Research Data Canada – Infrastructure Subcommittee

Activities 2 to 5: Research Data Infrastructure gaps in Canada, and how to address them

Research Data Canada (RDC) has established the Infrastructure subcommittees (RDC-I) to tackle research data management infrastructure issues. RDC-I established a workplan in March 2013.

This is the second report from the Infrastructure subcommittee. A first report entitled "Activity 1: Identify infrastructure available for data in Canada" was submitted in September 2013 (See reference 2). The present report completes the first document by summarizing the findings of the subcommittee related to activity 2: "identity gaps in data support infrastructure in Canada"; activity 3: "Identification of the degree of awareness of the importance of research data in other efforts to address wider research infrastructure"; activity 4: "Gather capabilities, assumptions, lessons learned, future plans for data preservation" and activity 5: "Evaluate alternate strategies for data infrastructure such as commercial clouds, etc."

The document presents the results of the RDC-I team investigations, to the best of their knowledge. It does not claim to be exhaustive but makes every effort to cover the subject with ample details.

The members of the RDC-I team, throughout this exercise, have been: Jim Ghadbane (CANARIE), David Moorman (Canada Foundation for Innovation), John Morton (Compute Canada), Francis Ouellette (Health, Ontario), Benoît Pirenne (Natural Science, Ocean Networks Canada, Chair), David Schade (Astronomy, NRC), Ray Siemens (Humanities & Social Sciences, University of Victoria), Scott Tomlinson (Aboriginal Affairs and Northern Development), Mark Wolff (CANARIE).

Summary of findings and recommendations

The following is a brief summary of the committee's main findings and recommendations. Please refer to the text for more details.

- Network capacity is excellent for most of the country's users
- Storage capacity is relatively easy to obtain, but not (yet) for archival purposes
- Many research data management software modules — collectively referred to as 'middleware' — are missing. Middleware is the key element that leverages raw storage and networking infrastructures to form what we refer to as "Data Stewardship Facilities"
- Data Stewardship Facilities (DSF) should play the following roles:
  - Recommend, help implement and support discipline-specific data description standards and implement archiving policies at the jurisdiction (country) level
  - Provide long-term research data storage and retrieval services, supported both by up-to-date software systems and knowledgeable personnel

  - Be the recipient of data products from on-going or completed research programmes, following funding agency requirements
  - Manage the technological transitions caused by short time scale evolutions in the areas of hardware, software and data description standards, as dictated by the realities of digital technologies.
- Digitization of existing analog material is still a work in progress, in particular in the social sciences and humanities

The concept of Data Stewardship Facilities has the potential to address issues such as those related to: too many data repositories, abandoned data, poorly described results, untraceable sources, unreadable digital media and inaccessible records, to cite just a few. Successful single-discipline DSF examples already exist in Canada. Possible governance, management and implementation models for DSFs will be the object of a future RDC report.

I. Context and definitions

This report on gaps and alternative strategies assumes that the following five concerns have already been addressed by the parties interested in preserving research data:

- Data are organized in a way that is useful for the purpose of the project and exist in some form that makes them suitable for computer-based management
- Data that should be preserved have been identified
- Data that should be made accessible (and by when) have been identified (data policy)
- Data that are identified for preservation and sharing are suitably organized
- Preferred method(s) for making data available have been agreed upon

This document relies on definitions summarized in the Research Data Canada Glossary of terms (See reference 4). For other terms, specific to this document, see the table below.

Term	Definition in the context of RDC
Middleware	From Wikipedia: "Middleware is computer software that provides services to software applications beyond those available from the operating system. It can be described as "software glue". Middleware makes it easier for software developers to perform communication and input/output, so they can focus on the specific purpose of their application."
Data	From the U.S. National Science Foundation: "The recorded factual material commonly accepted in the scientific community as necessary to validate research findings. This includes original data, but also 'metadata' (e.g. experimental protocols, code written for statistical analyses, etc.). It is acknowledged that there are many variables governing what constitutes 'data', and the management of data, and each area of science has its own culture regarding data."

Table 1: Definitions of terms used.

