Partnership
The Canadian Journal of Library and Information Practice and Research

vol. 9, no. 1 (2014)

Developing a Research Data Management Service – a Case Study

Jeff Moon
Data & Government Information Librarian
Queen's University
moonj@queensu.ca

Abstract

Publicly-funded, researcher-generated data has been on the front burner lately, driven by a variety of factors, including evolving funding-agency policies and journal publisher requirements. In this context, Queen's University Library (QUL) developed and implemented a Research Data Management (RDM) Service to meet researchers' needs. This process is described here, framed around four main themes: planning, building, educating, and doing.

Keywords

research data management; RDM

Background

Research data can be defined as "That which is collected, observed, or created in a digital form, for purposes of analysing to produce original research results" (The University of Edinburgh). Research Data Management is what we do with research data to ensure it is complete, accurate, well-documented, compliant with ethics standards, open, and in preservation-friendly format(s).

Research Data Management fits into a broader Research Data Life Cycle that can be represented graphically in many ways. One illustration of this life cycle is shown in Figure 1.

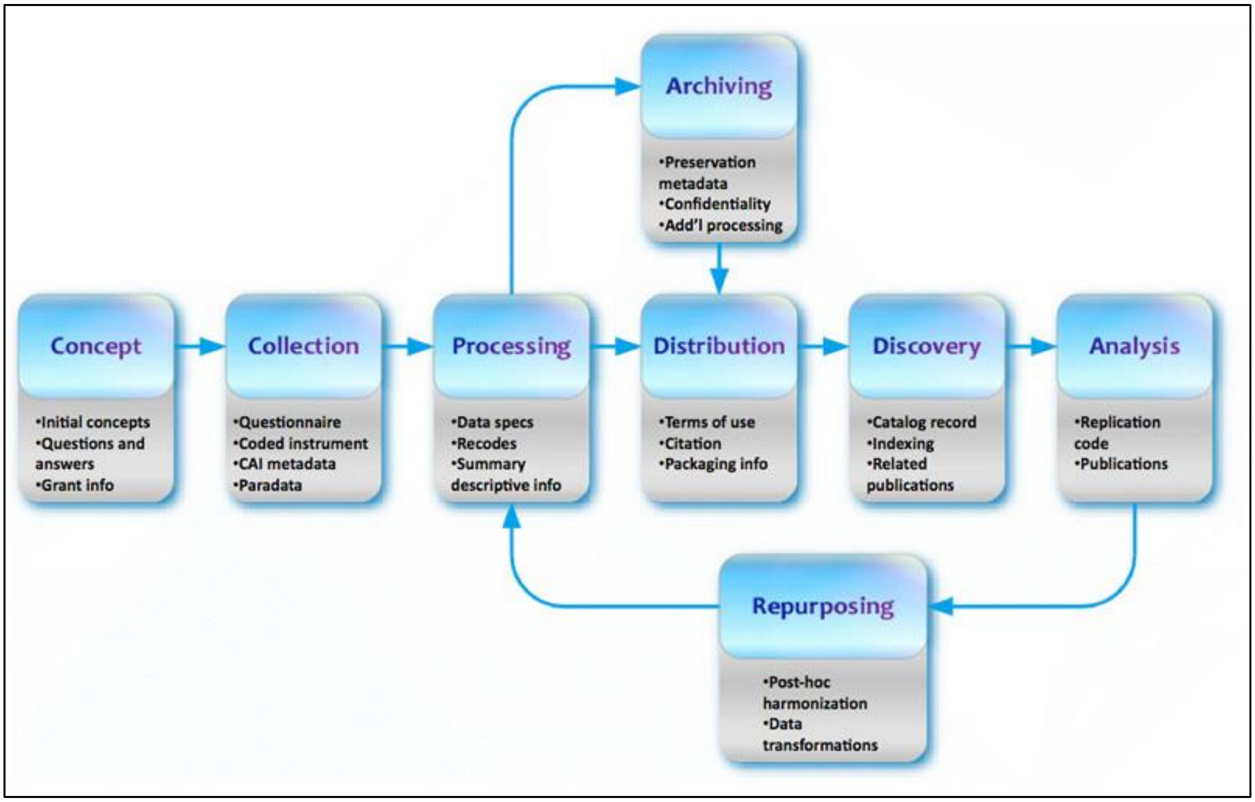


Fig. 1. Research Data Life Cycle (DDI Alliance)

