The Study on Distribution of Research Data Repositories Based on Resources Directory Websites

Zhi-Feng ZHOU[1,2,*]

[1]School of Information Management, Wuhan University, Wuhan, China

[2]The Principal's office, Wenzhou University, Wenzhou, China

[a]wzulib@126.com

*Corresponding author

Keywords: Research data repositories; Research data; Scientific data; Data management

Abstract. This paper thinks Research Data Repositories are vital platforms and tools for sharing and accessing research data. This paper analyzes the resources directory websites of research data repositories and discusses the distribution of research data repositories by online survey and statistical analysis of data. This study can provide valuable information for stakeholders of research.

Introduction

With the rise of e-science and data-intensive research, research data have always been at the core of much scientific research, so more and more universities and research institutes are starting to build research data repositories in order to achieve the purpose of allowing permanent access to research datasets in a safe and reliable information environment. [1][2][3]

Open access and sharing of research data is actually needed

The long-term preservation and access to research data is a challenge for all stakeholders in the scientific community. However, the long-term preservation and the principle of open access to research data can offer broad opportunities for the scientific community.

In addition, as more and more funders and academic journals adopted data policies that require researchers to deposit underlying research data in data repositories, the question how to choose an appropriate repository becomes more and more important. In general, data policies required fund grantees and paper authors to ensure the accessibility of research data generated within the scope of a project or as the basis of a publication. For example, The National Science Foundation (NSF) requires applicants "to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants."[4] The NSF further requires that measures for the implementation of this policy be specified in a "Data Management Plan". [5] In Nature journal, related dataset of papers must be made freely available to readers from the date of publication, and submission of several types of dataset to some public repositories is mandatory.[6]

Research Data Repositories are vital platforms for sharing and accessing research data

In order to promote data sharing of scientific research community, researchers would require infrastructures and platforms that ensure a maximum of accessibility, stability and reliability of research data. Such infrastructures and platforms are being increasingly summarized under the term Research Data Repositories (RDR). [7] RDR store a wide variety of file formats under different conditions for access and reuse. RDR and their services are mostly characterized by the various

scientific discipline. The researchers hope that RDR can become the right-hand man of scientific research, but it is difficult for them to find an appropriate repository for the storage of their data.

Research purpose of this paper

Research data are varied and ubiquitous, and RDR were also established by various institutions dispersedly. Thus it is difficult for researchers, funding agencies, publishers and academic research institutions to select appropriate repositories for research data management. One of the obstacle to wise choice is the lack of knowledge on already existing RDR by stakeholders of research data management, especially the information of distribution of RDR. Therefore, in order to meet the demand of stakeholders of research data management, this paper attempted to obtain the distribution of RDR through the online survey.

Research methods and design

Although the distribution of RDR is irregular in Cyberspace, the emerging of one-stop resources directory websites of RDR offered an opportunity for obtaining RDR information. These resources directory websites can gather major RDR worldwide. By these resources directory websites, stakeholders may select to appropriate RDR according to their needs. We can also study the distribution of RDR by analyzing these resources directory websites.

Selecting Samples of resources directory websites

(1) Sample one: Registry of Research Data Repositories(Re3data) [8]

Re3data, which is a global registry of research data repositories, is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG). Re3data covers research data repositories from different academic disciplines. Re3data aims to promote a culture of sharing, increased access and better visibility of research data.

Re3data presents repositories for the permanent storage and access of data sets to researchers, funding agencies, publishers and academic research institutions. Re3data helps researchers to find appropriate repositories for the storage and access of research data. Further, it can be used by funding agencies to promote permanent access to research data from their funded research projects. In addition Re3data offers publishers and research institutions an online tool for the identification of research data repositories where researchers can deposit and share their research data.

