CONSOLIDATED RESPONSE

Questions from Innovation, Science, and Economic Development (ISED)
Leadership Council on Digital Research Infrastructure (LCDRI)
Data Management (DM) Position Paper

Question 1: Alignment with ARS and Broader Objectives

"Develop a national software framework that supports the development of innovative local and discipline-specific tools to support researchers with RDM workflows." How would this align with advanced research software development (as a distinct pillar of DRI) and broader objectives around developing generic software tools?

The RDMC national software framework would align directly with Advanced Research Software (ARS) development and broader software initiatives, building a more data-centric, interoperable research software framework for researchers.

ARS can be divided into three types: i) system software that is necessary for computational systems and storage to function at a basic level; ii) middleware that allows researchers to access systems, share data, and manage authentication and reporting; iii) research applications that researchers use to perform their computations, run their models and to analyze, visualize, and manage data. In the RDM context, an effective RDM software development framework would provide software components across all three ARS types to support researchers in finding, accessing, and reusing current and historic data easily and efficiently. In some cases, this ARS for RDM would be more generic or of use across a broad spectrum of disciplines, while in other cases, it would be more targeted to meet the specialized requirements of a particular project.

Some examples of RDM-specific software that would improve RDM practices and allow efficient workflows for researchers would include:

- active data management planning tools accessible throughout a project;
- deposit services to facilitate publication and reusability of data in the most appropriate repository platforms;
- curation tools to facilitate rich description of data and subsequent discovery and reuse;
- middleware to enable the transparent flow of data through all stages of the research lifecycle, increase citability and create rich metrics for the creation, use, and reuse of research data;
- automated metadata extraction tools functioning at the system level as files are created and modified;
- preservation tools such as a Format Policy Registry (FPR) that would operate in concert with archival storage infrastructure to provide a standards-based approach to recognizing contributed data files, automatically extracting additional metadata, and converting files to longer-term preservation formats where appropriate;
- federated discovery tools to help researchers find appropriate data for reuse; and,
- facilitating educational programs with an emphasis on RDM, such as the Software and Data Carpentry programs.

Through its RDM-specific ARS framework, RDMC would work with discipline-specific communities and local front-line service providers to identify the RDM needs of researchers and then support the collaborative development of software solutions to address them. A good example of this is the proposed Format Policy Registry (FPR) that is mentioned in the examples above. In consultation with researchers, RDMC would facilitate the development of a single centralized FPR service for use at all stages of the data storage continuum, from active to archival. Simply adding a data file to a specific location in a national storage infrastructure could trigger a series of machine-driven events that would add substantial value to all stages of the research lifecycle, without the researcher having to do anything more complicated, reducing research administration burden and improving data stewardship.

RDMC's RDM-specific ARS framework would coordinate and integrate with a future ARS framework for the DRI ecosystem as a whole. To help researchers easily and seamlessly meet RDM needs throughout the research lifecycle, it would emphasize the importance of integrating RDM-specific software components into ARS that is related to other components of the DRI ecosystem (e.g. ARS tools that drive instrumentation, perform data transformations or reductions, or integrate research processes) and ensure that these other tools would, in their design, incorporate and interoperate with RDM applications.

In many cases these RDM functions would be implemented as software services (ie. "hidden" computer-to-computer functions) that could be integrated into existing platforms and applications. This would make it easier for researchers to access and use these tools and facilitate application of RDM best practices, as they would be a seamless component of the larger software ecosystem.

It is important to note that RDMC would generally not undertake software development in-house, but would work with existing infrastructure providers, software developers, and research software funders. The proposed approach would also reflect the FAIR principles that make research data findable, accessible, interoperable, and reusable, which are already being used to frame the short-term CANARIE RDM software program.

In order to ensure that researchers are comfortable with using RDM-specific ARS, RDMC would facilitate training programs that would be delivered in conjunction with community partners and designed for graduate students and late-career researchers alike. This training would give additional opportunities to build bridges to ARS within the broader DRI ecosystem.

Lastly, Canadian researchers will not only need access to RDM-specific ARS that is integrated with the broader DRI ecosystem ARS, but also to ARS that is coordinated internationally. To this end, RDMC would also participate more actively in international efforts to build similar platforms, including the European Open Science Cloud, the U.S. National Data Services, and the International Science Gateways initiatives.

Question 2: More Information on Network of Experts

Can more information be provided on the network of experts? Specifically, what role would they fulfill (providing advice, training, guidance to support personnel, providing direct support for researchers, or both?); where would they be based; who pays their salaries; what is the rationale behind the increasing number of experts (due to increasing need for curation experts?), and how would that number continue to grow?

