Data Management in the United States and Canada: Academic Libraries’ Contribution

Mayu Ishida
MLIS Student
University of British Columbia

For LIBR 559L: Issues in Scholarly Communications and Publishing, 2011 Summer Term 1

Introduction

It is now possible to amass and preserve an unprecedented amount of data in research efforts because of advancing information technologies. Improvements in data collection devices in addition to the increasing capacity and decreasing cost of data storage are contributing to what is called “data deluge” (Hey and Trefethen). The amount of research data is beyond what can be managed manually. Therefore, to exploit large volumes of data for research and learning, it is necessary to establish repositories and an information infrastructure that allows scholars to preserve and access research data with tools for search, analysis, and visualization.

The data-centric approach to research is becoming prominent in the sciences as well as the social sciences and the humanities (Borgman, Scholarship in the Digital Age 6). Scientists are using linear accelerators, sensor networks, and satellites to generate large datasets. Social scientists are studying volumes of statistical data from governments and online surveys. Humanities researchers are mining large bodies of text, images, video, and audio in digital formats. By working with big data, researchers in a variety of disciplines are discovering patterns and correlations that have not been conceived before. This focus on research data is also changing the culture of scholarly communication. For example, open access journals published by the Public Library of Science request that researchers should make available the datasets on which their manuscripts are based (Nelson 163). As a rising number of journals are published in a digital format, it is now feasible to include in an article links to models and datasets available online or to package all relevant scholarly artifacts in one digital object.

As research increasingly revolves around data, data management becomes an integral part of the scholarly communication life cycle (“Scholarly Communication”) and requires further study and improvement. In this literature review on the data sharing trend in the United States and Canada, we analyze data management issues and proposed solutions including the role of academic libraries in the preservation of faculty’s research datasets and scholarly legacies.

Why do we want to share data?

The call to preserve research data and make them accessible to the scholarly community derives from the tradition of “open science”, which dates back to the fourth century (Borgman
Data Management in the United States and Canada: Academic Libraries’ Contribution
Mayu Ishida

35-36). It is a notion that the whole of society benefits from the wide and rapid exchange of ideas within the scholarly community, and the rigorous validation of new knowledge by scholars. Furthermore, emerging modes of scholarly communication such as open access literature (“Why Open Access?”) and knowledge commons (“Open Knowledge Commons”) stem from the notion of open science.

Data sharing is also promoted by government funding agencies in the United States and Canada. They argue that the results, including data, of publicly funded research projects should be made available to taxpayers. They also hope to increase the return on investment by fostering research efforts based on existent publicly-funded data. In 2003, National Institutes of Health (NIH) in the United States published the final NIH Statement on Sharing Research Data (“Final NIH Statement on Sharing Research Data”) to confirm its support of data sharing (“NIH Data Sharing Information”). Another American public funding agency the National Science Foundation (NSF) went one step further and introduced data management plan requirements in January 2011 (“NSF Dissemination and Sharing of Research Results”). Research proposals submitted to NSF must include a data management plan that the research proposal will comply with NSF’s data sharing policy.

In Canada, data sharing is part of the guidelines for access to research results endorsed by three public funding agencies: the Canadian Institutes of Health Research (CIHR), the Social Sciences and Humanities Research Council (SSHRC), and the Natural Sciences and Engineering Research Council (NSERC) (“Science.gc.ca Access to Research Results: Guiding Principles”). SSHRC states that SSHRC-funded data should be made accessible to the public “within two years of the completion of the research project for which the data was collected” (“SSHRC Research Data Archiving Policy”). CIHR obliges researchers to “deposit bioinformatics, atomic, and molecular coordinate data into the appropriate public database (e.g., gene sequences deposited in GenBank) immediately upon publication of research results” and to “retain original data sets for a minimum of five years” (“CIHR Policy on Access to Research Outputs”).

Data management challenges

1) Lack of data and metadata standards

Guidelines and mandates for data sharing aim to raise awareness of data management in the scholarly community and reduce volumes of hidden and dispersed datasets that are stored and often forgotten on researchers’ personal computers or local servers. However, the guidelines only represent an educational and promotional approach to the implementation of the data sharing policy and do not have the power to enforce the policy on researchers. Although mandatory, the NSF requirements for data management plans specify neither standards for data and metadata nor policies for access, re-use, and re-distribution; rather, they ask researchers to define such standards and policies in a supplementary document to a research proposal (“NSF Grant Proposal Guide Chapter II.c.2.j”).