II. Gaps

In the following, we present some of the most important gaps that the team has identified and that pertain to the ability to deliver research data management services, including preservation and access. Right from the outset, we would like to dispel common misconceptions related to science data archival. Indeed, contrarily to the impressions that might have been left by hardware vendors, purchasing disks, tapes, computers and network access is clearly not sufficient for dealing with the preservation of research data. Buying a few disks to keep data would only transpose a pre-existing entropy. Similarly, moving to commercial clouds for storage without building the capacity for data management, preservation, and access merely moves the site of data entropy. As we will see later in this report, the key gaps have more to do with appropriate software and people expertise.

Data Management Systems galore

Anyone with any research project producing any data volume will quickly realize that some form of data management is going to be necessary to support the goals of the project. And that some form of infrastructure to link data its users will be necessary. With data originating from a variety of sources, following various formats and types, and with few existing systems available to bring order, we witness more of a plethora than a dearth of data management systems. The previous report of this working group clearly points this out (Activity 1). Indeed, it would appear that many funded science projects have considered data management as an afterthought or have considered that their needs are so different from everybody else's that they felt compelled to create a home-grown, stop-gap system to deal with them. These systems have therefore each been created to satisfy immediate and specific needs. Moreover, they were developed only to the extent that the PI of the project has lent his/her data/software people both credence and resources.

This current ad hoc situation is the source of a number of issues:

- we end up with many different, small systems having each their own purpose, unclear extensibility and, quite likely, their own access methods and idiosyncrasies;
- access to the data is not uniform or/and may not be public;
- standards, if they exist, can be costly to implement and therefore those small systems provide incompatible data sets that cannot be seamlessly used with those coming from other, similar sources.

The source of many of the gaps mentioned above is a **lack of ready-to-use discipline-specific common facilities that can support small and medium research projects and their output. Those common facilities, if they did exist for more disciplines, would provide reliable and secure data collection, storage, distribution and stewardship**. This gap forces individual projects to re-invent their own system for every new project, for each science discipline.

An encouraging counter-example in Canada in the field of astronomy is the Canadian Astronomy Data Centre (CADC) that has, over the past couple decades, demonstrated that it was able to deliver

the key services of reliability, security, interoperability and access method and has, over the years, convinced more and more of the smaller observatories and experiments to entrust their data to them. This was however, a fairly isolated case of a grass roots initiative that quietly found its way to success. It was not imposed to anyone or necessarily well supported by funding agencies.

Common facilities such as the CADC will be hereafter referred to as “Data Stewardship Facilities”.

Infrastructure issues

In order to reliably satisfy their mandate, Data Stewardship Facilities have infrastructure needs. Infrastructure is necessary to successfully carry out the four activity areas: *data collection/acquisition, storage, distribution and stewardship*.

Acquisition and distribution clearly require **networks**. Acquisition moreover requires **software** that will specifically speak the language of the data producers (instruments, etc.). These will evolve over time and the software will have to be adapted.

Data distribution not only requires suitable, specific user interfaces and application programming interfaces, it also requires data transport protocols and data encoding mechanisms. Those are typically **software**-based. Both user interfaces and application programming interfaces will have **usable lifetimes that are shorter than that of the data they are to serve**. (E.g. the character-based terminals that were prevalent until the early nineties have all but disappeared now. Today’s web-based technologies will also someday cede their place to approaches we cannot yet fathom). So again here, supporting software will have to evolve with technology to remain relevant and usable.

Clearly, storage requires technology, but as the past few decades have demonstrated, the typical **storage technology** lifetime is usually much shorter than that of the data it supports. Trying to keep old technology around past its prime is a recipe for increased management costs. Therefore, proper long-term thinking needs to be present and resources to evolve this part of the infrastructure must be available.

Data archival is linked not only to simple storage, but involves questions of **retention protocols** and practices, data management **planning** and the **regulatory apparatus** around mandatory deposit of data, in other words the implementation of **data archival policies**.

Data stewardship is the activity that allows a dataset to remain usable by having agents (**software systems** and/or **human experts**) continuously ensuring the quality of data and of their attendant metadata. The agents’ role is to verify consistency and completeness, quality and description. User support is also a critical item if there is a desire that non-specialists also be in a position to understand and use a specific dataset.