The Network of Experts (NOE) is a key element in the RDMC model, as it helps level the RDM playing field in Canada by enabling a highly efficient pooling of resources and expertise to facilitate more equitable access to RDM infrastructure and services for Canadian researchers, across all disciplines, regardless of where they are located or the size of their institution. For instance, the NOE will help to support and build the skills of front-line staff in smaller institutions that may not have the critical mass or money required to fund full-time, expert personnel, thereby ensuring a basic level of consistent support and practice for researchers across the country. The NOE will also allow the RDM community to benefit from economies of scale and more equitable, shared access to specialized RDM support. For example, it would be cost prohibitive and inefficient for every university across Canada to provide full-time access to an expert in high energy physics-related RDM, as it is unlikely that this resource would be required by any one institution on a full-time basis.

The NOE would be funded through the RDMC, but integrated effectively to leverage and support the expertise found at both regional and local RDM delivery layers. It would also work closely with the RDM communities of practice that are envisioned as another important pillar of RDMC. Central to its mandate would be the provision of a variety of support roles related to implementation of the five core RDM functions that were identified in the LCDRI report: policy; standards and protocols; processes and procedures; leadership, advice, support, and training; and tools and platforms to help enable consistency of practice and approach to RDM across Canada.

While the NOE would be coordinated nationally by the RDMC secretariat, and tasked with supporting national initiatives and serving researchers across the country, it is envisioned that these experts would be deployed regionally and institutionally. Embedding experts in this way would improve RDMC's awareness of, and responsiveness to, researcher needs, as well help it to build strong linkages with front-line service providers. This, in turn, would enable RDMC to support the development of a more consistent and strong RDM culture within institutions and regions across the country and to provide evidence-based, well-informed leadership in the development of a national approach to implementation of the five core RDM functions.

The types of experts who would be hired as part of the NOE would reflect the diversity of expertise that is required. For example, the following types of experts could be engaged as part of the NOE:

- new hires to facilitate consistent and shared data management practices across Canada and to fill gaps where no or limited expertise exists within institutions;
- partial FTEs, (shared institutional/RDMC positions) to leverage local expertise to be deployed nationally on behalf of the NOE;
- RDM Internships for graduates and postgraduates from library, information technology, software development, and data management programs, as well as from discipline-specific communities, to support talent development and training; and
- RDM Champions who are data management experts or data-savvy researchers from a wide variety of research domains, who would provide specific expertise within discipline-specific communities of practice.

These experts would be paid by RDMC through full- or part-time salary dollars or stipends, depending on what form of payment is most appropriate to the type of expert who is being hired.

Decisions about when and where to deploy RDM experts as part of the NOE would be determined in consultation with the RDM community, using factors that would reflect the needs of researchers and front-line service staff. We are proposing a gradual growth in the NOE. This reflects our belief that it will be important to strengthen our understanding of researcher and front-line service needs before hiring and deploying resources. We also expect need to grow as a stronger RDM culture is developed and new RDM policies, such as the Tri-Agency Policy on Data Management, are implemented.

