Survey of Canadian and International Data Management Initiatives

By Diego Argáez and Kathleen Shearer

on behalf of the CARL Data Management Working Group

(Working paper)

April 28, 2008

Introduction

Today, research is increasingly data intensive and researchers rely on access to large and complex data sets. Unprecedented access to data and digital tools is increasing the efficiency of research and enabling new discoveries. “Ensuring research data are easily accessible, so that they can be used as often and as widely as possible, is a matter of sound stewardship of public resources.” (OECD 2004)

Indeed, the effectiveness of our research enterprise depends on how this data is managed. In the past, data archiving has been managed at the level of the discipline, community, or individual researcher. However, given the substantial cost of creating data collections and the complexity of managing and preserving them, this approach is no longer considered adequate.

Research libraries have a role to play in this emerging data-intensive environment. A 2007 CARL survey found that most CARL members are interested in managing research data, but few have a formal data archiving policy. CARL has formed a Research Data Management Working Group to assist members in collecting, organizing, preserving and providing access to the research data and to formulate a cooperative approach for CARL.

The purpose of this report is to provide an overview of the types of data management activities being undertaken in Canada and internationally. This review documents the various options available for libraries, and will pave the way for a more detailed investigation by the Working Group of the potential roles for libraries.

Scope of Review

In the National Science Foundation report, *Long-lived Digital Data Collections* (http://www.nsf.gov/pubs/2005/nsb0540/), three types of data collections are identified (NSF, pg. 20-21):

> Research data collections are the products of one or more focused research projects and typically contain data that are subject to limited processing or curation. They may or may not conform to community standards, such as standards for file formats, metadata structure, and content access policies. These collections are supported by relatively small budgets, often through research grants funding a specific project.
>
> Resource or community data collections serve a single science or engineering community. These digital collections often establish community-level standards either by selecting from among preexisting standards or by bringing the community together to develop new standards where they are absent or inadequate. The budgets for resource or community data collections are intermediate in size and generally are provided through direct funding from agencies.
>
> Reference data collections are intended to serve large segments of the scientific and education community. Characteristic features of this category of digital collections are a broad scope and a diverse set of user communities including scientists, students, and educators from a wide variety of disciplinary, institutional, and geographical settings. In these circumstances, conformance to robust, well-established, and comprehensive standards is essential, and the selection of standards by reference collections often has the effect of creating a universal standard.

The purpose of this review was to identify best practices and possible models for long-term archiving and access to data in libraries. Thus, the review focused on identifying different models that fall into the later two categories. As well, for the most part, data centres or services that did not host data, but exclusively provide access to data were omitted from this review (with the exception of ODESI).

There are thousands of data collections around the world. The intent of this review was not to provide an exhaustive list of data archiving initiatives, but instead to identify the various models for managing research data currently being employed. Because this is a rapidly evolving area, a list of 'demonstrator' and 'in development' projects that were encountered during the review has also been included in the Appendix of the report.

Methodology

The review included a scan of both Canadian and international initiatives, identified through the existing literature and Internet searches. For the review of Canadian initiatives, a survey was also sent out to the CARL directors to identify any other relevant initiatives.

Wherever possible, the following information about each initiative was documented:

- Name of initiative
- Organization and country
- Types of data collected
- Scope of data collected
- Discipline
- Metadata
- Technologies
- Staffing
- Funding
- Curation and preservation
- Data acquisition
- Access
- Services
- Organizational models

Results

Digital data comes in many forms and exists along a continuum of analysis (from raw data to highly processed data). Data management activities differ greatly according to data type and the mandate of the organization involved. These differences should be kept in mind when looking at the variety of data management activities in use.

Types of data: All types of data were included in the review including numbers, images, video or audio streams, software and software versioning information, algorithms, equations, animations, or models/simulations. The types of data collected by an individual data archive are very discipline specific. While there are a few archives that collect a large variety of data types within a single archive (DANS in the Netherlands, Arts and Humanities Data Service in UK), for the most part, data archives have been specifically designed to collect and manage a restrictive number of data types in a given discipline.

Scope of data: The scope of data archives differ according to the archive's mandate and discipline. Generally speaking, national archives in the social sciences tended to collect data related to that individual nation, while scientific archives are more likely to collect data from outside national borders. Some archives collect raw data, others focus on post-analysis data only, and some collect both.

