IJoL International Journal of Librarianship

ISSN: 2474-3542 Journal homepage: http://journal.calaijol.org

The Development of Academic Data Services in Canada and China: Profiles of Data Services at Fudan University and the University of Windsor

Kristi Thompson and Shenqin Yin

Abstract:

The following notes provide a comparative perspective on the development of data services at Fudan University in China and at the University of Windsor in Canada. The two Universities are very different and operate in different environments. Fudan University is a major research university, one of the most respected academic institutions in China, a member of the elite C9 group (People's Daily, n.d.) of, and is located in Shanghai, one of the world's largest cities. The University of Windsor is a mid-size, mid-ranked comprehensive University in Canada, situated in Windsor, a smaller Canadian city on the U.S. border adjoining the much larger Detroit. Canada has one of the world's wealthiest economies; China's economy is still developing. The academic environments in the two countries are quite different. However, while the institutions may be dissimilar, the needs of their academic data users are not. We hope this article will help illustrate some of the broad commonalities in academic data services that stretch across borders, and will encourage institutions in disparate situations to recognize the feasibility of collaborating on common solutions.

To cite this article:

Thompson K., & Yin, S. (2017). The development of academic data services in Canada and China: Profiles of data services at Fudan University and the University of Windsor. *International Journal of Librarianship*, *2*(1), 73-78. https://doi.org/10.23974/ijol.2017.vol2.1.32

INTERNATIONAL JOURNAL OF LIBRARIANSHIP, 2(1), 73-78

ISSN: 2474-3542

The Development of Academic Data Services in Canada and China: Profiles of Data Services at Fudan University and the University of Windsor

Kristi Thompson, University of Windsor, Windsor, ON, Canada

Shenqin Yin, Fudan University, Shanghai, China

The following notes provide a comparative perspective on the development of data services at Fudan University in China and at the University of Windsor in Canada. The two Universities are very different and operate in different environments. Fudan University is a major research university, one of the most respected academic institutions in China, a member of the elite C9 group (People's Daily, n.d.), and is located in Shanghai, one of the world's largest cities. The University of Windsor is a mid-size, mid-ranked comprehensive University in Canada, situated in Windsor, a smaller Canadian city on the U.S. border adjoining the much larger Detroit. Canada has one of the world's wealthiest economies; China's economy is still developing. The academic environments in the two countries are quite different. However, while the institutions may be dissimilar, the needs of their academic data users are not. We hope this article will help illustrate some of the broad commonalities in academic data services that stretch across borders, and will encourage institutions in disparate situations to recognize the feasibility of collaborating on common solutions.

Background and Research

Fudan University currently has three separate institutions engaged with providing academic data services: the Library, the Social Science Data Centre, and the Institution for Big Data. Early data support included purchasing resources such as statistical yearbooks, economic and social statistical databases such as ChinaInfobank, CNKI (China Economic and Statistical Database), and CSMAR (China Financial and Economic Database). The library offered data usage training and support to faculty and students. The first data librarian was hired in 2016.

In 2011, the Social Science Data Center of Fudan University was established, and it started up the project of developing a social science data platform in April 2012. The development plan of the project proceeded in five phases: Survey and Research, Selection and Appraisal, Internationalization and Secondary Development, Policy Development, and Pilot Run. The initial phase included an investigation into 35 social science data centers around the world by conducting a literature review, doing website research and visiting selected locations. Sites visited included NORC at the University of Chicago, the Interuniversity Consortium for Political and Social Research (ICPSR) at the University of Michigan, CHRR (Center for Human Resource Research) of Ohio State University, and IQSS (The Institute for Quantitative Social Science) at Harvard University, in addition to centers in the UK and Australia. The working team conducted surveys on organization structure, funds, data collections, software platforms, metadata, policy, and data curation and management services.

The Institution for Big Data is the newest of the three support services and was founded in 2015 (Yang, 2015), together with the School of Data Science. It provides collection, storage and analysis services for “big data” collections, usually conceived of as datasets ranging in size from gigabytes to terabytes.

At the University of Windsor, the process of developing academic data services began early but proceeded in fits and starts. As with Fudan, at first data services largely consisted of purchasing individual data files for researchers. Data was purchased primarily from Statistics Canada, Canada’s national statistical agency on tapes, which were then loaded to local computer systems. Unlike at Fudan this was initially done by Information Technology Services (ITS), however in 1996 Statistics Canada launched the Data Liberation Initiative, an agreement under which universities paid a single fee to access the entire Statistics Canada data collection. This subscription model included support and training, and at many Canadian institutions, including Windsor, data support moved to the library (Boyko and Watkins, 2011). Data services remained basic, primarily data references and provision of data files, and were provided by a single librarian who also had a number of other responsibilities. In many institutions, this service was provided by the government information librarian; at Windsor the librarian was also responsible for business and economics.