Question 3: More Information on Network of Repositories

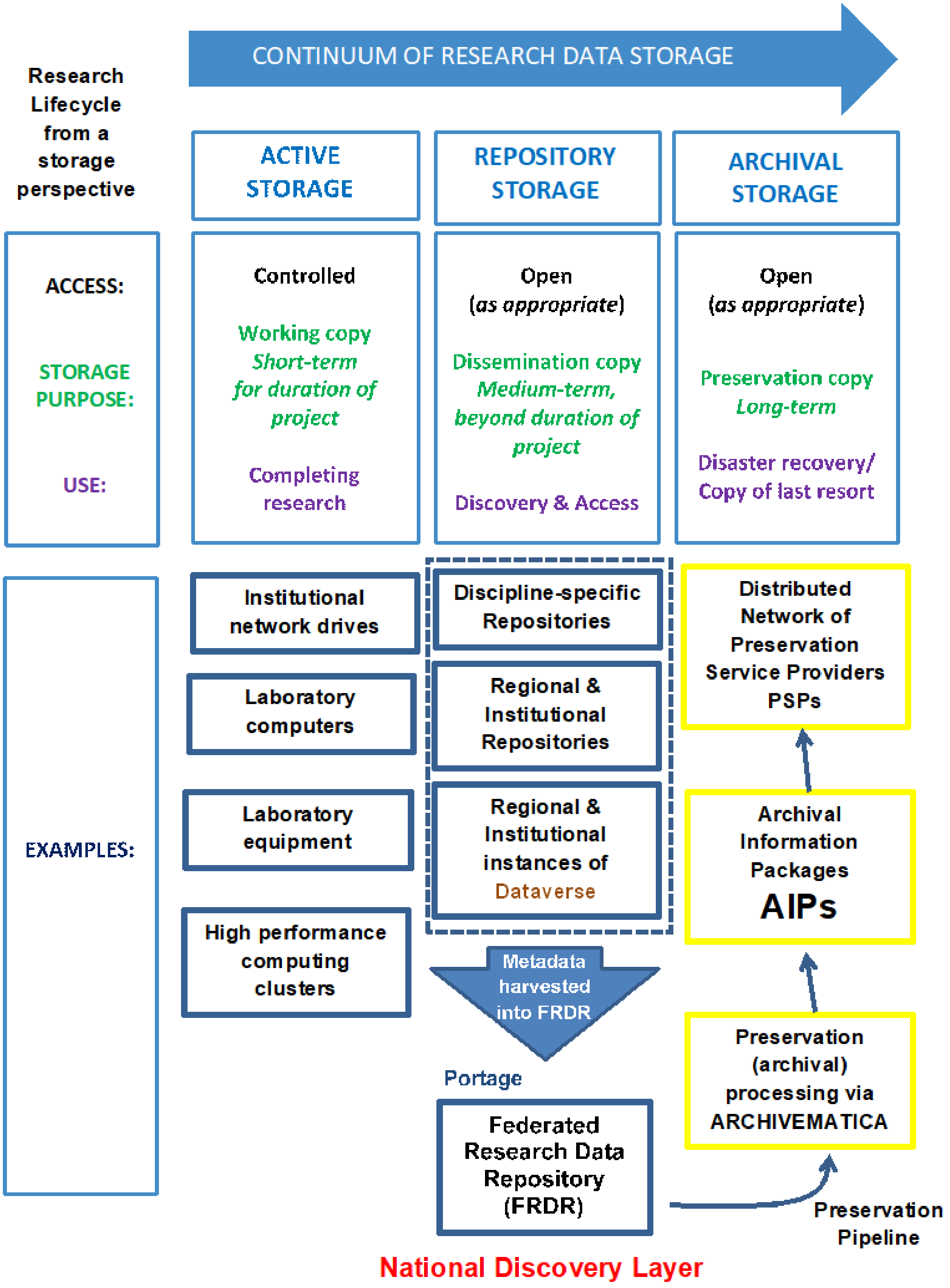
Can more information be provided on how establishing a network of repositories will be implemented? How will some repositories "provide specific services to other repositories"? What would RMDC funding support data repositories (leasing repositories/technologies? Supporting operational costs?) and how will funds be distributed among repositories?

Repositories in the continuum of research data storage

Data repositories are an important bridge between active and archival storage. As illustrated in Figure 1, the three primary types of storage are: **active**, **repository**, and **archival** storage. ARC has a key role in providing the active storage component of this continuum, while DM is more focused on the provision of repository and archival storage. The primary purpose of repository storage is to support data dissemination by ensuring that research data is stored securely and can be discovered and accessed appropriately. Repositories also serve as an important pathway to archival storage. The Portage/Compute Canada Federated Research Data Repository (FRDR) is designed to support both of these core repository functions.

Archival storage is intended to preserve a 'copy-of-last-resort' for the long-term. Specifically, "archival storage is the function of the archival system which manages the long-term storage and maintenance of the content under the stewardship of the archive. In addition to storage, regular maintenance activities such as format migrations, media refreshment, error checking, and disaster recovery plans are a very important part of this service which help to enable long-term access[1].

[1] As described in the Open Archival Information System-- Reference model



CONTINUUM OF RESEARCH DATA STORAGE

Research Lifecycle from a storage perspective	ACTIVE STORAGE	REPOSITORY STORAGE	ARCHIVAL STORAGE
ACCESS:	Controlled	Open (*as appropriate*)	Open (*as appropriate*)
STORAGE PURPOSE:	Working copy *Short-term for duration of project*	Dissemination copy *Medium-term, beyond duration of project*	Preservation copy *Long-term*
USE:	Completing research	Discovery & Access	Disaster recovery/ Copy of last resort

EXAMPLES:

Active storage: Institutional network drives; Laboratory computers; Laboratory equipment; High performance computing clusters

Repository storage: Discipline-specific Repositories; Regional & Institutional Repositories; Regional & Institutional instances of Dataverse

Metadata harvested into FRDR

Portage

Federated Research Data Repository (FRDR)

National Discovery Layer

Preservation Pipeline

Preservation (archival) processing via ARCHIVEMATICA

Archival Information Packages AIPs

Distributed Network of Preservation Service Providers PSPs

Figure 1. Continuum of Research Data Storage

What factors need to be considered and what challenges would need to be addressed when implementing a network of repositories?