Disciplines: Data archiving initiatives were identified in four broad disciplinary categories: (1) Arts and Humanities, (2) Social Sciences, (3) Health Sciences, and (4) Natural Sciences and Engineering.

Metadata: Most repositories adhere to discipline specific metadata standards for descriptive metadata. In the social sciences, data centres regularly employ the Data Documentation Initiative (DDI) standard. No general model for the representation of scientific study metadata exists. However, there is a nascent movement to develop a common set of metadata so that datasets from different scientific disciplines are interoperable.

Technologies: The technologies in use varied. Many archives are using technologies developed in-house, especially for data analysis purposes. There are also large technology companies that have also been closely involved in the development of repositories and tools (ie. Microsoft).

Curation and Preservation: The focus of many data archives is to curate data in order to facilitate greater analysis. In many cases it is not clear what preservation activities are being undertaken by the archive. Quality control, data storage and backup, and descriptive metadata are the most commonly cited practices. In terms of preservation, many archives refer to best practices, such as the Open Archival Information System reference model, but it is unclear to what degree these types of rigorous standards are being adhered to within the repository. Costs and reproducibility of data are both considerations for preservation.

Staffing: Staffing at the archives reviewed ranged from 5 full time employees to over 50 full time employees, depending on the size of the archive.

Data acquisition: Data is acquired in a number of ways:

- In the Natural Sciences, data is generally acquired directly from scientific instruments such as telescopes, although some larger archives also welcome researcher deposited data
- In the UK, the government-run data centres do require researcher deposit
- In the Arts and Social Sciences, it is usually a combination of researcher deposit and acquisition (sometimes by payment) of external data sets from other organizations
- In the Health Sciences, where there is a tradition of journals requiring data deposit before publishing related articles and the responsibility often lies with the researcher to deposit the data, are much higher rates of researcher deposit

Funding: For the most part, the data archives reviewed are funded by one, or a combination of the following:

- One or more granting agencies
- A government agency
- A university department
- Licensing revenues

Because long-lived data collections are international in scope and span beyond the life of a given research project, procuring sustainable funding for long-term preservation remains a challenge.

Access: Many initiatives offer access to at least some data sets for free. The various models employed are as follows:

- Institutional membership or license required OR pay per view access for non-members.
- Free access for research or education purposes AND Paid access for commercial purposes
- Free access to some data; pay per cost for other data Or no access to other data outside of community.
- Access to community members only. No external access.

Most data archives require users to sign licensing agreements, which govern things such as confidentiality, acknowledgements, and redistribution of data. In terms of personal data, all archives that dealt with this type of data had mechanisms in place to anonymize the data before it was made available to others. In some cases, data that could not be appropriately anonymized was simply not available. A number of archives also enabled researchers to restrict access to their data for a given period of time after deposit, to allow them to publish their findings before the data was released.

Services offered by archives vary. They can be grouped into four categories:
Deposit services-providing researchers with help structuring, tagging and depositing data
Discovery services-search and retrieval. These services are mainly technology driven
Data analysis services-such as visualization, grouping. These are also mainly technology driven, and there is a strong emphasis on these types of services in the science
Expert guidance and support services-these encompass assisting data creators with developing data management plans, data preparation, and metadata tagging.

Organizational models: Four types of organizational models represented:

- Stand-alone data centre
- Centralized resource centre, and part of a distributed network
- Series of test beds supported by a coordinating centre
- Distributed networks of interoperable data centres

Other Trends of Note

- Big commercial players are becoming interested in collecting and storing research data (Microsoft, Google)
- Organizational models are moving away from a centralized stand-alone model towards distributed networks of interoperable data centres
- New functionalities will require interoperability between different disciplines and different types of data
- The UK, Netherlands, and EU are examining the role of institutional repositories for data collection and management
- Metadata is recognized as being key for long-term data access and usability (including documenting the research process, not just the data itself)
- The suite of data analysis tools are growing and becoming more complex. (ie. the use of GIS in many fields).
- The focus is on curating data for reuse, not for long-term preservation.
- Data creators need support in structuring and tagging their data so that it can be understood by others.
- Open access to data is becoming more common.

Roles for Libraries

It is clear that if libraries must work closely with the research community. Any services provided must be driven by and reflect the needs of the research communities they seek to support. Based on this review, a number of preliminary areas are identified for which research libraries could contribute to data management.