Various stakeholders in the University expressed an interest in providing expanded data services, and in 2003 a group including several faculties and departments, ITS, the Library, and other offices on campus compiled a proposal for developing a Research and Instructional Data Centre. Authors of the report reviewed service models at major universities including Harvard, Stanford, Emory and, in Canada, McGill and Calgary. The initial proposal was quite ambitious and covered archival support as well as statistical and software consulting, and in 2005 a more modest proposal for a centre in the library was accepted by Library administration. The new proposal reviewed current needs and goals from a library perspective and suggested starting with two staff positions, a data librarian and a data analyst, with the idea that additional staff including technicians and archivists could added later. Services were to include support for statistical analysis and software in addition to the provision of data files from Statistics Canada and other sources. [1]

At the time there were relatively few available models providing integrated data, software and statistical consulting support from a single library service point. In 2006 the library filled the two positions with staff members hired from Data and Statistical Services at Princeton University, and the Academic Data Centre opened; the new staff developed a service based on the one deployed at Princeton (see Edelstein and Thompson, 2004). In 2012 a Geospatial analyst was hired as well.

[1] Thanks to Katharine Ball and Cathy Maskell for providing the authors with recollections and documentation of the early development of data services at Windsor

Infrastructure, Software Selection and Appraisal

In 2013, Fudan decided to test deploy four social science platforms, including Fedora commons, Dspace, Nesstar, and Dataverse. After six-months of testing and evaluation, it was decided that Dataverse, a platform developed at Harvard University, met their requirements fairly well, although it still lacked some features. However, as an open source system, Dataverse could be expanded and customized. Fudan selected Dataverse as their social science data platform to be localized and secondarily developed with customizations to meet local needs.

The Fudan working team implemented a Chinese interface for Dataverse, which is the first interface other than English to be available. The Fudan University Dataverse Network (http://dvn.fudan.edu.cn) is based on DVN3.3. It uses a specialized language search, providing a Chinese search engine with Chinese word segmentation, navigation with Chinese character index, and supports online analysis in Chinese without character disorder, which is unique to this particular Dataverse Repository Archive. Fudan has contributed their international code to the Dataverse Network so it will be available to other institutions, and is currently collaborating with Harvard's Dataverse Development Team to integrate their code with Dataverse 4.0. After about a year of technical work and policy development, the Fudan University Dataverse Repository pilot was launched in June 2014, and the public launch took place in December 2014. About 18 representatives of the media were present to take part in the ceremony and report on this achievement.

At Windsor, the initial focus was on providing in-person and classroom support to users of statistics and data. In 2010, the library was successful in its application to host a secure enclave for analysis of restricted government data, and opened the Statistics Canada Research Data Centre, Windsor branch. These services proved very popular (Thompson, 2012) and left little time for local technology development, and after 2012, budgetary constraints meant the library was in a hiring freeze. However, Windsor was (and still is) a member of the OCUL (Ontario Consortium of University Libraries), and these libraries pooled their resources with some additional government funding to provide a Nesstar implementation housed at the University of Toronto, Canada's largest university. Branded ODESI (Ontario Data Documentation, Extraction Service and Infrastructure), the service launched in 2008 with a collection largely consisting of Statistics Canada data. Further sources including large Canadian and International polling data collections were added later, and the collection continues to expand. Nesstar was selected in part because it implements the open DDI (Data Documentation Initiative) open standard for metadata, and an important part of the OCUL implementation was developing a Canadian Best Practices document that could then be shared with the rest of Canada.

Nesstar turned out to be highly suited for discovery of and access to curated, published data such as government surveys and commercial polls, but OCUL found it less useful for long-term preservation of institutional research data and unsuited for self-archiving by individual faculty members. This led to other options being explored, and like Fudan, in 2011 the Ontario university consortium launched an instance of Dataverse as a pilot (Barsky et al, 2017). Ontario university libraries, including Windsor, were quick to launch institutional Dataverses on the platform, and in 2012 the pilot concluded with Dataverse becoming a core service. As with Nesstar, the technical infrastructure is provided as a central service, but support for end users is provided locally by each institution. Windsor is also piloting the use of a new OCUL service, the Ontario Library Research Cloud, a secure storage network for big data files.

Data Curation Services

The Fudan Dataverse repository currently has datasets ranging from research findings to working papers, journal papers, and social science datasets. The majority of current datasets are survey and census data deposited by Fudan University affiliated researchers working in demography, economics, social science, geography, etc. File types include text, data files (i.e. dta, spss, xls, csv, etc.), image files (i.e. jpg), and GIS (Geographic Information Systems) data. In addition to original research data collected by Fudan University faculty, the archive also houses and distributes population census data, statistical yearbooks, and other government data. In addition to census and geospatial data, widely used collections include the Fudan Yangtze River Delta Social Transformation Survey and Fudan Energy Data Curation and Management.

Data needs to be findable to be useful, and the creation and maintenance of descriptive metadata is necessary to keep a data repository from turning into a data graveyard. The Fudan University Dataverse team provides many services to encourage researchers to move their data from local hard disks to be safely preserved in the Dataverse repository. Services include:

- Curation & Data Management Services: helping the researchers describe their data and other files, converting files to preservation-friendly formats, cleaning the original data to make it easier for others to reuse, etc.
- Branding & Customization: helping make customized banners for many of their researchers' Dataverses.
- Outreach: techniques include social media such as WeChat, a series of three videos prepared by faculty and students, and a regular newsletter.

