Submitted on: 19.07.2016

RDM: Exploration and Practices of Academic Libraries - Partnerships, Collaboration, Expertise

Nie Hua
Peking University Library, China
hnie@lib.pku.edu.cn

Abstract:

In recent years, RDM has become increasingly important under the context of Open Science and Open Data. The re3data.org already has over1500 data repositories registered. Based on the practice of Peking University Open Research Data Project, this paper will discuss issues of awareness and demands of data providers and users, data policies of both institutional and funding parties, investigation and selection of data management platform, localization of OSS, and the support and collaboration within institutions such as between administrative units and data owners.

Peking University Library started planning the Open Research Data Project at the beginning of 2014. However, it seems that we lacked positive demands from the researchers, necessary institutional level policies and partners to cooperate with. After seeking support both internally and externally, and raising awareness of research data as an important format of academic output, Peking University got granted a NSFC (National Science Foundation of China) project with a part of it goes to development of Research Data Platform.

With collaboration of the Institution of Social Science Survey and University Research Administration Units and after one year's development, Peking University Open Research Data Platform launched at the end of 2015. Services provided include completed data capturing, managing and publishing, sharing and storing, DOI registration and authoritative data citing, access and copy right control mechanism, data version archiving, usage statistics, tracking and reporting, online data analysis and visualization plus digital fingerprint, Chinese and English bilingual interface. The Dataverse based platform opens fully to the academics and 2 months after its launch, 15 dataverses created including China Survey Data Archive, China Family

Panel Studies, China Health and Retirement Longitudinal Study, Center for Bioinformatices, PKU, Visualization and Visual Analytics Research Group PKU, etc. More and more datasets have been capturing and open to reuse and cite.

1. Background：

The "Science as an Open Enterprise" report published by the Royal Society in 2012 highlights the need to grapple with the huge deluge of data created by modern technologies in order to preserve the principle of openness and to exploit data in ways that have the potential to create a second open science revolution. Open inquiry is at the heart of the scientific enterprise. Publication of scientific theories, and of the experimental and observational data on which they are based, permits others to identify errors, to support, reject or refine theories and to reuse data for further understanding. Science's powerful capacity for self-correction comes from this openness to scrutiny and challenge. [1]

In recent years, RDM has become increasingly important under the context of Open Science and Open data. While scientists have become more open among themselves, as well as with the public and the media, more experts in managing and supporting the use of data are required. Librarians, as an important stakeholder in the open access movement and in scholarly communication, are called in the team of practices and developments of open research data.

Re3data.org has already reached a milestone in identifying and listing 1500 research data repositories since its launch in 2012. [2] However, there is only less than a dozen research data repositories on re3data.org related to China.

Peking University Library (PKUL) has long been paying close attention to scholarly communication and open access while seeking opportunities to raise awareness, foster collaboration and initiate projects in and out of the university. Since 2010，as one of the important and innovative functions of academic library to support the dynamic changing environment of scholarly communication, Peking University Library initiated multiple initiatives of open access and scholarly communication: PKU Institutional Repository[3], PKU Open Journals[4], Scholars @ PKU[5] and PKU Open Research Data[6]. Among them, the PKU Open Research Data focuses on facilitating more effective and efficient data sharing and preservation practices, and provides incentives for making data easily accessible among researchers.

2. Need Assessment

Prior to the PKU Open Research Data project, PKUL conducted a campus wide survey of RDM requirements among researchers and research teams in 2013. The purpose of the survey was to understand the real needs of researchers in order to design research data management services. Issues covered by the survey included: awareness and practices of research data management and data sharing; features of specific scientific data of the research teams; current status of data management of the research team; re-use of the data; and the expectation to research data management services. The results showed that 87.5 percent of respondents were willing to share research data under certain conditions. The biggest motivation was related to the value of shared data, positive relation between data usage and the citation of the academic outputs and the exposure of the data. While the biggest concern was possible plagiarism which due to shared data.

Interviews with researchers conducted by the library showed three major trends of research data management. Firstly, research data sharing behavior showed very strong disciplinary-oriented features. For data-driven and data-intensive disciplines, which already practice open access maturely, data sharing usually follows certain standards and norms. Secondly, researchers generally require an embargo period with data sharing. Almost all researchers interviewed stressed that data should be published after their results are published, due to concern of possible plagiarism. Thirdly, data sharing behavior is mostly spontaneous, and lacks proper incentives and necessary maintenance, and lacks a well-managed mechanism for data citation, academic recognition and feedback.

