National Diet Library Newsletter

No. 194, June 2014

Libraries in the "Big Data" era: Strategies and challenges in archiving and sharing research data

Reader Services and Collections Department
and
Digital Information Department

This is a translation of the article in Japanese of the same title in NDL Monthly Bulletin No. 639 (June 2014).

Contents

1. Introduction

With increased attention on using the "Big Data," what is going on the field of research data in distribution of academic information? So far libraries have provided only literature as the outcome of academic activities, not including data. What role should libraries have in the circulation of academic information now? How do libraries have to change their style for collecting and maintaining resources? For getting some answers to these questions, we planned an "International Symposium: Libraries in the "Big Data" era -- strategies and challenges in archiving and sharing research data."[1] We are introducing details of the symposium held on February 5, 2014, at the Tokyo Main Library, National Diet Library.

2. "Big Data" on scientific research

"Today there are more scientists who analyze and examine data than those who experiment by themselves," said Dr. Masaru Kitsuregawa (Director General, National Institute of Informatics; Professor, Institute of Industrial Science, University of Tokyo) in the discussion in the symposium. This would result from the situation in which many scientists come to use huge amounts of data generated from giant instruments in the research of what is called "Big Science."

The laboratory of Dr. Kitsuregawa has collected Internet resources from webpages, blogs and Twitter for 15 years. These large amounts of Internet resources provided a dissemination pattern of information at the time of the earthquake disaster. The laboratory also has supported building and maintaining the infrastructure of DIAS2 which conducts

the Environmental Information Integration Program. The flow volume of rivers can be predicted based on the large volume and variety of data of earth observation from DIAS.[2] "This prediction is useful to adjust the water volume discharged by dams to prevent flooding. Research data is getting to play various kinds of roles in our society as well as in the science area," said Dr. Kitsuregawa.

3. Importance of sharing research data

About the importance of sharing research data including "Big Data," Dr. Yasuhiro Murayama (Director of Integrated Science Data System Research Laboratory, National Institute of Information and Communications Technology (NICT); Visiting Professor, Research Institute for Sustainable Humanosphere, Kyoto University) said in his lecture that we needed archiving and sharing research data to secure the trustworthiness of scientific articles. Scientists judge the reliability of the results of each research from the related articles and data. Dr. Murayama indicated that a significant proportion of articles which have been cited many times cannot be reviewed on their research data based on the articles introduced in "Nature."[3] Our society needs data such as huge earthquakes or global warming to make political decisions. However we cannot get any data about past disasters and climate change if we did not cope with the data at that time. Therefore a system for archiving and sharing data is really important to secure the reliability and advancement of science and to promote the development of society, said Dr. Murayama.

Also Dr. Hiroki Sato (Professor, Interfaculty Initiative in Information Studies, Institute of Social Science, University of Tokyo), who introduced a data archive for the social science field in Japan as a case study, said that sharing data enabled researchers to discuss comparing another hypothesis based on the same data and for young researchers to access the data made at a huge cost.

4. Global activities for research data management

World Data System (WDS) founded by the International Council for Science (ICSU) in 2008 is an organization to promote universal and equitable access to data and to ensure long-term data stewardship for sharing data, holding up an open data policy. The International Program Office of WDS is located in NICT which Dr. Murayama belongs to. The International Program Office launches a project on "data publication" that enables them to publish data in cooperation with academic publishers, libraries and others.

RDA (Research Data Alliance) was set up in 2013 as a consortium to accelerate archiving and sharing research data. RDA has already twice held international conferences with working groups for research of data management.[4] Dr. Murayama said that specialists in library science play a crucial role in the conferences.

5. TIB's action for research data

In Germany, there are several national libraries including the German National Library of Science and Technology (TIB), whereas the NDL is the sole national library in Japan. The TIB is responsible for resources in the science and technology field and also has a role in the University Library Hanover.

Dr. Peter Loewe (Head of Development, TIB; Visiting Scientist [High Performance Computing], GFZ German Research Centre for Geosciences) said in the keynote lecture that the research cycle started with data generated by experiment; traceable information is created by analyzing and interpreting data; and information change into accessible knowledge by publishing (see figure). In addition to keeping this research cycle, it needs

to enhance the role of existing data centers and to use persistent and actionable identifiers for data. These days DOI (Digital Object Identifier) is used for scientific data as well as for articles or literature. DOI enables eternal access to data while general URLs cannot ensure access for the long term. The TIB became a DOI registration agency for primary data ahead of the rest of the world. This main function transited to a consortium, "DataCite," and the TIB continues to work as a part of the function at this point.

Dr. Loewe also introduced the RADAR (Research Data Repositorium) project operated by the TIB. The RADAR project started in FY2013 and will operate for up to three years, working with research institutions, academic organizations and libraries. The 75 % of research data stored on personal or institutional single hard drives can never be opened or used by other people. these data is near lost in a manner of speaking. The RADAR project aims to reduce this percentage and increase research data published and articles linked to research data by means of providing infrastructure to facilitate research data management.

6. State of data sharing and its issue

Next we shall look at case studies in Japan.

In the agricultural research area, Mr. Takuji Kiura (Senior Researcher, Agroinformatics Division, Agricultural Research Center, National Agriculture and Food Research Organization) introduced an actual case. "Most data are still in the hands of researchers, although a part of the statistical data or meteorological data is getting to be shared. ISO standards for agricultural machinery describe the guidelines of data processing. Therefore agricultural machines have a system for recording data automatically. However the data cannot be shared, because they have been accumulated only in computers of the companies who provide their machines or display terminals. There are some cloud services in the agricultural area, but the data in that cloud are incompatible with the software of other companies," said Mr. Kiura.