Outreach has been very successful and the repository now houses over 5,700 projects collected from 1,319 researchers that are affiliated with the university.

The Fudan University Dataverse team has taken a leadership role in promoting data management using the Dataverse platform in China. They have conducted two Chinese domestic seminars, reported on Dataverse and data curation platforms at 14 nationwide academic conferences between 2013-2017, and are founding members of the "China Academic Library Research Data Management Implementation Group"[2] formed to promote the development of Chinese domestic Research Data Management.

Windsor and other Ontario universities offer data management services that are broadly similar to those at Fudan: data documentation, file conversion and cleaning, Dataverse branding and customization, and outreach through web, email and social media such as Twitter. While

[2] The nine university members of this group include Peking University, Tsinghua University, Zhejiang University, Wuhan University, Beijing Institute of Technology, Shanghai Jiaotong University, Shanghai International Studies University, Tongji University, and Fudan University

uptake of Dataverse by Windsor researchers has been relatively slow thus far, Canada's funding agencies are starting to insist that data collected as part of funded research needs to be archived and made accessible. This mandate promises to ramp up demand considerably. With Dataverse hosting already in place, Windsor should be well situated to meet this demand and the data team has collaborated with other campus organizations to organize a series of presentations and workshops to promote data deposit.

The Windsor data team contributes to the Portage Network, probably Canada's closest equivalent to China's Research Data Management Implementation Group. Windsor's data librarian also is leading a team of data librarians in a project to recover, document and publish a major collection of government health survey files (Cooper, Thompson and Trimble, 2017), using Dataverse as a staging repository. The team intends to publish the data openly on both Nesstar and Dataverse once it is in useable form, and is simultaneously developing a guide to data rescue to share with the data community. This group project encapsulates much of what is happening today in data librarianship – a person at one institution wanted to provide some files to local researchers, and found that the best way to do this was to enlist a larger group to do the work and share the results.

Academic data service is fundamentally the process of making data as accessible as possible to as wide a range of researchers as possible. So it is perhaps not surprising that the common element that most unites the experiences of these two very separate institutions is that of sharing. Both institutions have focused on working with open standards and open source software (particularly Dataverse and DDI), both have been involved with contributing code, practices, documentation or standards to their communities, and both have found that the best way to advance local goals is to work formally and informally with alliances of other institutions. In a rapidly evolving area such as data, often the best way for an individual institution to make progress is to help the entire community move forward as well.

The authors of this article are hoping to continue their investigation into and knowledge-sharing around the development of data services in China and Canada in the fall of 2018, when Fudan University has offered to host two librarians from the University of Windsor as visiting scholars. Look for the results of their research to be published in a future issue of the International Journal of Librarianship!

References

Barsky, E., Laliberte, L., Leahey, A., & Trimble, L. (2017). Collaborative research data curation services: A view from Canada. In L. R. Johnston (Ed.), *Curating Research Data, Volume One: Practical Strategies for Your Digital Repository* (pp. 79-101). Chicago: Association of College and Research Libraries.

Boyko, E., & Watkins, W. (2011, November). The Canadian data liberation initiative: An idea worth considering? Retrieved from http://www.ihsn.org/sites/default/files/resources/IHSN-WP006.pdf

Cooper, A., Thompson, K., & Trimble, L. (2017, May 24). Documenting data rescue. The Ontario data community data rescue group and the data rescue & curation guide for data rescuers. Presented at *IASSIST 2017*. Retrieved from https://works.bepress.com/kristi_thompson/6/

Edelstein, D., & Thompson, K. (2004). A reference model for providing statistical consulting services in an academic library setting. *IASSIST Quarterly*, *28*(2), 35-38. Retrieved from http://scholar.uwindsor.ca/leddylibrarypub/9/

People's Daily Online (English ed.). (n.d.). China's Ivy League: C9 League. Retrieved from http://en.people.cn/203691/7822275.html

Thompson, K. (2012, June 6). "Bringing them in: What 5 years of data can tell us about growing a data service." Presented at *IASSIST 2012*. Retrieved from http://www.iassistdata.org/downloads/2012/2012_pk1_thompson.pdf

Yang, M. (2015, October 9). Fudan University launches big data institute. *Shanghai Daily (Online ed.)*. Retrieved from http://www.shanghaidaily.com/metro/Fudan-University-launches-Big-Data-Institute/shdaily.shtml

About the authors

Kristi Thompson is the Data Librarian at University of Windsor with additional responsibilities in Information Services and Systems, and the guest editor of this issue of the International Journal of Librarianship. She previously co-edited *Databrarianship: The Academic Data Librarian in Theory and Practice for ACRL* with Lynda Kellam, and has published and spoken widely on academic data librarianship.

Shenqin Yin is an assistant Director of the Social Science Data Research Center of Fudan University, a member of Dataverse Advisory Team. She also supervises graduate library students. Her research areas include research data management, open government data and public governance, especially population data, youth development data and user information behavior data.