(2) Sample two: Databib [9]

Databib, which catalog the research data repositories in the whole world, is a online tool or directory for helping researchers find, identify and locate repositories of research data. The development of Databib was originally sponsored by a Sparks! Innovation National Leadership Grant from the Institute of Museum and Library Services(IMLS). Databib attempts to address these needs for the stakeholders of research community, including researchers (data users and data producers), publishers and professional societies, librarians, research funding agencies etc..

Over 600 research data repositories have been cataloged in Databib, with more being added every week. Information professionals and users catalog and curate metadata records that describe research data repositories.

Selecting analysis Methods

This paper carried out appropriate research through online survey and statistical analysis of data. We browsed and retrieved the two websites by the user's perspective, and collected related data.

Re3data and Databib provide overview of existing RDR and related information.

Online survey time: 18-20 March, 2014.

Results and discussion

In March 19, 2014, Re3data lists 586 research data repositories, more than 400 of these are described in detail by a comprehensive vocabulary.

In March 20, 2014, there were 624 research data repositories total in Databib.

By analysis of the related information retrieved, we can find distribution of the RDR. Then we mainly analyzed distribution of the RDR from four aspects (the countries, subjects, content types, repositories types).

The distribution of countries on RDR

As shown in Table 1 and Table 2, RDR have appeared in every continents and regions all over the world, so we think the concept of research data management has been widely recognized. According to the distribution of countries on RDR, the top three are the United States, United Kingdom and Canada, accounting for 67.58% in Re3data and 72.76% in Databib.

The reasons for this distribution are as follows: firstly, these countries with higher economic and technology level have the strength to build RDR for storage and sharing of research data; secondly, the research data management as concept and technology first appeared in European countries and the United States, and RDR have become one of important infrastructures in e-science.

Table 1. The distribution of countries on RDR (Re3data)

Serial number	Countries	The Number of repositories	Proportion (%)	Serial number	Countries	The Number of repositories	Proportion (%)
1	United States	270	46.08%	25	South Africa	3	0.51%
2	United Kingdom	90	15.36%	26	Republic of Korea	3	0.51%
3	Canada	36	6.14%	27	Lithuania	3	0.51%
4	France	34	5.80%	28	India	3	0.51%
5	Germany	25	4.27%	29	Czech	3	0.51%
6	Japan	23	3.92%	30	Slovenia	2	0.34%
7	Belgium	23	3.92%	31	Romania	2	0.34%
8	Netherlands	19	3.24%	32	Panama	2	0.34%
9	Switzerland	17	2.90%	33	New Zealand	2	0.34%
10	Denmark	14	2.39%	34	Mexico	2	0.34%
11	China	13	2.22%	35	Indonesia	2	0.34%
12	Australia	12	2.05%	36	Senegal	1	0.17%
13	Spain	10	1.71%	37	Polynesia	1	0.17%
14	Russian	10	1.71%	38	Poland	1	0.17%
15	Luxembourg	9	1.54%	39	Israel	1	0.17%
16	Italy	9	1.54%	40	Ethiopia	1	0.17%
17	Austria	9	1.54%	41	Estonia	1	0.17%
18	Sweden	6	1.02%	42	El Salvador	1	0.17%
19	Norway	6	1.02%	43	Costa Rica	1	0.17%

20	Greece	5	0.85%	44	Burkina Faso	1	0.17%
21	Finland	5	0.85%	45	Brazil	1	0.17%
22	Ireland	4	0.68%	46	Benin	1	0.17%
23	Hungary	4	0.68%	47	Azerbaijan	1	0.17%
24	Ukraine	3	0.51%	——	——	——	——

Table 2. The distribution of countries on RDR (Databib)