Concerning data sharing in the social science area, Dr. Sato introduced the activities of the Social Science Japan Data Archive (SSJDA) which collects, stores and provides micro-data generated from social survey, canvass, statistical survey and so on. "It takes much effort for us to mask and organize each micro-data item so that we can ensure access to data by general software while preventing a particular person or an institution from being identified. Therefore an urgent issue is to foster a ‘data librarian' who can curate data and conduct reference research through data," said Dr. Sato.

7. For maintaining high quality of sharing data

"It is an important point how to make easily usable data as well as how to collect and archive data," remarked Dr. Kitsuregawa. "A key point for promoting utilization of data is IT technology, especially data codes. For archiving high-quality data, we have to consider how to make standards of data codes as well as who will develop the infrastructure of data," observed Dr. Kitsuregawa.

"Related to this point, open data would have an effect on improving the quality of data," said Dr. Sato. "Additionally, now researchers maintain codebooks voluntarily because scientific research grants do not include the cost of making a codebook. Therefore, the data archives should have a structure for proper data curation."

In addition, Dr. Kitsuregawa pointed out that we need a system to give high evaluations to researchers who make data usable for many people, because it costs a lot to process and

provide data in a form that anyone can use.

8. A future role of libraries for data management

Dr. Loewe introduced the project "Radieschen (Rahmenbedingungen einer disziplinübergreifenden Forschungsdateninfrastruktur)"[5] which aims to consider the future scenarios for science in German infrastructure including libraries' perspective. The Radieshen project forecasts that libraries evolve into innovative, interlinked centers for information and competence, and data scientists will work in libraries in fields like curation, quality assurance or archiving, and also libraries will replace the scientific publishers of today.

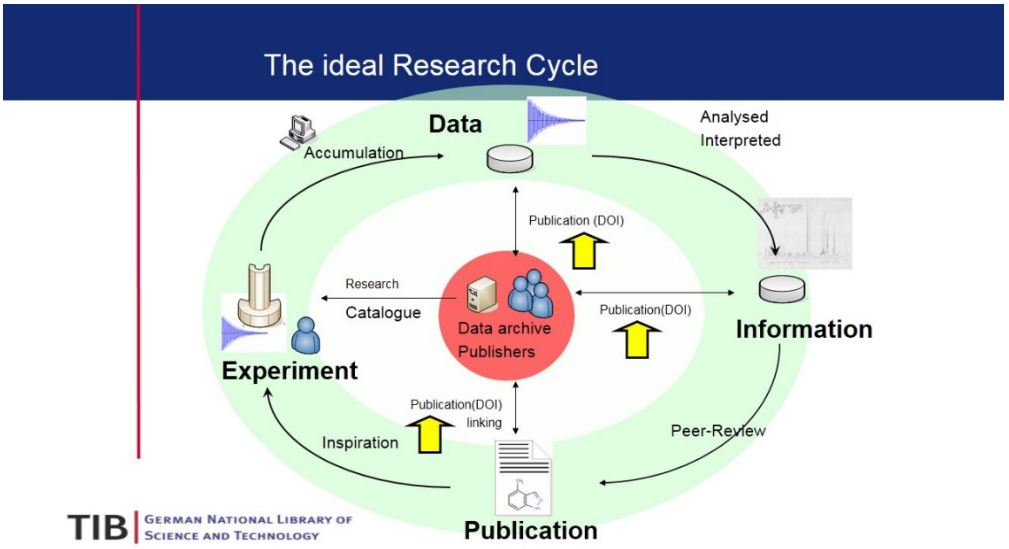
There are specialized institutions and specialists for data curation overseas. Dr. Murayama said that librarians could be specialists to organize and classify data and data codes and to conduct research reference by using data, although design of data codes needs specialists in the IT field and data analysis needs specialists in each academic field.

Dr. Kitsuregawa urged the need for building a national data center for research data, and demanded that in the future libraries should play the role of supporters for searching the location of data and data codes like libraries do for books.

Last Dr. Loewe insisted that as previously stated, some 75 % of data stored on single hard drives is being lost and this means that we are exposed to losing data again and again in the future. Therefore we should build a sustainable infrastructure for data on a world scale.

9. Conclusion

The NDL already holds some research data in collections stored on CDs or DVDs. We believe that libraries should expand their role by publishing metadata not only for articles but for any related reference data to identify them, as well, and should also develop a system to enable anyone to review the data. Such a system would create new knowledge. The NDL is now considering its role in the development of a system for sharing and archiving research data thorough the NDL Great East Japan Earthquake Archive.[6] Our challenge has just begun.



[1] Related documents include following web page:
http://www.ndl.go.jp/en/event/events/20140205sympo.html
[2] Data Integration and Analysis System:
http://www.editoria.u-tokyo.ac.jp/projects/dias/?locale=en
[3] C. Glenn Begley, Lee M. Ellis. Drug development: Raise standards for preclinical cancer research. Vol.483.(29 March 2012). pp.531–533. doi:10.1038/483531a.
[4] The 3rd conference of RDA was held in March 2014, after our symposium.
[5] Requirements for a multi-disciplinary research data infrastructure.
[6] http://kn.ndl.go.jp/node?language=en