1. Data repositories - building and managing institutional data repositories, in order to eventually achieve a transparent system of grid-like libraries and library data services supporting data science and curation

2. Metadata development - Metadata are an essential component of research data. Research libraries can lead development of standardized, ontologically rich automated metadata for such datasets. Developing and managing metadata already are established tasks in the library community - although current practices will not handle the scale envisioned.

3. Interoperability - Libraries have a long history of working together. Access and cross-domain usage of distributed data collections will require application of uniform methods of description when the data are created.

4. Support and Training - Support for personal information management: as datasets and associated information becomes increasingly portable.

5. Preservation - Existing practices within data repositories focus on the immediate use of data, not use over the long-term. There is a need to build capacity in terms of preservation activities.

A number of reports discussing the roles for libraries in terms of data management have been published recently, in particular, a 2006 ARL report, *Agenda for Developing E-Science in Research Libraries* (http://www.arl.org/bm~doc/ARL_EScience_final.pdf) and the UK report, *Dealing with Data: Roles, Rights, Responsibilities and Relationships* (http://connect.educause.edu/Library/Abstract/DealingwithdataRolesright/44533). These reports provide important context for any further discussions of the findings outlined here. In the coming months, the CARL Data Management Working Group will further develop the roles identified here and make recommendations to CARL members for library engagement in this area.

Summary findings for **25** current **Canadian** data management initiatives in terms of ***access***, ***services*** offered, ***data acquisition***, ***organizational models***, ***preservation*** and ***funding***.

Access	**# of data initiatives**
Consortium model, access for staff and students from participating member institutions	**2**
Reserved for University (one institution) staff and students	**2**
Free, public use microdata files and aggregate data. Datasets held back until they have been anonymized.	**1**
Reserved for researchers who go through application and approval process	**5**
Free	**12**
Some free content, some accessed for a fee	**2**
"At cost" per requested data set	**1**
Two levels of access: free to staff and students at participating institutions paying membership dues; researchers who go through application and approval process	**1**

Services	**# of data initiatives**
Search and retrieval	**11**
Online analysis tools	**6**
Online analysis tools supported by Nesstar software	**3**
Consulting, educational services on data management practices, staff available to offer assistance, etc.	**5**
Data deposit	**4**
Data preparation, anonymization, formatting, etc.	**4**
Online help files and tutorials	**1**
Various data reports/publications or supporting documents that provide context for the datasets	**3**
Custom data reports for a fee	**2**
Access to real-time data and images	**2**

Data acquisition	**# of data initiatives**
Most data sets already held by the lead institutions, and others acquired through external sources	**1**
Computer-assisted phone interviewing (CATI)	**2**
Survey questionnaires	**6**
Data acquired from external sources (e.g. - Statistics Canada, a polling firm, etc)	**7**
Survey data extracted from administrative files	**4**
Researcher deposit	**3**
Scientific instruments	**7**
Unknown	**1**
National census records	**5**

Organizational models	# of data initiatives
2 lead institutions, supported by External Advisory Group and Partners Operational Management Group	1
Stand-alone project supported by partner universities, federal gov't and other research institution	1
Stand-alone data centre/project	7
Stand-alone data centre/project with involvement from parent institution and other data mgt org	1
Federally-mandated organization with various national stakeholders	1
Distributed network	6
Demonstrator project	1
Stand-alone data centre/project, also part of distributed network	8

Preservation	# of data initiatives
Quality control of structured annotation (metadata)	11
Curation method unknown	10
Some form of archival backup	3
Data Management Archive System (DMAS) for both live and archived data	2

Funding	# of data initiatives
Funding from a granting agency	8
Source of funding unknown	3
Partially funded by parent/member institutions	2
Publicly funded	17
Funding provided by more than one granting agency	5
Funding from participating institutions' membership dues	2
Private sector support	8
Additional funding from a variety of research organizations	4

Summary findings for **17** current **international** data management initiatives in terms of ***access***, ***services*** offered, ***data acquisition***, ***organizational models***, ***preservation***, and ***funding***:

Access
In the case of **12** projects, the data is freely available to anyone. One initiative requires researchers to submit an application before accessing the data at no cost (Henry A. Murray Research Archive), one of the projects makes the data freely available but only to researchers from affiliated institutions (Australian Social Science Data Archive) and non-affiliated researchers pay for access "at cost" per requested dataset. In another project, free access is given in conjunction with certain licensing terms and the provision of attribution for the data used.

