hand, and the most important thing is to consider whether metadata and log files

could be generated automatically during processing the data. Those who are unfamiliar with the software in data information center should take the initiative to receive various trainings to enhance their data analysis skills.

F. Data preservation

Data preservation means not only to duplicate data as described above, but also to preserve the data after the accomplishment of the project, consideration from following aspects need to be taken into account : 1) maintain long-term data access; 2) able to be retrieved in the future; 3) prevent loss of data; 4) provide data support for research findings once being challenged; 5) develop long-term preservation plan; 6) designate person in charge of the persistence of data; 7) storage duration; 8) location for long-term data storage, such as data warehouse; 9) data sharing method; 10) laws governing the preservation of data files and the requirements of funding agencies or publishers.

G. Data public

At present, more and more project funding organizations require the publication of research data of its funded projects, at the same time, publishers also require publication of research data on published papers, top journals like Nature and PLOS all require the publication of scientific data in theses. Elsevier and Springer publishing group are also trying to make the data of the paper public. As a researcher, the motivations for the disclosure of data are as follows: first, to increase the visibility of the project; second, to increase the citation rate of research results; third, to obtain more cooperation opportunities with colleagues. As is mentioned previously, data are of different access rights, and it needs to be considered when making public the data that, for data including privacy information, culture and nationality sensitive information, and information protected by intellectual property which is not sure to be released or not, appropriate anonymous measures and different access rights should be implemented to limit the publication. For some confidential information, it is generally not disclosed.

H. Data dissemination and re-use

For unconcealed research data, researchers in the project team and other scholars in this field are likely to reuse the data, thus relevant regulations on data reuse and dissemination should be established under the guidance of funding agencies or research data management policies, such as who can reuse data? How others reuse data? Is it charged for data reuse? Can others disseminate data? In the meantime, for some researchers who are interested in reusing, the relevant person in charge of research data can give them some practical guidance [7].

III. CONCLUSION

At present, research data management in China is still in the groping phase, with the complicated management process, how to effectively manage the research data produced in the project with limited funding and giving full play to the important value of research data require a standardized management process to comply with [8]. The design of a standardized management process aims to ensure the integrity, security, availability, long-term preservation and reusability of the research data so as to save others' research time, and to improve the efficiency of research and speed up the process of scientization.

REFERENCES

[1] Gonzalez, A., & Peres-Neto, P. R. (2015). Data curation: act to staunch loss of research data. Nature, 520(7548), 436.

[2] Ministry of Science and Technology. China will speed up the construction of the national scientific data center.http://news.sciencenet.cn/htmlnews/2017/8/384142.shtm.

[3] Zheng Shurong, Zhao Peiyun(2003). Scientific data sharing management: problems and solutions. China Science and Technology Achievements, 2003(23):8-10.

[4] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., & Baak, A., et al. (2016). The fair guiding principles for scientific data management and stewardship. Scientific Data, 3(160018 (2016)), pp.167-172.

[5] JISC.How and why you should manage your research data: a guide for researchers.https://www.jisc.ac.uk/guides/how-and-why-you-should-manage-your-research-data.

[6] Curtin University. Research data management: Data description .http://libguides.library.curtin.edu.au/c.php?g=202401&p=1333152.

[7] Perrier, L., Blondal, E., Ayala, A. P., Dearborn, D., Kenny, T., & Lightfoot, D., et al. (2017). Research data management in academic institutions: a scoping review. Plos One, 12(5).

[8] Chu Wenjing, Xu Wenxian(2009). Research on digital resource procurement process. Information studies: Theory & Application, 32(4),pp.75-78.