Characteristics Analysis of Research Data Repositories in Humanities and Social Science - Based on Re3data.Org

Zheng Li [a], Wenying Liu [b *]
Xiamen University Library, Xiamen,361000, China.
[a] LZ82@xujc.com, [*, b] liuwy@xmu.edu.cn,

Keywords: research data, data repositories, data management, humanities and social science.

Abstract. Based on the open data storages catalog and registration system of the re3data.org platform, this article analyzes the characteristics of data repositories in humanities and social sciences field, in order to promote the development of humanities and social sciences data management services in China.

1. Introduction

Research data is an important foundation for scientific work. The diversity of these data reflects the breadth of different scientific disciplines, research interests, and research methods. Research data may include measurements, laboratory values, audiovisual information, texts, survey data, collections of objects, or samples created, developed, or evaluated during scientific work. Test methods such as questionnaires, software, and simulations may also produce important results for scientific research and should therefore be classified as research data. Long-term archiving and accessibility of research data contribute to the traceability and quality of scientific work, enabling researchers to carry out other tasks.

Re3data.org is a global registry of research data repositories that covers research data repositories from different academic disciplines. It presents repositories for the permanent storage and access of data sets to researchers, funding bodies, publishers and scholarly institutions1. Until March 13, 2018, there were 2034 data Repositories registered at re3data.org, of which 600 were Humanities and Social Sciences, accounting for 29.5% of total Repositories. This paper uses these 600 data Repositories to study the characteristics of the humanities and social sciences data Repositories from the subjects, countries and time of construction, institution responsibility type data and database accessed.

2. Characteristics Analysis of Data Repositories in Humanities and Social Sciences

2.1 Subjects

The 600 data repositories are all multi-disciplinary. As can be seen from Figure 1, the number of data repositories in the field of social sciences and economics is the most, while the number in literary studies, theology and philosophy is the least.

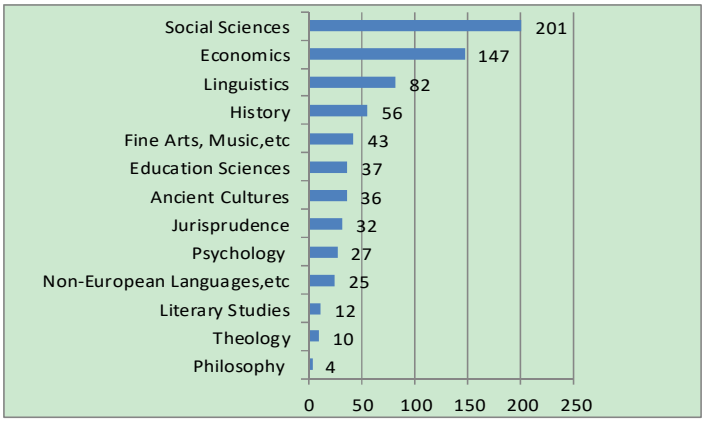

Figure 1. Discipline distribution of data repositories

2.2 Construction Time

Among all the 600 data repositories, 48 have a clear representation of construction time, the earliest can be traced back to 1556 --the SLUB2 commissioned by the German Research Foundation, which contains a collection of contemporary art, photography, industrial design and commercial art as well as other valuable documents such as history technology. The other data storage established before the 19th century is SSRQ-online3, which is a series of legal sources edited by the Swiss Bar Association since 1898 and has published more than 100 volumes and more than 70,000 pages of source materials until now. We disclosure the number of data Repositories which describes the construction time in Figure 2, it shows clearly that since the 1960s, the number of data Repositories has been on the rise.

Compared with developed countries in Europe and America, China's humanities and social sciences data service started late, and data Repositories were established after 2010. For example, National Data4 was established by the National Bureau of Statistics in 2013 to provide monthly, quarterly and annual census, regional, and international socio-economic statistical data. And in 2015, in order to meet the needs of research data management, Peking University Open Research Data5 platform was online, which provides preservation, management and distribution services for research data.