Serial number	Countries	The Number of repositories	Proportion (%)	Serial number	Countries	The Number of repositories	Proportion (%)
1	United States	343	54.97%	19	New Zealand	2	0.32%
2	United Kingdom	69	11.06%	20	Norway	2	0.32%
3	Canada	42	6.73%	21	Russia	2	0.32%
4	India	25	4.01%	22	Spain	2	0.32%
5	Australia	20	3.21%	23	Austria	1	0.16%
6	China	20	3.21%	24	Brazil	1	0.16%
7	Germany	17	2.72%	25	Cyprus	1	0.16%
8	Japan	14	2.24%	26	Egypt	1	0.16%
9	France	7	1.12%	27	Estonia	1	0.16%
10	Sweden	6	0.96%	28	Ireland	1	0.16%
11	Switzerland	6	0.96%	29	Kenya	1	0.16%
12	Denmark	5	0.80%	30	Lithuania	1	0.16%
13	Netherlands	5	0.80%	31	Luxembourg	1	0.16%
14	Belgium	4	0.64%	32	Pakistan	1	0.16%
15	South Africa	4	0.64%	33	Panama	1	0.16%
16	Greenland	3	0.48%	34	Slovenia	1	0.16%
17	Finland	2	0.32%	35	Unspecified	10	1.60%
18	Italy	2	0.32%	——	——	——	——

The distribution of subjects on RDR

In Re3data, the staff marked more than one subject for every research data repositories. According to statistics, Re3data has 152 subject categories. This paper lists the top 30 subject categories marked frequency (See Table 3).

As shown in Table 3 and Table 4, the subject categories of RDR have mainly included categories of natural science. For example, life sciences, biological sciences, geosciences, geography, environmental sciences, etc.. Compare to the disciplines of the humanities and social sciences, the disciplines of the natural sciences have stronger demand on RDR. In the research process of the disciplines of the natural sciences, the massive research data would be generated.

Table 3. The distribution of subjects on RDR (Re3data)

Serial number	Subjects	The number of repositories	Proportion (%)	Serial number	Subjects	The number of repositories	Proportion (%)
1	Natural Sciences	311	53.07%	16	Astrophysics and Astronomy	76	12.97%
2	Life Sciences	235	40.10%	17	Basic Biological and Medical Research	74	12.63%
3	Geosciences	191	32.59%	18	Engineering Sciences	64	10.92%
4	Geography	191	32.59%	19	Zoology	54	9.22%
5	Social Sciences	169	28.84%	20	Geophysics and Geodesy	51	8.70%
6	Humanities and Social Sciences	169	28.84%	21	Geophysics	51	8.70%
7	Humanities	169	28.84%	22	Microbiology, Virology and Immunology	48	8.19%
8	Biology	162	27.65%	23	Immunology	48	8.19%
9	Physics	146	24.91%	24	Economics	44	7.51%
10	Medicine	131	22.35%	25	Virology	40	6.83%
11	Behavioural Sciences	110	18.77%	26	Plant Sciences	35	5.97%
12	Oceanography	98	16.72%	27	Water Research	34	5.80%
13	Atmospheric Science and Oceanography	98	16.72%	28	Empirical Social Research	33	5.63%
14	Atmospheric Science	98	16.72%	29	General Genetics	27	4.61%
15	Chemistry	85	14.51%	30	Computer Science, Electrical and System Engineering	26	4.44%

Table 4. The distribution of subjects on RDR (Databib)

Serial number	Subjects	The number of repositories	Proportion (%)	Serial number	Subjects	The number of repositories	Proportion (%)
1	Biological Sciences	149	23.88%	12	Education	5	0.80%
2	Environmental Sciences	83	13.30%	13	Language and Literature	3	0.48%
3	Geosciences	63	10.10%	14	Fine and Performing Arts	3	0.48%
4	Mathematical and Physical Sciences	57	9.13%	15	Engineering	3	0.48%
5	Health and Medical Sciences	52	8.33%	16	Communications and Information Sciences	3	0.48%
6	Social Sciences	46	7.37%	17	Law and Legal Studies	2	0.32%
7	Multidisciplinary	33	5.29%	18	Business	2	0.32%
8	Ecosystem Sciences	15	2.40%	19	Philosophy and Religion	1	0.16%
9	Agriculture	13	2.08%	20	History	1	0.16%
10	Area, Ethnic, and Gender Studies	9	1.44%	21	Unclassified	75	12.02%
11	Interdisciplinary	6	0.96%	——	——	——	——

The distribution of data content types on RDR

Data formats of each research data repository are not single. As shown in Table 5, RDR can support multiple content types, including the structured data and non-structured data.