The lack of standards for data and metadata is due to a multitude of definitions of data arising from individual disciplines. Data can take many forms and may mean many things, depending on context and interpretation (Borgman, Scholarship in the Digital Age 120-121). For example, rock samples are data to geologists while recorded interviews are data to sociologists. An
increasing number of datasets are born digital and stored electronically such as online survey forms and signals from remote sensors. Data are also collected for many purposes by many collection methods (Borgman, “Research Data: Who Will Share What, with Whom, When, and Why?” 3). Due to a variety of formats, purposes, and methods in data collection, it is implausible to set a universal definition of data in a data sharing policy or mandate. However, as the awareness of data management increases in the scholarly community, each discipline may discuss and eventually establish feasible standards for data and metadata standards. NSF may also monitor data and metadata practices in individual disciplines through submitted data management plans and iteratively refine its requirements for data and metadata standards. Compliance with a data sharing policy will become less ambiguous as data and metadata standards emerge in individual disciplines and these standards are incorporated in a data sharing policy.

2) Data repositories and information infrastructure still in development

While introducing data sharing policies and mandates, the public funding agencies in US and Canada do not themselves offer data repositories with long-term data management services. SSHRC instructs researchers to deposit their datasets at institutional repositories at Canadian postsecondary schools who are members of the Canadian Association of Public Data Users (CAPDU) (“SSHRC Research Data Archiving Policy”). CIHR also refers researchers to Canadian postsecondary schools as well as data repositories established for health science disciplines such as GenBank (“CIHR Policy on Access to Research Outputs”). However, these data repositories do not necessarily have long-term funding sources or sustainable business models to guarantee perpetual access to deposited datasets. They may also pose additional policies and mandates. An institutional repository may be underfunded or understaffed if it is not a high priority to the university library (Salo), and hence a CAPDU member university may not be ready to receive large volumes of datasets.

It is ideal that the government should commit long-term funding to a national data repository where researchers can deposit publicly funded research data. Such a centralized data repository can provide easy access for the public and should accept datasets that are too big to store locally at institutional repositories. Alternatively, existing repositories may be networked and incorporated into an expansive information infrastructure. Such infrastructure initiatives are led by governments in Europe and the United States: DRIVER, funded by the European Commission (“DRIVER Search Portal”), and Cyberinfrastructure, funded by NSF (“NSF Cyberinfrastructure: A Grand Convergence”).

Moreover, legal complications may influence the development of repositories and infrastructure. If a dataset is generated in inter-institutional or international research collaboration, it may be difficult to ascertain copyright ownership or licencing for the dataset, and to decide in which repository the dataset should be deposited. International agreements should be reached on the intellectual property issues in data sharing, and repositories should be designated where multinational research efforts could deposit their datasets without legal impediments. Legal concerns may also affect repository design and capability. Due to concerns about privacy and patient confidentiality, a data repository needs to be capable of granting different degrees of access to data regarding human subjects. For example, users may freely access aggregate clinical data where it is impossible to identify individuals whereas the users may have restricted access to clinical data associated with unique individuals (Nelson 170). Such sensitive data may also indicate security issues; if researchers are to deposit their datasets in repositories outside their institutions, it may become unclear who is responsible for safeguarding the datasets.
Whether a centralized or distributed model is implemented for data management, it should be economically, legally, and technically sustainable, and there should be data repositories ready to receive publicly funded data as well as an infrastructure to provide access to such data.

3) Reluctance to share data

Besides establishing standards and infrastructure for data management, the culture of data sharing should be fostered in the scholarly community in support of data-centric research endeavours. Arguments for data sharing highlighted in policies and mandates do not explicitly relate to the incentives of researchers to share their data (Borgman, “Research Data: Who Will Share What, with Whom, When, and Why?” 7); rather, the arguments describe public goods, scientific advancement, and benefits to data users who wish to reproduce research results and ask new questions on existing data. To implement data policies, it is necessary to create compelling incentives for data producers to share their datasets.

One way to reward researchers for sharing their datasets is to recognize their shared datasets as contributions to the scholarly community, as units of scholarly communication, so that the shared datasets will be considered in tenure and promotion. According to William Michener, the director of e-science initiatives for University Libraries at the University of New Mexico, many repositories will remain empty unless the scholarly culture changes to consider data as valuable as publications (Nelson 163). In some disciplines, data sharing is already contributing to the reputation of data producers; in the Human Genome Project, a shared dataset itself is regarded more highly than any single journal article based on the dataset (Witt 194). Assigning a Digital Object Identifier to a shared dataset makes it easy for other researchers to cite the shared dataset and also makes it possible to measure the impact factor of the shared dataset.