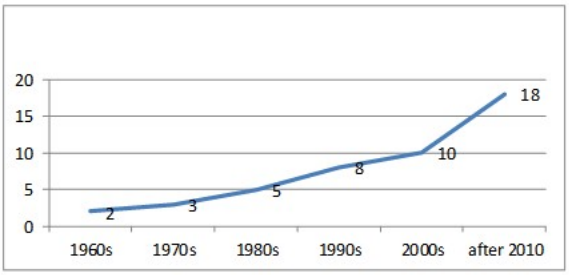

Figure 2. construction time distribution of data repositories

2.3 Construction Countries

We can see from Figure 3, the United States, Germany, the United Kingdom, and the European Union established the most number of data Repositories in humanities and social sciences. Meanwhile, there are only three Repositories in China's mainland and one in Taiwan. There is still a big gap between China and Europe, as while as the United States.

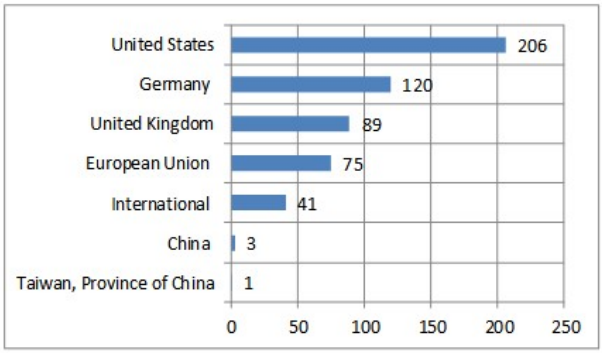

Figure 3. Construction countries distribution of data repositories

2.4 Data and Database Access

The openness of data repositories can be reflected in three aspects (Table 1): Database access, Data access and Data upload. Access levels are divided into: open, closed, and restricted. Fully open means that there are no access barriers. Restricted use means that external users can overcome access barriers. Not opening means that external users cannot overcome entry barriers. Embargoed is only applicable to the level of access to research data sets and means that external users cannot overcome access barriers until the data is released for open or restricted access.

From the perspective of database access permissions, most of the databases are completely open, accounting for 95%, and only 0.3% of databases are not open, 4.7% of databases are restricted; Judging from the data access rights, nearly half (46%) of the data is completely open, restricted data

use accounts for 36.5%, another 8.5% of data is not open, and 8.6% of data is forbidden; From the perspective of data upload permission, only 3% of the data is completely open, and more than half (66%) of the data are uploaded, the other 31% of data uploads are not open.

Table 1. openness of data repositories

	Database access	Data access	Data upload
open	95 %	46 %	3 %
closed	0.3 %	8.5 %	31 %
embargoed	-	8.6 %	-
restricted	4.7 %	36.5 %	66 %

2.5 Institution Responsibility Type

According to the analysis, Funding institutions build the most number of data repositories, such as the NSF and NIH, the European Commission, the German Research Foundation, the UK's ESRC and JISC, followed by the universities' general organizations, for example, University of Michigan, University of California, Cambridge University, etc. We list the relevant policies of research data management from the top (Table 2), these policies are an important guarantee for scientific data management.