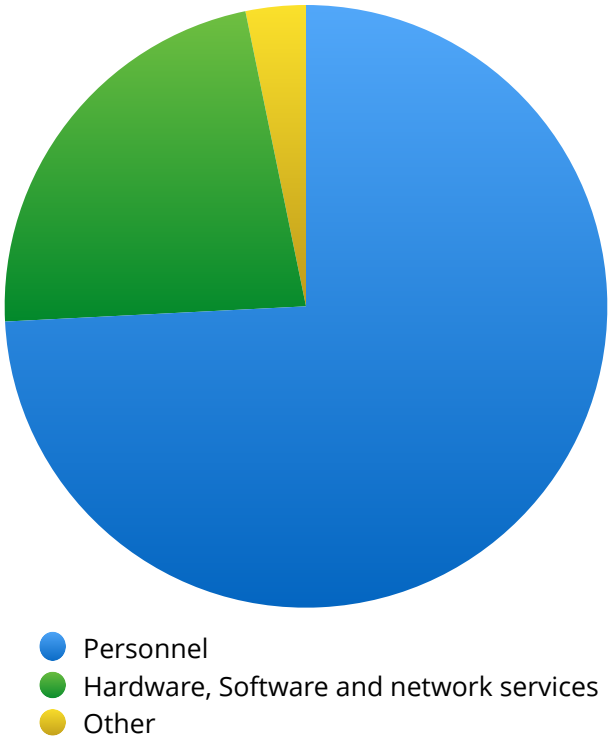
An encouraging sign that lessons are being learned in this area can be observed in a recent proposal for a new data management facility for the Canadian High Arctic Research Station (CHARS) (See reference 5). The proposal makes it clear that its proponents are aware of all the issues addressed

by RDC. Indeed, most of the aspects the proposal are in line with current thinking related to science data management facilities. The proposal, however, could take a longer term view rather than focusing too narrowly on infrastructure implementation details that depend on today's technologies.

In summary, most of the challenges of Data Stewardship Facilities revolve around networks, storage technology, the ubiquitous software and the expertise around the data. Software is particularly discipline-specific, is subject to regular evolution and adaptation and will become the most significant expense of the entire facility in the long run. Data storage costs are volume-specific, but are likely to be significantly lower than the attendant personnel costs in charge of software and

Equipment vs. People

Experience in the natural sciences (see reference 7) shows that, in the long run, the data stewardship and management cost breakdown follows more or less those proportions:



operations.

When it comes to the specifics of the Canadian landscape in terms of core infrastructure, we are fortunate to be able to rely on unified, national organizations with presence and representation from coast to coast. CANARIE, with the help of provincial partners, delivers the digital

communication capabilities that enable researchers to connect with data and peers. Compute Canada is offering storage and compute capabilities to all researchers in the country through reliance on pre-existing regional consortia and by managing a unified and centralized competition process for access to any of the country's participating resources. Whereas the CANARIE model satisfies most needs when it comes to research data transport, Compute Canada, which has been considered a possible solution for a long term storage platform in the country, cannot currently provide that service. The main reasons for this is that the Compute Canada storage that scientists can compete for has to be (re)justified yearly, and that justification has to be supported by an ongoing stream of scientific outcomes, publications and activities. At this time Compute Canada is therefore not a solution for the long term storage of data that is likely to be dormant at the end of a research project. Compute Canada's approach in this area is understandable as they do not want to be in charge of looking after data that may not be well managed and that may not be described following recognized standards, etc. A solution to this issue would be to explicitly introduce a mandate in the Compute Canada's portfolio, whereby storage is made available on a long term basis to well managed, described and accessible data collections, even if they are not necessarily currently producing streams of scientific publications. Of course, alternatives for storage exist through commercial cloud services. Those however often suffer from the issue of unclear data residence, which is a problem if one considers the legal obligations associated with e.g., health data that must remain in a given Province, or at least in the country, out of privacy concerns.

From a Humanities and Social Sciences perspective, it might be useful to consider the cyberinsfrastucture report (see Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences — reference 8). Responses to the current tri-council discussion documents from Humanities and Social Sciences communities have drawn on and pointed to it quite extensively, as well as a similar report from the European Science Foundation (see "Humanities Researchers and Digital Technologies: Building Infrastructures for a New Age" — reference 9), and work such as the CFI-funded "Canadian Writing Research Collaboratory" (CWRC — reference 10) project has drawn on it quite extensively.