There are different types of repository services and this reflects the diversity of communities using or providing them. Any national framework for a network of repositories must recognize this reality. It must also ensure that repositories are interoperable and federated across Canada, and, ideally, internationally. In addition, the framework must situate repositories as part of a broader national storage strategy that provides access to a continuum of research data storage that is optimized for the specific use case. This would enable shared discovery of Canadian research, while maintaining rich domain-specific views of research data in various disciplines.

Canada currently faces a number of challenges in relation to the development of this essential network of repositories. First, there is a lack of repository storage for researchers to use. For instance, Isaac Tamblyn's Computational Laboratory for Energy and Nanoscience at UOIT creates 50 TB of data related to molecular structures annually, which is then used as input to machine learning algorithms. The data is stored in a traditional database using Compute Canada resources, but that storage is only accessible on a maximum 3-year allocation: the hunt for funding to maintain the growing data repository illustrates why this is such a critical gap in so many research labs. How does a researcher like Dr. Tamblyn provide open access to this data beyond a 3-year allocation and in a form that can be accessible to researchers and machine processes alike?

Second, in cases where data repositories do exist (e.g., UBC's Open Science Framework implementation, McGill's C-BIG repository at the local level; OCUL's Scholar's Portal Dataverse platform and the Atlantic Research Data Repository at the regional/provincial level; the Canadian Astronomy Data Centre at the national level; and CCIN's Polar Data Catalogue at the international level), they serve their communities well, but operate largely in isolation from one another, causing data to be siloed. Further, without national cost-sharing support and coordination, they are only able to meet the needs of a small percentage of Canadian researchers.

Third, there are also a number of specialized repositories developed by Canadian researchers that offer, or have the potential to offer, access to invaluable collections of data. However, these repositories may be vulnerable because they have been established and maintained by individual researchers without a strong background in RDM best practices, such as proper documentation, backups, and adherence to standards. These collections of data, typically, have no plans in place for long term sustainability, such as "Data rescue" initiatives that would migrate the data to more stable, shared platforms.

The CARL Portage and Compute Canada FRDR system is an example of a recent effort to try and tackle these challenges by aggregating metadata from multiple repositories, facilitating the transfer of 'big data', and ensuring long-term data preservation through Archivematica. While this effort represents an important step in the development of a national federated approach that can be used to fulfill both repository and archival storage needs in a seamless and transparent fashion, stable long-term funding for this initiative is needed if it is to fully achieve its potential for Canada's research community.

How would the proposed funding be used to address these challenges?

The funding that is proposed in the DM position paper would be used to address these challenges by supporting and coordinating the following activities across the country:

- the development of a coordinated and sustainable approach to storage throughout the research lifecycle that recognizes the need for all three types of data storage;
- the development of mutually-accepted standards and protocols to facilitate the flow of data and metadata;
- lease/license infrastructure, technology platforms, and services necessary to support RDM nationally (e.g. license Globus services). Current information technology makes it possible for a national organization, like the proposed RDMC, to develop, coordinate, and support comprehensive RDM services in Canada without having to directly own and operate them;
- implement succession policies to ensure that all repositories are backed up in the event a repository is forced to cease operating;
- supporting “data rescue”;
- training researchers on the use of data repositories and related tools; and,
- given the constant change in the world of standards and storage, provide funds to spur innovation and ensure that the national federated network of repositories is able to evolve with international communities of practice.

In addition, some funding may flow to repositories that provide services to researchers at universities or research centres that do not have a local data repository of their own - not every institution will need a data repository. To maximize efficiencies, we envision an environment in which repository infrastructure and services are shared across institutions and discipline-specific communities, with the potential for targeted repositories eventually taking on different domain specializations.

Lastly, RDMC repository-related funding would include support for a coordinated approach to the acquisition of storage (cloud-based and local) to benefit from economies of scale. Cloud computing services can be leased from multiple platform providers in customizable configurations, providing a viable and effective means of offering national RDM services. Infrastructure providers can be institutional, regional, national, or international, as well as based on commercial and open source frameworks. This new IT environment provides RDM administrators with an opportunity to provide more cost-effective and efficient services by leveraging the assets and talents of others, rather than having to establish and maintain their own tools and platforms. Where needed, it also provides secure local storage for highly sensitive datasets, where access via the internet is an unacceptable risk.

An issue that was not raised in the DM position paper, but that is important to signal as part of the federal research portfolio, is the existence of a number of key, federally-funded data repositories that are at risk. These repositories are currently funded through research grants awarded to researchers for specific research projects, leaving data in these repositories vulnerable as operational funding for them disappears at the end of a granting cycle. Stabilizing these repositories by making their current funding sustainable is critical to ensuring that the significant investment that has been made in them is not lost and that important data are protected and available for current and future researchers. RDMC could be an appropriate vehicle for administering funds for, and facilitating the maintenance of, these repositories as part of its mandate. This would help to ensure that these key repositories are integrated as part of a broader network/community of practice, operating according to international standards and best practices. It will also send an important signal to the international community that Canada is a leader in good public stewardship of data - in some cases, these key repositories are high-profile, internationally-recognized, and essential to maintaining Canada's research reputation on an international stage.