Table 2. policies of research data management of part institutions

Institution name	number	policies of research data management
National Science Foundation (NSF)	44	NSF issued the "Project Management Guide"[6]in 2010, requiring that: Starting from January 18, 2011, all project applications must include a data management plan of no more than two pages, and a common data management plan template has been developed.
Michigan,Inter-University Consortium for Political and Social Research (ICPSR)	30	ICPSR[7] maintains and provides an extensive archive of social science data for research and teaching. Since 1963, ICPSR has provided quantitative method training to promote effective data use. The ICPSR Social Research Quantitative Method Summer Course offers comprehensive courses in research design, statistics, data analysis and methodologies. In order to ensure that mathematical resources are available to future generations of academics, ICPSR plans and preserves data and migrates it to new storage media and file formats as a technology change. In addition, ICPSR provides user support to help researchers identify relevant data for analysis and conduct research projects.
European Commission	26	The GRDI[8] 2020 project released the "Global Scientific Data Infrastructure: Major Data Challenges" report, which presents policy challenges and recommendations for building a global scientific data infrastructure, including: ① The need to develop scientific data infrastructure to support open-linked data spaces, support for data, and Interoperability between documents, support for data-intensive research, support for multidisciplinary and interdisciplinary research, and scientific ecosystems. ② To develop new data tools, data models and query languages. 3 Develop professional data experts and research teams.
National Institutes of Health (NIH)	25	In 2003, the "Data Sharing Policy and Implementation Guidelines" [9]was issued. Requirements: ① All scientific researchers who applied for funding to the NIH for more than $500,000 from the 1st of October 2003 must submit a data sharing management plan or data sharing instructions. ②Applicants can apply for financial support for data management. ③NIH requires that the data types be transferred for 3 years after completion of the project. ④Sharing methods include personal websites, institutional websites, or storage to databases. ⑤ Assessment of implementation of data management plan by project staff.
German Research Foundation(DFG)	20	In 2010, the German Federation of Scientific Organizations [10]expressed its support for the "Research Data Processing Principles" in the practice of long-term archiving of research data, open access, and compliance with conventions of various disciplines. DFG General Principles for Study Data Processing: ①The applicant should specify in the proposal the research data to be generated or evaluated in the scientific research project. ② It must be ensured that access to data can still be guaranteed when the rights to use research data are transferred to third parties (usually publishers) through publishing. ③ Study data should be archived in the investigator's own institution or appropriate national infrastructure for at least 10 years.

3. Suggestions

The United States, Germany and United Kingdom lead the construction and sharing of scientific research data repositories in the humanities and social sciences. Through the analysis of the characteristics of data repositories, China can learn from the mature construction experience of developed countries and build high-impact data warehousing with its own domain characteristics and integrate data management services.

First, the biggest obstacle of research data management is the lack of user awareness; the scientific researchers have no enthusiasm to deposit scientific research results. So, we have to intensify publicity efforts and enhance openness and sharing awareness.

Second, data policy is the basis for data sharing. The U.S., Germany and British governments have intensively release open data related policies and strategic plans to support and promote data sharing and improve data utilization. We can use this as a reference; formulate relevant policies to promote open scientific development.

Third, a good institutional cooperation atmosphere is an important factor in promoting open data. Data management is a huge project that is difficult to accomplish with only one effort. The construction of integrated data infrastructure will not only help enhance information exchange and create a good environment for cooperation, but also enable the sharing of data and results in the research process.

Fourth, talent is the guarantee of career development. As a new business, whether the open data can be successfully implemented depends to a large extent on the support of professional teams. Since it involves the general public including scientists, publishers, and scientific researchers, etc., professionals who are engaged in this task must not only have profound professional skills but also have good communication and communication skills. Therefore, it is imperative to promote personnel training and team building.

References

[1]. re3data.org. [EB/OL]. [2017-10-23]. https://www.re3data.org/about

[2]. Slub Collections. [EB/OL]. [2017-10-23]. https://www.slub-dresden.de/en/collections/

[3]. SSRQ-online. [EB/OL]. [2017-10-23].https://www.ssrq-sds-fds.ch/online/

[4]. National Data. [EB/OL] [2017-10-23]. http://data.stats.gov.cn/

[5]. Peking University Open Research Data. [EB/OL]. [2017-10-23]. http://opendata.pku.edu.cn/

[6]. NSF Data Sharing Policy. [EB/OL]. [2017-10-23].https://nsf.gov/bfa/dias/policy/dmp.jsp

[7]. Inter-university Consortium for Political and Social Research. [EB/OL]. [2017-10-23]. https: //en.wikipedia.org/wiki/Inter-university_Consortium_for_Political_and_Social_Research

[8]. Global Research Data Infrastructures: The GRDI2020 Vision. [EB/OL]. [2017-10-23]. http: //www.grdi2020.eu/Repository/FileScaricati/fc14b1f7-b8a3-41f8-9e1e-fd803d28ba76.pdf

[9]. NIH Data Sharing Policy. [EB/OL]. [2017-10-23]. https: // grants. nih. Gov / grants/ policy/ data_ sharing/

[10]. DFG Guidelines on the Handling of Research Data. [EB/OL]. [2017-10-23]. www. dfg.de /download/pdf/ foreordering/./guidelines_research_data.pdf