3. PKU Open Research Data

3.1 Objectives：

The findings of the survey and interviews show a strong case for open data while there was only very limited subject-specific or research team-oriented data storage efforts and data services either on the national level or the institutional level. The PKU Open Research Data project was a respond to these challenges, with its goal to develop infrastructure to support PKU researchers in managing their data more effectively, and providing services ranging from advice to storage repositories.

The objectives of PKU Open Research Data project could be summarized as:

- Publish high-quality research data to disseminate academic outputs through open platform;
- Promote open science to encourage reuse and reproduction of research data;
- Foster practices and metrics of research data citation;
- Explore data publishing and long-term preservation solutions;
- Foster innovation and cross-disciplinary integration.

3.2 Cooperation:

Lyon[7] maps potential roles of the library to a research lifecycle model in 10 stages and identify potential partner services at several points: 1. RDM requirements gathering – through auditing (with academic departments); 2. RDM planning – advocacy and guidance to researchers at all levels including PGR (with doctoral training centers); 3. RDM informatics – technical advice on data formats and metadata; 4. research data citation; 5. RDM training – training to researchers including PGR (with doctoral training centers); 6. research data licensing; 7. research data appraisal – guidance on which data to keep; 8. research data storage (with IT services); 9. research data access; 10. research data impact (with research support offices). As the one of the most active advocators of RDM on the campus, we understand the importance of cooperation and worked very hard in seeking potential partners and supporters in the University.

Based on previous successful experience of cooperation and through in depth discussion, PKUL and the Peking University Institute of Social Science Survey (ISSS) reached a consensus to cooperate on developing research data archives and services together. ISSS acts as a social science data survey coordinator and interdisciplinary empirical research platform that enables Peking University as well as other research institutions around the world to study China's social problems and conduct social science research, mainly through undertaking large scale social survey projects. ISSS aims to provide leadership and training in data access, curation, and methods of analysis for the social science research community.

In 2014, the Peking University was approved by the National Natural Science Foundation of China with the China Survey Data Archive (CSDA) project, which aims to develop and support a data repository under the Peking University Management Science Data Center, an affiliated organization of ISSS. This became another opportunity to promote cooperation between PKUL and ISSS, along with involvement of the administrative units of the University such as the Office of Science Research and the Office of Social Science. According to the assigned responsibilities of cooperating parties, Peking University Management Science Data Center, under the supervision of ISSS, is responsible for research data collection and cleaning, standardization and analysis, and data repository platform testing and feedback. Peking University Library is responsible for requirements and functional design as well as the development and maintenance of data repository, data storage, classification, management and associated services.

3.3 Data Repository Platform:

One of the major tasks of PKUL is choosing, utilizing, deploying, and developing the best possible data repository platform to meet the objectives of PKU Open Research Data Project. The PKUL Team made a comprehensive investigation on several major repository platforms of research data repositories, including both commercial software and open source software, on issues of business model, subject coverage, core functionality and service models. PKUL has long been active in taking advantage of open source software. We adopted DSpace for our Institutional Repository, and Open Scholars for Scholars @PKU. This time open source software became the preferred choice again due to its open architecture, lower cost, rich and useful functions, and active development community. Among the several most frequently adopted OSS research data repositories, Dataverse, Data Conservancy, CKAN and Dspace had been chosen for local implementation and trial.

Dataverse was finally chosen because it meets Peking University's core requirements for open research data platform. The criteria we valued Dataverse most included: metadata standard and good interoperability, completed and flexible permissions management and data access control, data publishing function associated with persistent identifier and version control, and online data analysis and visualization.

As Dataverse 4.0 has significant improvements in data discovery, interface and user experience, and support of multidisciplinary data, we migrated to Dataverse 4.0 from Dataverse 3.3 in June 2015.We also completed major localization development including bilingual interface and bilingual content by adding second language information blocks for dataverses and datasets, improved user group management functions, added data request function associated with application of joining user group, supported online view, export and visualization of usage statistics data, featured dataverses in home page, and DOI registration and SSO Authentication for PKU researchers.

3.4 Database Content (Research Data Deposited)

Peking University Open Research Data platform currently hosts 15 Dataverses, 77 datasets and 278 data files, which can be grouped into seven major categories: survey data （23）, open source coding and tools（6）, bioinformatics database (30), network data extracted from social media (10), climate data (3), linguistic data (4) and chemoproteomics data (1). Datasets were deposited by both research institutions and researchers from disciplines including Social Science, Computer Science, Health and Life Science, Earth and Environmental science, Atmospheric and Oceanic Science, Chemistry, etc. The most common file types include text files (txt, doc, pdf), data files (dta, spss, xls, csv, etc), image files (jpg, png, pdf), and spectrum data.