Examples of such Canadian repositories abound. Many are indexed in the international Registry of Research Data Repositories (re3data.org). Notable examples include: CBrain (McGill Centre for Integrative Neuroscience), Canadian Astronomy Data Centre, Polar Data Catalogue, Ocean Networks Canada, Genome Canada, and the Canadian Writing Research Collaboratory.

Question 4: Centralized vs Distributed Archival Storage

Why would RDMC not establish consolidated/centralized archival storage to achieve economies of scale and cost efficiencies (similar to Compute Canada process), particularly if the federal government was to provide 75% of funding for archival storage? What efficiencies would be lost if maintaining a distributed storage network?

RDMC proposes a national, decentralized-federated approach to archival storage for the following reasons:

1. recognition of the distributed nature and ownership of archival storage;
2. cost effectiveness;
3. risk mitigation; and,
4. regional talent development.

Recognition of the distributed nature and ownership of archival storage

Canada's research community is very diverse, with different needs and practices. Any solution proposed for archival storage must recognize this reality and develop an approach that will meet/respond to diverse researcher requirements and in which they all can have confidence. An example of this diversity would be the physics or genomics communities in which data storage is part of a broader set of global archival storage practices and systems. It would be very difficult to implement centralized archival storage in this context. In addition, certain communities such as those working in health research, have highly specialized security requirements that may make centralized archival storage difficult to realize. Lastly, researchers are often most comfortable adopting new approaches when the infrastructure and services are close to them. Given that a DM culture is relatively new in Canada, it will be important to establish an environment that is conducive to increasing researcher adoption and confidence. Having the flexibility of local/regional and discipline-specific archival storage solutions is important to realizing this goal.

Cost Effectiveness

Avoiding the prohibitive costs of establishing, operating, and maintaining a centralized national data archive is one of the key benefits of a decentralized-federated approach. Technology and expertise need not be centralized to be efficient and cost-effective, and this is particularly true in the context of data archiving. This view is supported by the Portage Preservation Expert Group's (PEG) soon-to-be released White Paper on data archiving. This report highlights that overseeing the provision of active, repository, and archival storage as part of a federated national storage strategy will introduce substantial efficiencies.

To achieve these efficiencies, RDMC would coordinate archival storage functions among new and existing organizations with capabilities and the associated expertise and experience to perform these functions effectively. So, rather than imposing homogeneity in archival storage architecture, RDMC would coordinate a strategic and diverse network of preservation service providers (PSPs), who are federated through a national strategy that identifies gaps and areas of overlap in the delivery of their services, and that defines a set of 'best practice' requirements, while leaving the operation and maintenance of individual PSPs to their host institutions. This approach would allow RDMC to leverage existing institutional and organizational capacity, expertise, and investment in support of the broader archival storage strategy for the country.

Risk Mitigation

A national strategy for archival storage would need to recognize the importance of risk mitigation as part of responsible public stewardship. One of the core value propositions of archival storage is its ability to respond to this requirement - a single centralized storage option is antithetical to this. Decentralization fulfils the "lots of copies keep stuff safe" (or LOCKSS) rule and the best practice of using geographically distributed storage locations to ensure data recovery in the event of a disaster. That the proposed decentralized network of archival storage implicitly achieves geographic dispersion of archived data is an important argument in favour of such an approach.

Regional Talent Development

A decentralized-federated approach would help leverage and grow regional and institutional expertise (HQP). As discussed above, researchers are most comfortable adopting new tools and approaches when the infrastructure and services are close to them. In this context, having well-trained and knowledgeable front-line service providers to help them with RDM is essential. A strategy that recognizes the need for regional talent development in the area of archival and repository storage would serve three purposes. First, it would help to ensure that researchers have equitable access to RDM expertise across the country, regardless of the location or size of their institution. Second, it would support first-hand knowledge development and skills training for these RDM experts who would work directly with both RDM infrastructure and researchers. And third, it would provide an important training and skills development opportunity for undergraduate and graduate students in this growth area. .

Question 5: Growth in Data Production

Can you describe the growth in the amount of data being produced in aggregate and in specific disciplines and the trends that make funding RDM initiatives necessary?