Planning

The decision to build an RDM service at Queen's University was influenced by several drivers, both external and internal. External drivers include funding agencies, publishers, libraries, and national and international initiatives. In Canada, funding agencies such as the Tri-Council are rewriting their policies to require researchers to deposit their data in open archives. Grant applications have sections dedicated to 'Data Management Plans' (DMPs) and examples of 'knowledge mobilization'. From a publisher's perspective, there is a growing understanding of the importance of 'open data' and data citation – "Many publishers are requiring that the data associated with publications be publicly available" (National Research Council 57). Libraries are seeking new ways to embed themselves in research and scholarly communications processes. National and international initiatives (e.g., Research Data Canada and CODATA, respectively) are exploring and addressing this issue as well. For example, the Research Data Strategy Working Group, in their 2011 report "Mapping the Data Landscape", suggests working with universities "to examine the possibilities of incentives for data management, such as those for publications/citations" (14).

Of these external drivers, the October 2013 Tri-Council consultation document, "Capitalizing on Big Data: Toward a Policy Framework for Advancing Digital Scholarship in Canada" was the most compelling.
This document promoted three main goals:

1. Establishing a culture of stewardship
2. Coordination of stakeholder engagement
3. Developing capacity and future funding parameters (Government of Canada).

The Tri-Council sought broad input on these goals, and Queen's University was not alone among Canadian stakeholders in providing feedback. Known respondents included the Canadian Association of Research Libraries (CARL), the Ontario Council of University Libraries (OCUL), Research Data Canada (RDC), and others. This suggests that the Tri-Council's goals resonated with both researchers and administrators and, from our perspective, that libraries had a central role to play.

Internally, development of our RDM service was informed and, in a real sense, accelerated by participation in educational opportunities such as the ARL/DLF E-Science Institute, that brought together "research library audiences seeking opportunities to boost institutional support of e-research and the management and preservation of our scientific and scholarly record" (Council on Library and Information Resources). Among other things, participants in this institute were asked to interview researchers and administrators at their institutions to gather their sense of gaps and relationships in the research life cycle. Shortly after this intensive program, Queen's participated in a CARL-sponsored Introduction to Research Data Management Services course that reinforced the need to act institutionally but to cooperate and coordinate provincially and nationally. Locally, formation of a Library Research Data Working Group sharpened our focus and resolve to act on what needed to be done at Queen's. And, finally, questions from researchers started to filter into the Library – directly from researchers asking about Data Management Plans and indirectly through University Research Services.

Based on these drivers and experiences, the QUL Research Data Working Group drafted a 3-year Research Data Management Plan[1] in 2013. Major sections of this report included:

I. Current Data Services
Like many Canadian post-secondary institutions, Queen's University Library has had a data service for many years, and Research Data Management has been a part of this service, albeit in an informal way. In the last three years, this has changed, with RDM becoming a growing part of the services we offer.

II. Drivers
Our decision to offer RDM services was guided by many internal and external drivers, but ultimately, its inclusion as a priority in the Library's 'Comprehensive Budget and Planning' document brought the service to life. Administrative foresight and budgetary support are key.

[1] Available on request

III. Data Services in 2016

Our vision for future RDM services followed the Digital Curation Centre's (DCC) lifecycle model, which includes the following discrete service steps:

- **Create or receive** – Receive data and metadata from researchers.
- **Appraise and select** – Evaluate data and select for long-term preservation.
- **Ingest** – Transfer data to an archive/repository.
- **Preserve** – Take steps to ensure long-term preservation and integrity/authority/usability of data.
- **Store** – Store data securely, according to relevant standards.
- **Access, use, and reuse** – Ensure data is accessible to authorized users and that users are aware of both the RDM service and data files available.
- **Transform** – Create new data by migration to new formats, creation of subsets/recodes, etc.

It became our goal to build a service that would fulfill these steps.

IV. Partnerships and Collaborations

Partnerships, both internal and external, have been key to the development and success of the RDM Service at Queen's University Library. Partners at Queen's include the Library (and various working groups), IT Services, Queen's Research Data Centre, the General Research Ethics Board, and more. External partners include OCUL, the OCUL Data Community, CARL, RDC, and more. This list illustrates how complex Research Data Management is and the ongoing need for highly trained and skilled professional and technical support to manage, sustain, and grow these collaborations and partnerships.

V. Implementation Principles

Implementation principles made clear our common understanding of how we would reach our goals. Principles agreed upon included the need for sufficient and sustained budgetary support, a philosophy of learning by doing, a commitment to grow through collaboration – not isolation, and a desire to promote the importance of data in a balanced academy.