For two of the international data management projects, access to the data is fee-based: Australian Social Science Data Archive – "at cost" per requested data set for non-affiliated researchers, and Natural Environment Research Council (NERC) Data Centres charge a fee depending on the use requested data will be put to.

One project has a data access model based on access to members paying consortium fees: Inter-University Consortium for Political and Social Research.

One gives access "at cost" to some datasets and free access in general to members of the higher or continuing education communities, and also offers a great deal of free material to the general public: Arts and Humanities Data Service.

Two of the data management initiatives provide access through a subscription model: Biology Image Library (individual, institutional and corporate rates), and Cambridge Structural Database.

In the Health Sciences, there is a tendency towards providing free access to data; **5** of the 6 documented international Health Science data management initiatives give access to their archived data sets free of charge.

The instance of free access to datasets in the other disciplinary areas:
Arts and Humanities: **1**
Social Sciences: **3**
Natural Sciences and Engineering: **3**

International initiatives (cont.)

Services	# of data initiatives
Educational services in data mgt practices	5
Assistance (e.g.- online help files or expert guidance)	3
Curation assistance (e.g.- recognized data documentation standards, quality control for deposited data sets, etc.)	6
Long-term preservation	3
Data deposit	6
Online search and retrieval tools	10
Online analytical tools (e.g.- data analysis without file downloading, visualization, graphical displays, etc.)	8
Data conversion	4

Data acquisiton	# of data initiatives
Researcher deposit	6
External sources (e.g.- universities, government organizations, market research companies etc.) and researcher deposit	6
External sources (e.g.- universities, government organizations, market research companies etc.)	1
Unknown	2
Researcher and journal deposit	1
Scientific instruments	1

Organizational models	# of data initiatives
Distributed network	6
Stand-alone data centre (some with units performing specific operations - e.g. in the case of the ICPSR, Collection Development, Collection Delivery, and Educational Resources)	5
Centralized resource centre, and part of distributed network	5
Series of test beds supported by a coordinating centre	1

Preservation	# of data initiatives
Some form of archival backup	**5**
Daily backup, archival backup storage, and migration planning	**2**
Unknown	**6**
Quality control of structured annotation (metadata)	**6**

Funding	# of data initiatives
Granting agencies and the host institution	**2**
Granting agencies, the host institution, and in-kind support from another organization (e.g.- an advanced computing agency)	**1**
Various institutions and granting agencies	**1**
More than one granting agency or institution	**4**
Consortium membership dues or subscriptions	**3**
Mix of self-funding (from *core budget*) and licensing of data products	**1**
Endowment	**1**
Single granting agency	**3**
Unknown	**1**
International voluntary contributions	**1**

International initiatives (cont.)

Out of **19** international demonstrators and data projects in development, the following observations are noted. Note: they are incomplete compared to the summaries for the other documented data management initiatives because they are in varying stages of development and/or not as much information was available for them on the Internet.

- For **three** initiatives, search and retrieval services are mentioned. E.g. - Online search and retrieval tools - five federated data grids that enable researchers from any World Universities Network (WUN) site to "seamlessly assemble" collections representing various resources from many institutions
- Interoperability is a common element in the planning of some of the projects . E.g. Development of middleware to support interdisciplinary collaborative projects. **Five** of the demonstrators/initiatives in development are being planned with interoperability
- For at least **three** of these international initiatives, Institutional Repositories will be a key facet of the infrastructure. In one case, one of the services will be the linking of publications with the datasets to which the publications refer
- For at least **five**, the intention is for the deposited data to be archived for long-term access. E.G. – one mentions archiving in IRs, for another the idea is to provide "durable archiving of digital data", and another project has plans for "context-specific tools and metadata for curation to "facilitate the subsequent re-use of the deposited information"
- At least **three** will provide data deposit services, one will offer a niche model for depositing orphan datasets
- **One** mentions planned services for online data extraction
- For **one** of the demonstrators, arrangements can be made for provision of raw images
- At lest **three** will include education/outreach activities in their operational models. E.g. "An attempt is made to make scientists more aware of the possibilities of a digital archive."