Growth in data production

Science in all areas has become a major producer and consumer of data. Whereas in the past researchers had to contend with the issue of data scarcity, the information age has brought on a data deluge, with data now being generated in unprecedented volumes and variety, and at increasing velocity (Costello & Vanden Berghe, 2006; Poole, 2015). Particularly in the physical sciences, satellites and remote sensing tools are being deployed on a global scale that dwarf traditional sampling methods (Costello & Vanden Berghe, 2006).[2] Additionally, new methods in science such as microarrays, combinatorial chemistry, and sensor networks produce new classes of born-digital data such as workflows, ontologies, supporting code, and other lab materials (Harvey, 2010; Pool, 2015). Well-funded RDM initiatives that are supported by a model of national coordination will be essential to meeting the myriad of challenges posed by increasingly data-driven research.

In this context, it is important to note that currently only a small percentage of data sets produced by researchers are made available and preserved over the long-term. Therefore, we need to build capacity to both manage what is currently being produced and to address future growth in data production.

Trends making funding for RDM initiatives necessary

Drivers affecting the need for a RDM are both top down, with journal and funder policies requiring data to be made discoverable and, where possible, openly available, and bottom up, as researchers are increasingly identifying the advantages of being able to integrate data from a variety of sources across disciplines. In addition, given the proliferation of fraudulent research and research journals, it has become increasingly important for data to be preserved in order to be able to test outcomes and reproduce results.

Journal publishers

Journal publisher policies are increasingly recognizing the value of data as a standalone research output,[3] with the number of dedicated “data journals” on the rise.[4] The number of journals requiring researchers to deposit data and related materials alongside publications is also on the rise, with some journals, such as *PLOS One*, refusing to accept submissions that are not supported by the accompanying data. While data reuse is still an emerging paradigm, it can be expected to rise as more research data becomes widely available. One of the challenges with this new approach is that

[2] See US Integrated Ocean Observing System, the United Nation’s Global Ocean Observing System, and Canada’s proposed Canadian Integrated Ocean Observing System (Bajona, 2017).

[3] Buckland (2011) notes: “The potentially useful record of science is increasingly not the written reports but (mainly non-textual) digital data sets of many kinds: the raw material, the operations upon it and progressively more refined derivations can be beneficially shared and built upon by other researchers’’ (102).

[4] See Nature’s *Scientific Data* journal.

academic journals, many of which are, or are published by, commercial entities, will become the owners of publicly-funded data that they will then be able to use as a very profitable source of revenue generation. This effectively means that public institutions and researchers will need to purchase back publicly-funded data at a potentially high cost.

Funding agencies

Over the past decade, there has been an international trend among funding agencies and governments toward the development of national RDM policies, and the data services that necessarily follow, in recognition of the need for publicly-funded data to be made discoverable and accessible (Shearer, 2015).[5] Within this global context, in 2016, the Canadian Tri-Agencies released a Statement of Principles on Digital Data Management that outlines funder expectations for RDM, and the responsibilities of researchers, research communities, institutions, and funders in meeting these expectations. The Tri-Agencies are now engaged in a consultation process around a draft data management policy that could require “all research data and code that support journal publications, pre-prints and other research outputs that arise from agency-supported research [...] to be deposited in an appropriate public repository or other platform that will ensure safe storage, preservation, curation, and (if applicable) access to the data” (Lucas, 2017). Funder mandates for the sharing of research data are expected to strongly impact the demand for robust RDM infrastructure and services.

Researchers

Researchers are also increasingly discovering the need for and advantages of being able to find, access, and reuse data. This, in turn, is driving a requirement for interoperable, sustainable, and agile RDM infrastructure and services. Usage statistics from Canadian repositories underscore this demand. For example, the Scholars Portal Dataverse reported over 46,000 research data downloads[6] over the 5 years of its existence, and UBC’s Abacus Dataverse Network reported over 45,306 downloads.[7] According to a recent report by the Portage Data Discovery Expert Group, there are roughly 170 data repositories in operation across Canada (Vejvoda et al., 2017). While these repositories are able to serve some of the needs of their respective researcher communities, support for national RDM coordination is needed to create the shared services and infrastructure required to ensure that data do not become siloed across many repository platforms that do not interoperate or conform to international standards and best practices.[8] A recent survey of Canadian researchers found that the majority of engineering/science researchers who participated were already depositing research data in external data repositories, but less than 40% felt their data were sufficiently documented for someone outside their lab to use (Sewrin et al., 2016).

[5] Examples include: the US National Data Service (http://www.nationaldataservice.org/), the Australian National Data Service (http://www.ands.org.au/), and the UK Data Service (https://www.ukdataservice.ac.uk/).

[6] As of 26 Oct 2017.

[7] As of 30 Oct 2017.

[8] See our response to Question 3 for more information on the proposed network of repositories and Appendix 1 for more information on FRDR.