Interoperability

Even in an ideal situation where only a few discipline-specific Data Stewardship Facilities exist -for example one for each large science area from astronomy to social sciences through biology and ocean sciences, etc.- the need for interoperability is still there as, internationally, groups have to remain compatible with their peers when it comes to data access, data transport and data formats.

Interoperability, simply put, is achieved when two or more organizations agree on data exchange methods between them. There are two main areas that need to be agreed upon: the data packaging (e.g., formats and semantic description of file content) and data access and transport protocols (how to search for, request and transport datasets). It turns out that interoperability is usually very discipline specific, with many groups having spent many years to try and **define suitable structures** that would address a sufficiently large fraction of all their requirements. Then it is up to the various participants to **implement the protocols and formats**, usually through software systems that

perform translation from one native format into the agreed upon standard. Those standards typically evolve to address new requirements and adapt to technology changes.

A Dearth of Digital Data

One of the impediments in the humanities today is the lack of well-preserved and described data itself — particularly that computationally-tractable data in humanistic fields which is or can be amalgamated in the way that in many science and applied science disciplines are taken for granted. There are **'disciplinary' holes in digital humanistic data**, and this might require some investigation. CRKN, the CFI-funded Synergies and TAPoR projects, the SSHRC ITST program, the earlier National Data Archives consultation, and other initiatives have been very helpful here, but additional investment (intellectually and otherwise) to bring the state of specifically humanistic data to the level of, say, astronomy or oceanographic sciences, would be beneficial. (See also reference 1).

On a related topic, recent outrage at the dismantling of science libraries at DFO sites (see e.g., the Tyee article mentioned as reference 3) has highlighted concerns regarding the comprehensiveness and quality of digitization prior to the destruction of old analog records.

III. Summary

Whereas in Canada today there is a significant digital networking infrastructure as well as ample storage capacity, the most important gap appears to be in the area of the “middleware” (or what is better described as research data management software modules). Middleware is a catch-all term for software types that include a very broad range of functions, from data flow management to visualization systems. The middleware gap can in turn be traced back to a lack of support for suitable software engineers and data experts whose task it is to implement the discipline-specific data formats and distribution standards, evolve the data acquisition, storage, search and distribution functions, and deal with the requirements of data quality and user support. In this respect, CANARIE Inc. can be commended for having identified such issues several years ago and funded groups of researchers and engineers in support of science to invest in middleware (See reference 6).

The management of the digitization of humanities and social sciences records to ensure that it is comprehensive and of high quality appears to require attention. If nothing else, the degree of completeness should be qualified, or better still, quantified.

In the area of storage capabilities, an expanded mandate for Compute Canada should include support for long term preservation of well-described digital data, even in the absence of on-going research through well-funded Data Stewardship Facilities.

Data Stewardship Facility implementation options will be the object of a future report.

Appendix I: Notes and References

1. History of Sciences Society Newsletter, *"Report on Data Management and Data-Management Plans for the History of Science Society Committee on Research and the Profession (September 6, 2013)",* Vol. 42, No. 4, October 2013.
2. Research Data Canada – Infrastructure Subcommittee, *Activity 1: Identify infrastructure available for data in Canada*
3. The Tyee — "*What's Driving Chaotic Dismantling of Canada's Science Libraries?"* http://thetyee.ca/News/2013/12/23/Canadian-Science-Libraries/
4. The Research Data Canada Glossary of Terms, November 2013, Draft 3B.
5. The Canadian Polar Data Network: *"Data and Information Management for Arctic Science and Technology: A Proposed Approach for the Canadian High Arctic Research Station",* July 2013
6. CANARIE Inc.'s "Network Enabled Platform" and "Research Platform Interface" programs support middleware development as described in the summary section. See: http://www.canarie.ca/en/programs
7. Digital Infrastructure Hardware vs. personnel cost source is from Ocean Networks Canada. The ratio is similar for the Canadian Astronomy Data Centre. More comparison with other disciplines, such as bioinformatics centre etc, would be interesting to have.
8. Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences
9. European Science Foundation: "*Humanities Researchers and Digital Technologies: Building Infrastructures for a New Age*"
10. The Canadian Writing Research Collaboratory (CWRC, pronounced "quirk") http://www.cwrc.ca/en/ is a Canada Foundation for Innovation supported project (among other funders).