Table 5. The distribution of content types on RDR (Rre3data)

Content types	The Number of repositories	Proportion (%)	Content types	The Number of repositories	Proportion (%)
Scientific and statistical data formats	370	63.14%	Audiovisual data	128	21.84%
Standard office documents	334	57.00%	Software applications	100	17.06%
Images	313	53.41%	Databases	83	14.16%
Plain text	309	52.73%	Networkbased data	55	9.39%
Raw data	276	47.10%	Source code	20	3.41%
Structured graphics	247	42.15%	Configuration data	14	2.39%
Structured text	227	38.74%	Other	140	23.89%
Archived data	151	25.77%	——	——	——

The distribution of repositories types on RDR

According to the analysis of Re3data and Databib, RDR can be divided into four types as follows: institutional RDR, disciplinary RDR, multidisciplinary RDR, and project specific RDR.

(1) Institutional RDR are run by the institutions such as universities or academic institutions. Such as Griffith University Research Data Repository, Purdue University Research Repository, Cornell University Geospatial Information Repository and so on.

(2) Disciplinary RDR collect research data of a certain discipline. For instance, the typical examples in the field of biology RDR are GenBank, NCBI Protein and TreeBASE.

(3) There are RDR serving multidisciplinary needs. Figshare is a RDR example that "allows researchers to publish all of their data in a citable, searchable and sharable manner."

(4) The landscape of RDR with a specific focus on the research data resulting from particular research projects is also diverse. Integrated Ocean Drilling Programm(IODP) can be named as being exemplary here.

Conclusion

With the development of e-Science, the number of RDR would continue to grow. As one of the infrastructure of research data management, its role would continue to highlight. Based on the survey and analysis on Re3data and Databib, we had generally understood the distribution of RDR. For researchers, this study can eliminate the blindness of searching RDR information and provide reference information to choose appropriate RDR. For librarians, this study can help them grasp the related RDR knowledge for providing information service.

Acknowledgement

The author wish to express their gratitude to the Wuhan university for "The tracking programs of research progress of the humanities and social sciences abroad" in 2013 (Project name: The latest research progress of international scientific data management).

References

[1] Hey, T., Tansley, S. & Tolle, K. (Eds.). (2010). The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA: Microsoft Research.

[2] E-Science and Data Support Services(2014-03-18). Available at:

http://www.arl.org/storage/documents/publications/escience-report-2010.pdf.

[3] Research data management at institutions:Visions, bottlenecks and ways forward(2014-03-18).

Available at: http://libraryconnect.elsevier.com/articles/RDM.

[4] National Science Foundation Proposal and Award Policies and Procedures Guide. Chapter VI - Other Post Award Requirements and Considerations(2014-3-20).

Available at: http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4.

[5] National Science Foundation Proposal and Award Policies and Procedures Guide. Grant Proposal Guide. Chapter II - Proposal Preparation Instructions(2014-3-20). Available at:

http://www.nsf.gov/pubs/policydocs/pappg uide/nsf11001/gpg_2.jsp#dmp.

[6] Availability of data and materials(2014-3-20). Available at:

http://www.nature.com/authors/policies/availability.html.

[7] Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.-J., Gundlach, J., Schirmbacher, P., Dierolf, U. (2013). Making Research Data Repositories Visible: The re3data.org Registry. PLOS ONE, 8(11), e78080. doi:10.1371/journal.pone.0078080.

[8] Registry of Research Data Repositories(2013-03-21).Available at: http://www.re3data.org.

[9] About Databib (2013-03-21). Available at: http://databib.org.