The National Digital Stewardship Alliance (NDSA) recently surveyed 133 institutions engaged in digital presentation activities to investigate how these organizations staffed and organized their digital preservation functions, and to identify any changes since the NDSA's 2012 survey (see Atkins et al., 2017; NDSA, 2012). They found that data holdings had grown since the 2012 survey, and while the majority of respondents expected less than 25% growth of repository holdings over the next year, respondents reported that they require nearly twice as many FTE staff to properly manage current holdings. A recent survey conducted by Portage examining staffing levels for RDM activities in CARL institutions found an average of only 1.42 FTEs directly supporting RDM functions, far below the NDSA average of 13.6 FTEs. Comparing the Canadian FTE levels to the NSDA survey, which looked at organizations primarily in the US, but also internationally, one could conclude that increased investment in RDM initiatives is needed to properly manage existing data holdings, without taking into consideration expected increases in research data deposits from the drivers outlined above.

Another trend driving RDM initiatives is the need to be able to re-access data in order to reproduce and verify results. In general, this is important to good public stewardship of publicly-funded research, but, as mentioned above, this has become increasingly important in the context of a growing proliferation of fraudulent research and research journals.

References

Atkins, W., Ghering, C., Kidd, M., Kussmann, C., Perrin, J., … Talbot, K. (2017). Staffing for effective digital preservation: An NDSA report. Retrieved from https://osf.io/3rcqk/

Costello, M. J., & Vanden Berghe, E. (2006). “Ocean biodiversity informatics”: A new era in marine biology research and management. *Marine Ecology Progress Series*. Retrieved from https://researchspace.auckland.ac.nz/handle/2292/7190

Harvey, R. (2010). Digital curation: a how-to-do it manual. *Neal Schuman, New York*.

Lucas, M. (2017, September). *Tri-Agency Data Management Policy Initiative.* Presented at Portage & Research Data Management in Canada - RDA 10th Plenary Collocated event, Montreal, QC. Retrieved from https://portagenetwork.ca/wp-content/uploads/2017/09/TC3-Data-Management-Policy-Initiative.pdf

Poole, A. H. (2015). How has your science data grown? Digital curation and the human factor: a critical literature review. *Archival Science, 15*(2), 101–139. https://doi.org/http://dx.doi.org.ezproxy.library.dal.ca/10.1007/s10502-014-9236-y

Sewerin, Cristina; Barsky, Eugene; Dearborn, Dylanne; Henshilwood, Angela; Hwang, Christina; Keys, Sandra; Mitchell, Marjorie; Spence, Michelle; Szigeti, Kathy; Zaraiskaya, Tatiana (2016). From Coast to Coast: Canadian Collaboration in a Changing RDM Seascape. Retrieved from https://tspace.library.utoronto.ca/handle/1807/72802

Shearer, K. (2015). *A comprehensive brief on research data management policies*. Retrieved from https://portagenetwork.ca/wp-content/uploads/2016/03/Comprehensive-Brief-on-Research-Data-Management-Policies-2015.pdf

Vejvoda, B., Ambi, A., Barsky, E., Lindstrom, K., MacDonald, H. … Thompson, K. (2017). *Portage Data Discovery Expert Group - Collections Development Working Group: Phase one report.* doi: 10.14288/1.0351978