Regarding the latter, the working group was in complete agreement that open access to research data was essential to peer review and extending research. Though not stated explicitly in the plan, the working group felt that well-documented, accessible, researcher-generated data should be given value in promotion and tenure considerations, in much the same way peer-reviewed journal articles and publications are.

VI. Advocacy, Outreach, Promotion and Education

There is a continuum of researchers who need to learn about RDM. At the beginning of this continuum are those writing funding proposals who need to submit RDM and knowledge-mobilization plans. From the perspective of RDM, this is the ideal point to make contact so that best practices can be introduced from the outset. At the other extreme are those who have research data in hand but no RDM plan in place. The majority of deposits received to date are concentrated here.

The biggest challenge experienced so far is targeting our message appropriately, in terms of content, timing, and audience, to achieve the best results. Taking content, for instance, how detailed and subject-specific do you make your message? Regarding timing and audience, when is the best time to reach your target audience: graduate students in the fall, when they arrive? Faculty in the spring, when classes end? Our experience with graduate student outreach sessions has been good. Graduate students are keen to learn about RDM and enthusiastic about spreading the word back to their professors and peers – and often, it is graduate students who are working most closely with the data.

VII. Staffing
Staffing for RDM Services will vary from institution to institution. Currently at Queen’s University, the service is staffed by the Data Librarian (who is also the Government Information Librarian and the Academic Director of the Queen’s Research Data Centre) and one half-time Data Technician. It is understood that, as the service grows, additional resources will be required.

VIII. Evaluation and Assessment
The RDM Service will be evaluated by a variety of metrics, including the number of data deposits received and processed, the number of workshops/sessions given, web traffic on the RDM guide page, use statistics from archives (ODESI, Dataverse), tracking RDM activities and workflows, and external collaborations.

Building

Workflow

A key component of the 3-year RDM Plan was the development of key activities and workflows and the assignment of timeframes and resources to each of these. Ultimately, we mapped out activities/workflow based on the DCC service steps (described above) and experience gained from several pilot data deposits. A simplified version of the current workflow is shown in Figure 2. This workflow purposely omitted mention of a collection development plan since we hadn’t yet written one. Our goal at this early stage was to experiment and see what kinds of data would come our way.