Several Dataverses are as following:

- China Family Panel Studies, CFPS
  http://opendata.pku.edu.cn/dataverse/CFPS
- China Health and Retirement Longitudinal Study, CHARLS
  http://opendata.pku.edu.cn/dataverse/CHARLS
- The Research Center For Contemporary China
  http://opendata.pku.edu.cn/dataverse/RCCC
- Center for Healthy Aging and Development Studies
  http://opendata.pku.edu.cn/dataverse/CHADS
- Data and Information Management Group, Peking University
  http://opendata.pku.edu.cn/dataverse/DAIM
- GIS software (Geosoft) Laboratory, Peking University
  http://opendata.pku.edu.cn/dataverse/pkugeosoft
- Visualization and visual analytics research group, Peking University
  http://opendata.pku.edu.cn/dataverse/DAIM

As an example, China Family Panel Studies (CFPS)[8] is a nationally representative, annual longitudinal survey of Chinese communities, families, and individuals launched in 2010 by ISSS. The CFPS promises to provide to the academic community the most comprehensive and highest-quality survey data on contemporary China.

3.5 Services provided and planned:

The PKU Open Research Data Platform provides open access of all research data deposited to users from all over the world, and more services are planned to promote the management, open access, reuse and citation of research data, along with long-term data storage and preservation.

- Data curation: with further cooperation with the Management Science Data Center, Peking University Library will work with researchers to improve the procedure of metadata cataloging, data creating, data format transferring and data depositing and publishing.
- Data affiliated publications tracking: clarify and establish correspondence relations between the research data and its affiliated publications, with the deposit of the publications into PKU institutional repository, to create two-way linkage between data and publications.
- User training: hold user training workshops for institutions and researchers, and to distribute usage statistics and functions updates of platform establish via mailing list of dataverse managers.

3.6. Usage and User Experiences:

Since its launch at the end of Dec 2015, PKU Open Research Data Platform has been providing full open access to identified data for more than 6 months. As of May 9th, 2016, 432 users have logged onto the platform. Among them 117 are IAAA user while 314 are registered users. The total downloads of research data is 3,762. Visitors are from 21 countries and regions all over the world. According to the registration information, visitors are from various disciplines and fields.

4. Understanding the barriers：

4.1 Policies

In May 2014, the National Science Foundation of China (NSFC) along with The Chinese Academy of Sciences (CAS) announced open access policy. The mandatory OA policies encourage authors to publish articles via immediate open access of the version of record (“gold OA”). Authors are also required to make articles available via self-archiving (“green OA”) regardless of OA publication.

However, so far there has not been any recommendations and mandatory requirements on research data by funding agencies, research institutions and scientific journals stand out here in China. It is very urgent and first priority to develop necessary legal frameworks that ensure the openness of research data for a various reasons.

4.2 Awareness

The importance of data, research data as well as research data management and openness has been aware among researchers and academic administrators. However, we observed a lack of organized activities and promotions on every aspects generally. It furthermore led directly the lack of institutional level strategic planning and implementation of RDM.

4.3 Collaboration

From the experiences of PKUL, it is very important to call attention and participation of related parties in, such as the researchers, academic departments, research support units, IT services, and the University administrators. Taking the rest of the university within our team.

4.4 Library's role: Engagement of Library

Despite all of the barriers and difficulties, there are tremendous needs for libraries services to foreground the RDM agenda, such as leading the institutional policies, raising data awareness, developing data curation capacity, teaching data literacy, and so on. The most important is that to ensure library's role in the research lifecycle, furthermore to foster a more open and rational scholarly communication ecosystem.

[1] Royal Society, Science as an Open Enterprise, 2012, https://royalsociety.org/~/media/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf

[2] http://www.re3data.org/

[3] http://ir.pku.edu.cn/?locale=en

[4] http://www.oaj.pku.edu.cn/OAJ/CN/OAJ/home.shtml

[5] http://scholar.pku.edu.cn/

[6] http://opendata.pku.edu.cn/

[7] Lyon L (2012) The informatics transform: Re-engineering libraries of the data decade. International Journal of Digital Curation 7(1): 126–138. DOI: 10.2218/ijdc.v7i1.220

[8] http://opendata.pku.edu.cn/dataverse/CFPS