However, data citation standards are yet to be established. The comparison of a sample of 20 style guides including the publication manual of the American Psychological Association, the MLA style manual and guide to scholarly publishing, and the Chicago manual of style, showed that the style guides do not indicate a consistent approach to the citation of research data (Newton, Mooney, and Witt). To address this issue, the International Association for Social Science Information Services and Technology created the Special Interest Group on Data Citation in 2010 (“IASSIST Special Interest Group on Data Citation”). It is proposed that we apply to data citation the source-tracking technology used at a music website ccMixer (Nelson 143). The technology automatically updates credit records for the samples ccMixer users add to new tracks in their mixes. A similar system that records a link to each repurposed dataset would be useful particularly to researchers who base their research on multiple existing datasets.

Researchers are concerned that data they share may be compromised (Borgman, “Research Data: Who Will Share What, with Whom, When, and Why?” 9), so guidelines against data misuse should be part of a data sharing policy. Moreover, researchers hesitate to share their data before they finish analyzing the data and produce publications based on the analysis (Borgman, Scholarship in the Digital Age 8). Researchers who share their dataset should have the right to publish the first article based on the shared dataset. However, it is recommended that the article be limited to the global analysis of the dataset, and that the embargo period be less than one year (“Prepublication Data Sharing” 170). Before data generation, researchers could publish a citable statement called a “marker paper” (“Prepublication Data Sharing” 170). It describes the researchers’ intentions for the resulting data and the details of the data generation
process such as the experimental design, the data and metadata standards, the quality-control system, and the expected timelines.

What can academic libraries do in data management?

1) Establish an interdisciplinary task force for data management

Research data management requires expertise from library science, archival science, information technology as well as knowledge of individual disciplines. Hence, a data management group should consist of experts from these fields so that they supplement each other’s knowledge and skill sets on data management issues. Archival concepts such as provenance, appraisal, and preservation are particularly crucial to data management. An academic library is a primary candidate to provide data management services since its workforce can collectively offer the required skill sets. Academic libraries may also have relevant experiences in digitization projects and digital library initiatives (Witt 195).

2) Learn from other institutions with data management experience

Academic libraries can play an important role in data management at a local institutional level, and some academic libraries are already offering data management services to faculty. The first step for an academic library new to data management is to learn from other academic and research institutions with data management experience.

Data sharing has been a priority to academic and research institutions in the United Kingdom since 1993 when the seven research councils established the Joint Information Systems Committee (JISC) (“Data’s Shameful Neglect”). JISC has also helped start the Digital Curation Centre in the United Kingdom, a national institution for the promotion of data sharing. The centre calls itself “the UK’s leading hub of expertise in curating digital research data”, and offers training programs and data management resources at its website.

The Association of Research Libraries (ARL) that consists of academic and research libraries in the United States and Canada conducted a survey on data support services at its member institutions and discovered that “research institutions are quickly rising to meet the challenges of managing data, especially in light of the anticipated federal government requirements for data plans as part of grant proposals” (Soehner, Steeves, and Ward 9). For further interviews and case studies, the survey contacted six respondents: Purdue University Libraries, the University of California, San Diego, Cornell University, Johns Hopkins University, the University of Illinois at Chicago, and the Massachusetts Institute of Technology (Soehner, Steeves, and Ward 7). They are considered to be leading academic libraries in data management, and it is worthwhile to examine their data management practices detailed in the ARL survey. The Canadian Association of Research Libraries (CARL) published a review of research data services by academic libraries in Canada and around the world. The review featured virtual research environment Islandora developed by the University of Prince Edward Island as well as discovery and access tool ODESI offered by the Ontario Council of University Libraries (Shearer and Argaez).
3) Identify skills transferable from a traditional library role and skills specific to data collection

By comparing the process of institutional data collection to that of traditional collections development, Newton, Miller, and Bracke identify skills transferable from the traditional librarian role as well as additional skills and knowledge required for the data-collating librarian role. As in the traditional collections development, the data collections development requires the identification, evaluation, negotiation, acquisition, and preparation of materials – in this case, research datasets (Newton, Miller, and Bracke 64). As in traditional collections development, the data-collating librarian should ensure that the resulting dataset collections meet the mission of the institutional repository and library by proposing and adhering to a data collection policy (e.g., collection criteria for a data repository prototype at Purdue University (Newton, Miller, and Bracke 67)).

Despite the similarities between traditional collections development and data collections development, the data-collating librarian needs to possess the following skills to build an institutional data repository (Newton, Miller, and Bracke 62-63):

- Ability to articulate the advantages of depositing a dataset in the data repository
- Fluency in the data repository technology
- Awareness of research activities on campus

When presenting a case for data sharing, the data-collating librarian needs to address the faculty's needs such as attribution, embargo, and interoperability. Although not every librarian will have experience in developing the data repository technology, the data-collating librarian still requires technical knowledge to answer the faculty's questions about the capability of the data repository. The data-collating librarian should also be able to collaborate with the data producer at multiple points throughout the data life cycle in order to assess the data producer's needs and inform the development of the data repository.