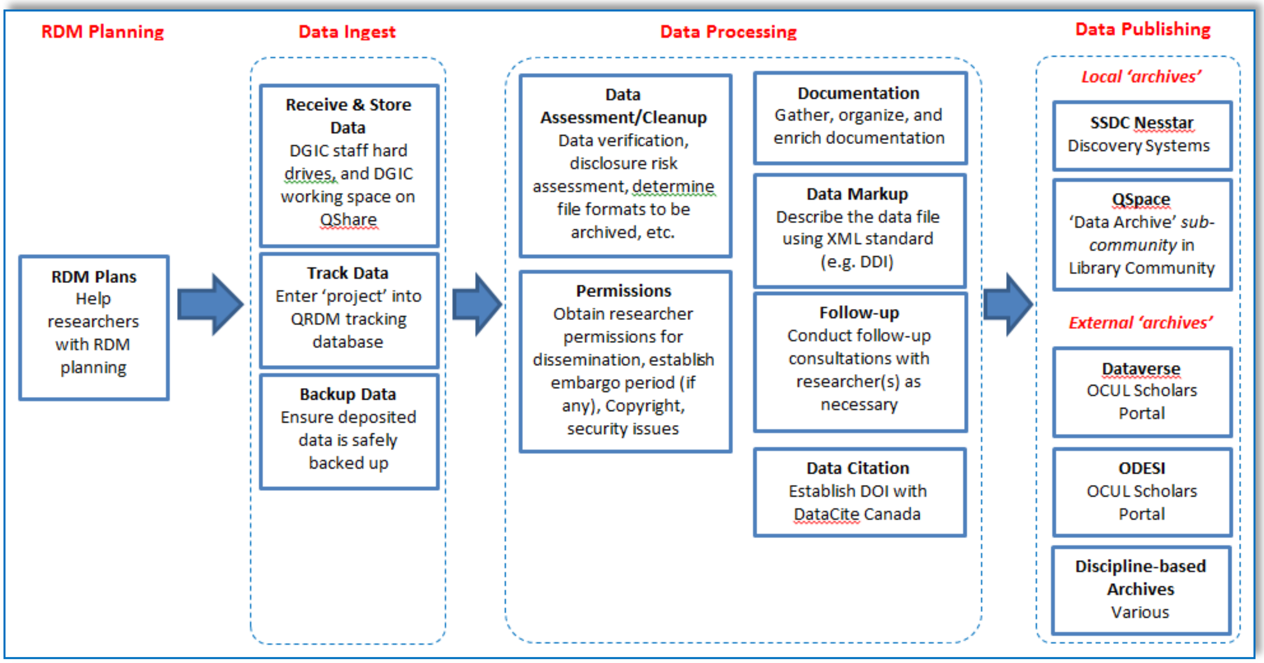

RDM Planning
Data Ingest
Data Processing
Data Publishing
RDM Plans
Help researchers with RDM planning
Receive & Store Data
DGIC staff hard drives, and DGIC working space on QShare
Track Data
Enter 'project' into QRDM tracking database
Backup Data
Ensure deposited data is safely backed up
Data Assessment/Cleanup
Data verification, disclosure risk assessment, determine file formats to be archived, etc.
Permissions
Obtain researcher permissions for dissemination, establish embargo period (if any), Copyright, security issues
Documentation
Gather, organize, and enrich documentation
Data Markup
Describe the data file using XML standard (e.g. DDI)
Follow-up
Conduct follow-up consultations with researcher(s) as necessary
Data Citation
Establish DOI with DataCite Canada
Local 'archives'
SSDC Nesstar
Discovery Systems
QSpace
'Data Archive' sub-community in Library Community
External 'archives'
Dataverse
OCUL Scholars Portal
ODESI
OCUL Scholars Portal
Discipline-based Archives
Various

Fig. 2. Research Data Management Workflow

Storage Options

Part of this workflow involved identifying potential data storage options. Our list ultimately included local options such as our institutional repository, QSpace, and consortial options such as ODESI and the Scholars Portal Dataverse. For the moment, we are using the following storage/archiving options, depending on the needs of the data deposit:

- **QSpace** – *Local*: Queen's University's institutional repository, based on 'D-Space'
- **Nesstar Webview** – *Local*: web-based data exploration tool
- **Dataverse** – *Consortial*: Scholars Portal's instance of this Harvard-developed data platform
- **ODESI** – *Consortial*: Scholars Portal's data portal
- **Discipline-based Archives** – *External*: Archives beyond our local and consortial reach but designed specifically for data in various disciplines

For each data project, the decision to use one or more of these options was influenced by the conditions associated with the deposit and the features of each system. These features are illustrated in Table 1.

Table 1. Features of Data Storage Options

	Local Control	**Flexible Access Control**	**Data-specific Metadata**	**Online Statistical Analysis Tools**	**Powerful Search Features**	**User-mediated Deposit**	**Archive Tailored to Discipline**
Institutional Repository	yes	no	no	no	no	no	no
Nesstar - local instance	yes	no	yes	yes	no	no	no
Dataverse - Scholars Portal	yes	yes	yes	some	yes	yes	flexible
ODESI	no	no	yes	yes	yes	no	somewhat flexible
Discipline-based Archive	no	unknown	yes	unknown	unknown	in some cases	yes

Data-deposit Form

To help inform these decisions, we asked researchers to complete an online data-deposit form (http://library.queensu.ca/webdoc/ssdc/data-deposit) that prompted them for basic metadata and permission to distribute their data. While local security policies prevent us from allowing researchers to actually deposit data using this PHP-driven form, it does provide an efficient means of gathering metadata in XML-compliant format. Data itself was delivered by other means (e.g., email attachment or in-person consultation). From there, we determined what storage option(s) worked best, based on the data in hand, researcher, needs, and features of the archiving solutions at our disposal.

Following the criteria in Table 1, we would evaluate:

- **Local Control** – Does the data need to be housed and controlled at Queen's for ethical, legal, or other reasons?
- **Flexible Access Control** – Is the data under an embargo? Does the researcher (or more than one researcher) need secure access to the archived data during the embargo?
- **Data-specific Metadata** – Does the data fit the Data Documentation Initiative (DDI) metadata standard, or is another standard more appropriate?
- Online Statistical Analysis Tools – Is there a need for web-based statistical analysis? What kind of analysis?
- **Powerful Search Features** – How important is it that the data be 'findable' via archive-specific search interfaces or broader search tools like Google?
- **User-mediated Deposit** – Is it likely, or desirable, that researchers be able to 'self-deposit' their data?
- **Archive Tailored to Discipline** – Is there a subject-specific archive that would suit the data, or will a more general archive be sufficient?

There have been instances when no one solution met all of a researcher's needs. In one case, for example, a researcher wanted to provide his research team with secure access to embargoed data *and* to provide web-based statistical analysis features. The best tool at our disposal for statistical analysis (tabulation, subsetting, etc.) was ODESI, using the Nesstar Webview interface, but this tool could not provide secure, targeted access to individuals on the research team. Dataverse, on the other hand, does the latter quite well but has limited statistical analysis capabilities. In this case, because the data are embargoed until 2018, Dataverse was the only viable option. Post-embargo, the data will be added to ODESI.

Metadata Standards

To date, we have dealt almost exclusively with numeric data. We have encountered some qualitative data embedded in numeric files (e.g., comments in text format), but no exclusively qualitative data have been deposited. No scientific or alternative format data (e.g., video, sound, etc.) have been deposited. As a consequence, we have only had occasion to use the DDI (Data Documentation Initiative) metadata standard – used for many years to document Statistics Canada and other social science data.

We anticipate soon receiving some mining data from a geology professor, which will be a chance to test how the DDI standard works with scientific data or to find a more appropriate metadata standard to work with. We will certainly try using DDI, given our familiarity with this standard and the availability of free, easy-to-use, mark-up software (i.e., Nesstar Publisher).

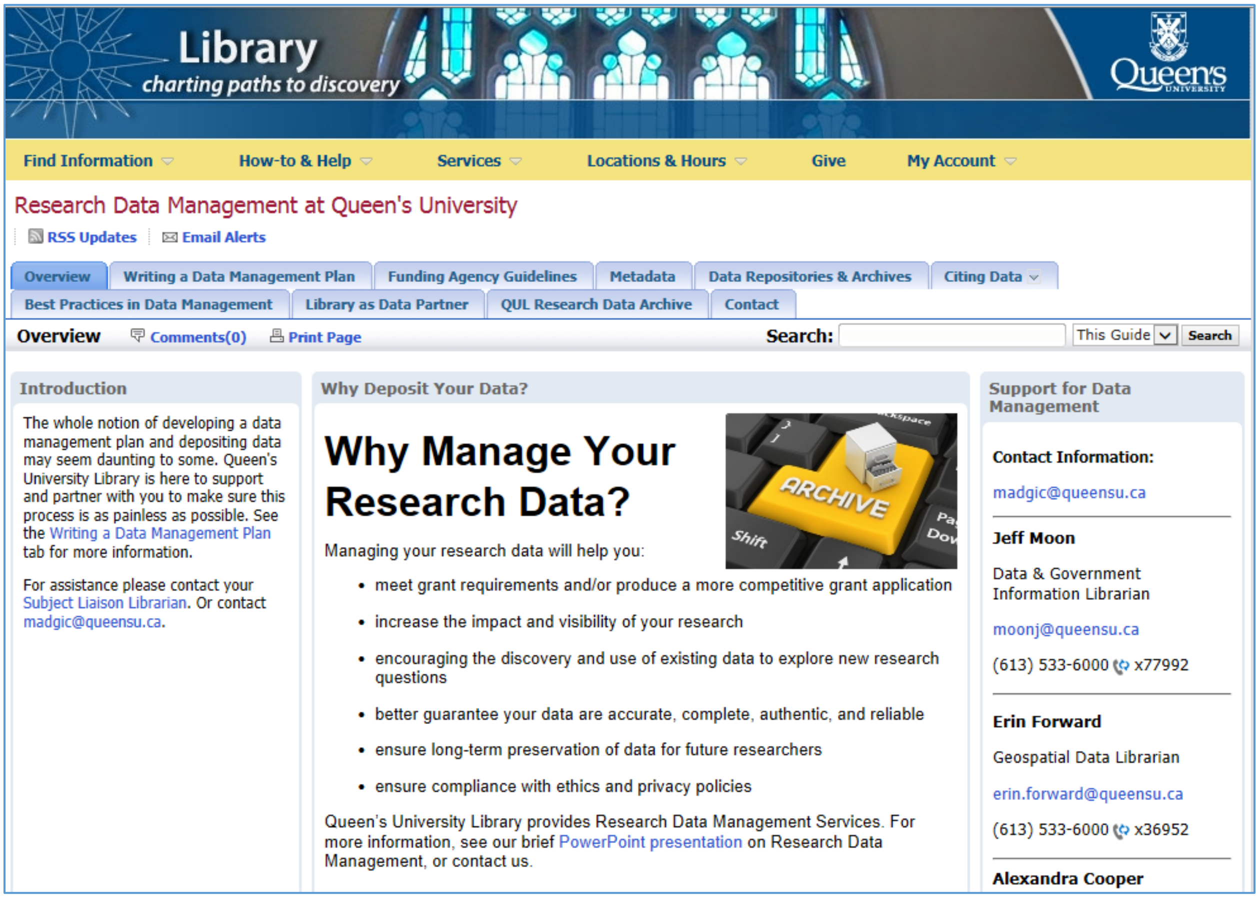
Regardless of the standard used, metadata should:

- Describe data in a standardized and structured form, usable by computers.
- Provide researchers with enough information to make sense of and use the data.
- Capture a dataset's origin, purpose, time reference, geographic location, creator, access conditions, and terms of use, etc.
- Ideally, come directly from researchers, as they know their data best.