4) Deploy existing library services for the promotion of data management

The existing library services can contribute to the development and promotion of institutional data collections (Witt 195). The cataloguing librarians can offer their expertise in classification and description, especially metadata standards. The public services librarians can provide access points to the data collections through reference and instruction services. In partnership with individual departments, the data librarians can teach students data literacy along with statistical literacy (Kellam) and encourage them to use the institutional data collections in their research. It is argued that "data management should be woven into every course in science, as one of the foundations of knowledge" (“Data’s Shameful Neglect”).

The subject librarians should be partners in data collections development since they can cultivate collaborative relationship with individual departments and can contribute their domain knowledge required in data management. Ongoing collaboration between the library and the faculty is essential to data collections development (Newton, Miller, and Bracke 63). By engaging in the faculty’s research activities, the library can locate a candidate dataset for the data collections and approach the data producers with a data management plan regarding the specifics of their research project. Based on the degree of trust they have in the library, the data producers will decide whether to share their dataset through the institutional data repository.
5) Address the concerns and needs of data producers

Researchers are wondering what rights they would have to relinquish when depositing their data in repositories and how they could track and format all their data when time and money are scarce (Nelson 160). The library can alleviate these concerns by offering its expertise in copyright, licensing, and knowledge organization. The library may communicate with funding agencies and ensure that the data management plans it prepares with the faculty meet the requirements of the funding agencies’ data policies. The library can also host a campus forum to raise awareness of data sharing in the university community and to promote the library’s data management services.

The institutional repository can be marketed as a content management system where the faculty could store their research data alongside related publications although it is difficult to capture elaborate relationships between data and publications in current metadata practices (Borgman, “Research Data: Who Will Share What, with Whom, When, and Why?” 13). The faculty may place in the institutional repository data from a research project that is put on hold until funding is restored to the project (e.g., the Advanced Life Support – NASA Specialized Center of Research and Training (Carlson, Ramsey, and Kotterman)).

6) Publish data management case studies

Academic libraries can help improve their data management services by reporting to each other ideas and practices for data sharing. They may frame their data management initiatives as empirical research efforts and publish the outcomes of their data management experiments in library and information science journals. They may also contribute to surveys and reports on data initiatives by library associations such as ARL and CARL. The experience of developing an institutional data repository benefits the analysis of data practices, and vice versa (Wallis et al. 338). Academic libraries entering the field of data management may extrapolate strategies appropriate to their institutions from the practices of other academic and research institutions.

Conclusion

All stakeholders need to prioritize, invest, and collaborate in order to successfully integrate the concept of data sharing in scholarly communication. The government should initiate the building of an information infrastructure that can overcome the economic, legal, and technical barriers to data sharing. The funding agencies and journals should work together to set consistent data sharing policies. Academic and research institutions should make data sharing part of their strategic plans to preserve, promote, and facilitate research through their libraries’ data management services. The data producers, data users, and scholarly societies should discuss feasible data and metadata standards and promote effective data use and management practices in the research community. Together the stakeholders could foster “a scientific culture that encourages transparent and explicit cooperation” (Nelson 169) in data sharing as a part of scholarly communication.

As the demand for assistance in data management increases, it becomes imperative to train librarians and information professionals in the field of data management (Borgman, “Research Data: Who Will Share What, with Whom, When, and Why?” 13). Library schools in the United States have now started offering curricula on data management and digital curation such as:
• Specialization in Data Curation at the University of Illinois Urbana-Champaign ("Master of Science: Specialization in Data Curation")
• DigCCurr (pronounced dij-seeker) at the University of North Carolina Chapel Hill ("DigCCurr Carolina Digital Curation Curriculum Project")
• DigIn at the University of Arizona (Fulton, Botticelli, and Bradley). DigCCurr differs from the other curricula in that it is at the doctoral level and aims to produce future faculty who will conduct research and instruct in digital curation ("About DigCCurr II").

As data sharing becomes prevalent in the scholarly community, it will facilitate research collaboration across institutions, nations, and disciplines, and will enable the scholarly community to tackle global issues such as climate change and poverty. Although data sharing initiatives in the United States and Canada are still in the early stage of establishing standards and infrastructure, they are gathering momentum and shifting the culture of scholarly communication.
Works Cited


“IASSIST Special Interest Group on Data Citation.” Web. 10 June 2011.


“Master of Science: Specialization in Data Curation.” Graduate School of Library and Information Science, the University of Illinois at Urbana-Champaign. Web. 12 June 2011.


Newton, Mark, Hailey Mooney, and Michael Witt. “A Description of Data Citation Instructions in Style Guides.” 7 Dec 2010 : n. pag. Print.

“Open Knowledge Commons.” *Open Knowledge Commons*. Web. 5 June 2011.