The DDI standard meets the first three of these criteria for the numeric social science data we've collected thus far. Any quality issues we've encountered stem from poorly-documented data rather than failings of the metadata standard itself.

RDM LibGuide

Given the range of issues and tools related to RDM, it soon became clear that a web page was needed to gather, organize, and present key resources to researchers (Figure 3). After a thorough survey of established RDM web pages, we sought and obtained permission to model our page on Boston College's Data Management LibGuide.



Research Data Management at Queen's University

Introduction

The whole notion of developing a data management plan and depositing data may seem daunting to some. Queen's University Library is here to support and partner with you to make sure this process is as painless as possible. See the Writing a Data Management Plan tab for more information.

For assistance please contact your Subject Liaison Librarian. Or contact madgic@queensu.ca.

Why Deposit Your Data?

Why Manage Your Research Data?

Managing your research data will help you:

- meet grant requirements and/or produce a more competitive grant application
- increase the impact and visibility of your research
- encouraging the discovery and use of existing data to explore new research questions
- better guarantee your data are accurate, complete, authentic, and reliable
- ensure long-term preservation of data for future researchers
- ensure compliance with ethics and privacy policies

Queen's University Library provides Research Data Management Services. For more information, see our brief PowerPoint presentation on Research Data Management, or contact us.

Support for Data Management

Contact Information:

madgic@queensu.ca

Jeff Moon

Data & Government Information Librarian

moonj@queensu.ca

(613) 533-6000 ☎ x77992

Erin Forward

Geospatial Data Librarian

erin.forward@queensu.ca

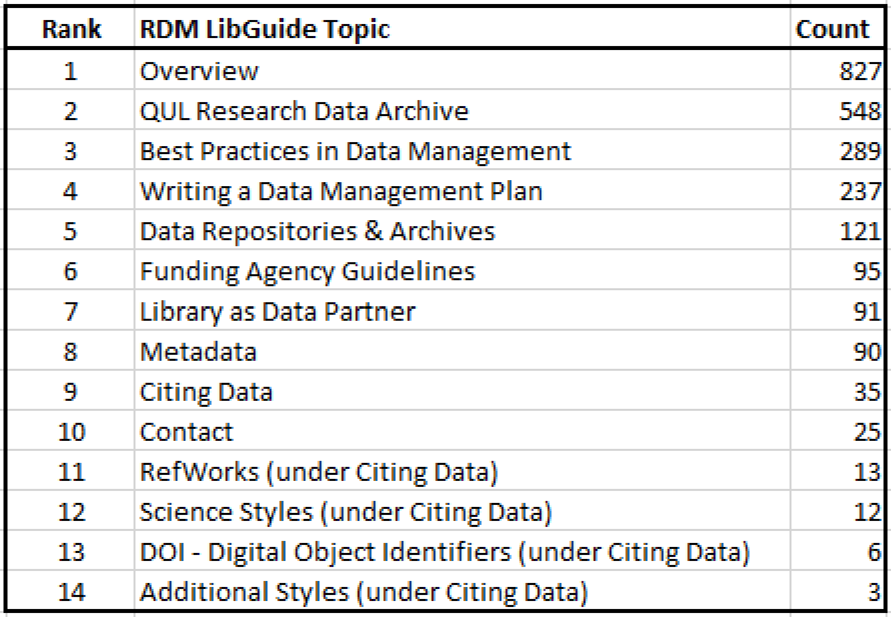
(613) 533-6000 ☎ x36952

Alexandra Cooper

Fig. 3. QUL LibGuide, Research Data Management at Queen's University
http://guides.library.queensu.ca/rdm

Major headings on this LibGuide were compared with our RDM workflow diagram (Figure 2) to ensure we had covered all the key bases. Since its introduction in the fall of 2013, this guide has been updated and streamlined based on identification of new and/or better resources and on our growing experience working with researchers and research data. We have used this guide in teaching and refer to it in all of our correspondence with researchers. Between September 2013 and April 2014, the guide was accessed 2,400 times. Statistics by topic tab are shown in Table 2.

Table 2. LibGuide Statistics by Topic Tab, Sept 2013-April 2014



Rank	RDM LibGuide Topic	Count
1	Overview	827
2	QUL Research Data Archive	548
3	Best Practices in Data Management	289
4	Writing a Data Management Plan	237
5	Data Repositories & Archives	121
6	Funding Agency Guidelines	95
7	Library as Data Partner	91
8	Metadata	90
9	Citing Data	35
10	Contact	25
11	RefWorks (under Citing Data)	13
12	Science Styles (under Citing Data)	12
13	DOI - Digital Object Identifiers (under Citing Data)	6
14	Additional Styles (under Citing Data)	3

Topping the list, not surprisingly, was the main RDM LibGuide overview page, followed by the QUL Research Data Archive where all completed projects are described and linked. Following these, there was roughly the same amount of traffic in Best Practices and Data Management Plans and diminishing activity for the remaining topics. As we get the word out about RDM, we expect activity levels for this LibGuide to increase.

Staffing and Expertise

The RDM Service depends heavily upon sufficient staffing and expertise. At Queen's University, the RDM Service was built using existing experienced staff, but we anticipate that strategic hiring/retraining and ongoing professional development will be needed. For new staff, pressure points will include the need to understand the structure, format, and features of datasets, knowledge of basic metadata elements and tools (e.g., DDI and Nesstar Publisher), familiarity with key software packages (e.g., SPSS, Excel), and a basic understanding of statistical analysis. Our staff need good interpersonal skills as well, given the need to interact with researchers over the course of a project. Plans are being developed to address these needs as demand for the service grows.

Collection Development Policy

A major challenge still facing our service is deciding what data we will collect. Framed broadly, collection decisions are based on a dataset's *manageability* and *value* (Figure 4). Manageability includes such factors as:

- How much data are there? (e.g., file size, number of files)
- How clean are the data?
- How complete and clear is the documentation?

- How familiar is the RDM service with the data discipline?
- What is the data format? Outdated? Open/archival? Proprietary? Tied to specific hardware/software?
- Are there access restrictions?
- How much work is needed for anonymization/de-identification?

Value is perhaps more difficult to define, but includes such factors as:

- Data provenance – Undergraduate? Graduate PhD? Masters? Post-Doc? Established researcher/research team? External agency?
- Reputation of researcher(s)
- Potential for future use
- Potential impact of future use
- Meeting funding requirements
- Reproducibility
- Cost of producing and/or reproducing the original data
- Scope of data – local, small sample; provincial/national, large sample; international

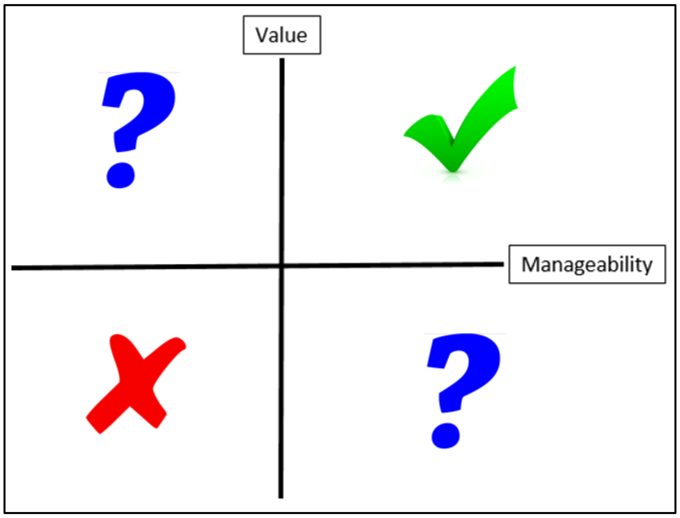


Fig. 4. Collection Decision Matrix

It is evident that ‘*high value*’, ‘*easy-to-manage*’ data would be likely candidates for collection, while ‘*low value*’, ‘*difficult-to-manage*’ data might be rejected. One could also decide to reject ‘*low value*’, but ‘*easy-to-manage*’ data.

The challenge lies in deciding what to do with ‘*high value*’, ‘*hard-to-manage*’ data. What is ‘*easy*’ and ‘*hard*’ to manage may change over time as we gain experience and as tools and standards are developed. The more difficult question is “what is valuable?” Ultimately, ‘value’ will be closely tied to the teaching and research mission of each institution.

At Queen’s University, work on a draft RDM Collection Policy is underway. In the interim, we’ve used experience, common sense, and curiosity to make collection decisions. The latter speaks to our desire to test-drive as wide a variety of data types and disciplines as possible.

Educating

Though still in the early stages of our RDM service, we see educating researchers, administrators, and our Library colleagues as key to success and growth. To date, we’ve shared the RDM message via:

- RDM LibGuide – This guide provides access to key information about RDM and the service we offer and currently hosts a list of completed projects.
- Seminars – We ran one major graduate seminar and several more targeted sessions with discipline-based groups of graduate students.
- Courses – We mentioned RDM in any graduate-level courses we were invited to give.
- Conferences – We described the development of our RDM service at the 2013 Consortia Advancing Standards in Research Administration Information (CASRAI) conference.
- University Administration – We added a representative from the Office of Research Services to our Research Data Working Group, and we mention RDM to administrators at every opportunity.
- Targeted marketing and word-of-mouth – We monitor research being conducted at Queen’s (via media, liaison librarians, and referrals) and proactively approach researchers about their particular project(s).

While all of these are important to our educational efforts, to date we have found the last approach to be the most fruitful. Major research projects tend to find their way into the news (Queen’s University, local, or national media), and targeting these researchers has been key to the growth we’ve experienced to date. In other instances, researchers who have made use of the service have referred others to us – this kind of peer endorsement brings us researchers who are predisposed to depositing data, making our work that much easier. And, we’ve been able to use our success with high-profile research projects as a ‘door-opener’ to conversations with other researchers.

We are also working to enlist and train liaison librarians. Their awareness of research being conducted in their disciplines makes them essential to our identifying and working with researchers. Direct involvement of liaison librarians has been limited to date, more from lack of time than perceived need. Overall, our experience promoting RDM has been positive; researchers, administrators, and librarians have been very receptive to the RDM message.

Doing

The future of our RDM Service will depend heavily on how successful we are at working with researchers to describe and deposit their data. We need to build a reputation for adding value and getting the job done. At Queen's University, we have accepted data deposits from such diverse disciplines as Nursing, Law, Economics, Sociology, and Biomedical and Molecular Sciences[2]. At the time of this writing, we have 11 completed projects, 20 projects in various stages of development, and a further 15 potential projects identified.

To date, all data deposits have been described using the DDI metadata standard, and all deposits have fit into one or more of our selected range of data repositories. We have yet to mediate data deposit into a discipline-based archive. We have worked with graduate students, established researchers and research groups, as well as retired researchers seeking to preserve their life's work.

Conclusion

Our longer-term success will depend on an evolution in thinking about research data. Working with local and external partners, we need to shift the perception of data from being just the raw material of scholarly output to being a valid and valued intellectual output in its own right. Funding council policy directives will contribute to this evolution, but this adjustment in thinking goes beyond money and speaks to the need for changes to traditional value systems (e.g., tenure and promotion) in academia.

This work can only grow as libraries and researchers become more aware of research data management, and as resources, tools, standards, and services expand through local, regional, and national initiatives and cooperation.

Works Cited

Council on Libraries and Information Resources. "Announcing 2012 ARL/DLF/DuraSpace E-Science Institute." Council on Library and Information Resources, n.d. Web. 17 Apr. 2014. <http://www.clir.org/about/news/pressrelease/escience-inst>

DDI Alliance. "What is DDI Diagram." DDI Alliance, 21 Nov. 2013. Web. 28 May 2014. <http://www.ddialliance.org/what>

[2] See the RDM LibGuide (http://guides.library.queensu.ca/rdm) under the "QUL Research Data Archive" tab.

Government of Canada. “Capitalizing on Big Data: Toward a Policy Framework for Advancing Digital Scholarship in Canada.” Government of Canada, 16 Oct. 2013. Web. 17 Apr. 2014. <http://www.sshrc-crsh.gc.ca/about-au_sujet/publications/digital_scholarship_consultation_e.pdf>

National Research Council. “For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop.” National Academies Press, 2012. Web. 1 May 2014. <http://openscholar.mit.edu/sites/default/files/%5Bvsite%3Asite-purl%5D/files/99-106.pdf>

Research Data Strategy Working Group. “Mapping the Data Landscape: Report of the 2011 Canadian Research Data Summit.” Research Data Canada, 2011. Web. 1 May 2014. <http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/obj/doc/2011_data_summit-sommet_donnees/Data_Summit_Report.pdf>

The University of Edinburgh. “Our Definitions.” The University of Edinburgh, 2014. Web. 17 April 2014. <http://www.ed.ac.uk/schools-departments/information-services/research-support/data-library/data-repository/definitions>