Question 6: Quantitative Examples of RDM Value

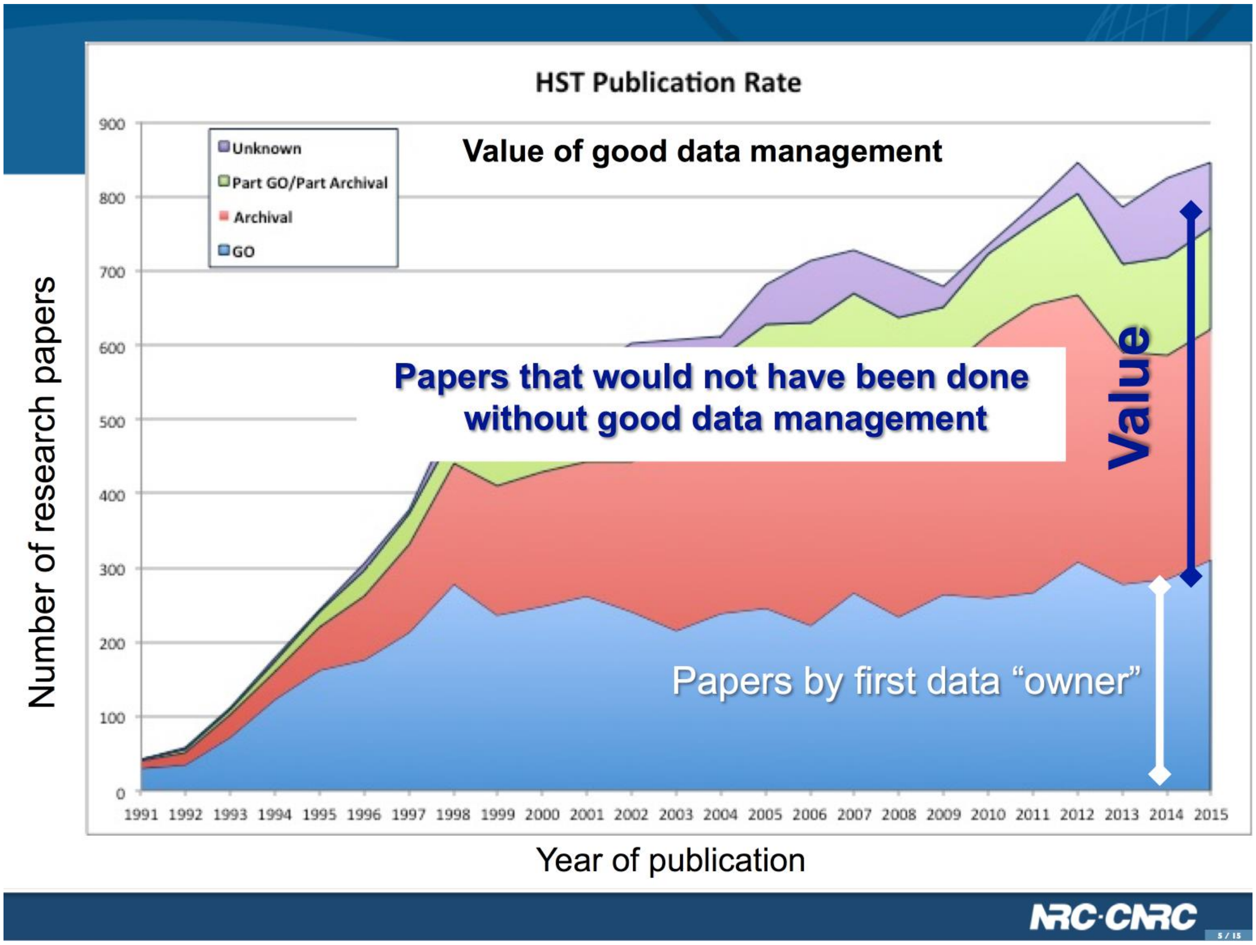
Are there quantitative examples of how effective RDM has strengthened scientific impact that speak to the great potential?

In much the same way as it would have been difficult for early proponents of the web to quantify its profound impact on all of us, one of the challenges that we have in answering this question is, of course, that Canada's RDM practice is still nascent and our infrastructure/services are not yet fully developed. As such, it is hard to quantify impact in the absence of the metrics that a more mature RDM platform would generate.

What we do know, however, is that RDM is essential to realizing the benefits of our new data-rich environment -- if this powerful resource is not managed properly and researchers are not supported in their use of it, we could find ourselves data rich, but information poor, losing all of the promise and potential that these data offer. RDM is critical to good public stewardship of the significant investment in research that governments make on behalf of Canadians. It enables appropriate dissemination of the data that underpins evidence-based research, which in turn supports reproducibility and validation of research and the sharing of data within and between disciplines.

We can also point to some quantitative examples from early adopter communities to demonstrate impact. At the recent Research Data Canada National Data Services Framework Summit (September 2017), one of the most compelling demonstrations of the impact and value of RDM and data sharing came from the Canadian Astronomy Data Centre. The speaker, David Schade[9], showed a slide illustrating data-sharing behaviour in the astronomy community. This slide shows not only growth in numbers of academic papers published by 'first data owner' but, most importantly, growth in the number of papers published by others who found and used these data.

[9] https://zenodo.org/record/1035837



There are other international examples of organizations that have tried to quantify the financial/efficiency impact of RDM by using open data as a measure. For example, in a 2016 review of the open data made available by the European Bioinformatics Institute (EMBL-EBI)[10], the authors noted the following impacts from making molecular data and services public and accessible to anyone:

- ○ *A contribution to the wider realization of future research impacts worth **£920 million every year**; and,*
- ○ *Annual direct efficiency impact estimated at between **£1 billion and £5 billion** per annum.*

In addition, in a recent report, the Open Data Institute suggests[11] that making all forms of data openly accessible (research data, as well as government and industrial data) would stimulate research and innovation by approximately 0.5% of GDP. For Canada that represents a **$90 billion impact annually**.

As RDM practice grows in Canada, RDMC would develop additional metrics and measures to help show the impact of RDM on research and societal outcomes.

[10] https://www.ebi.ac.uk/about/news/press-releases/value-and-impact-of-the-european-bioinformatics-institute

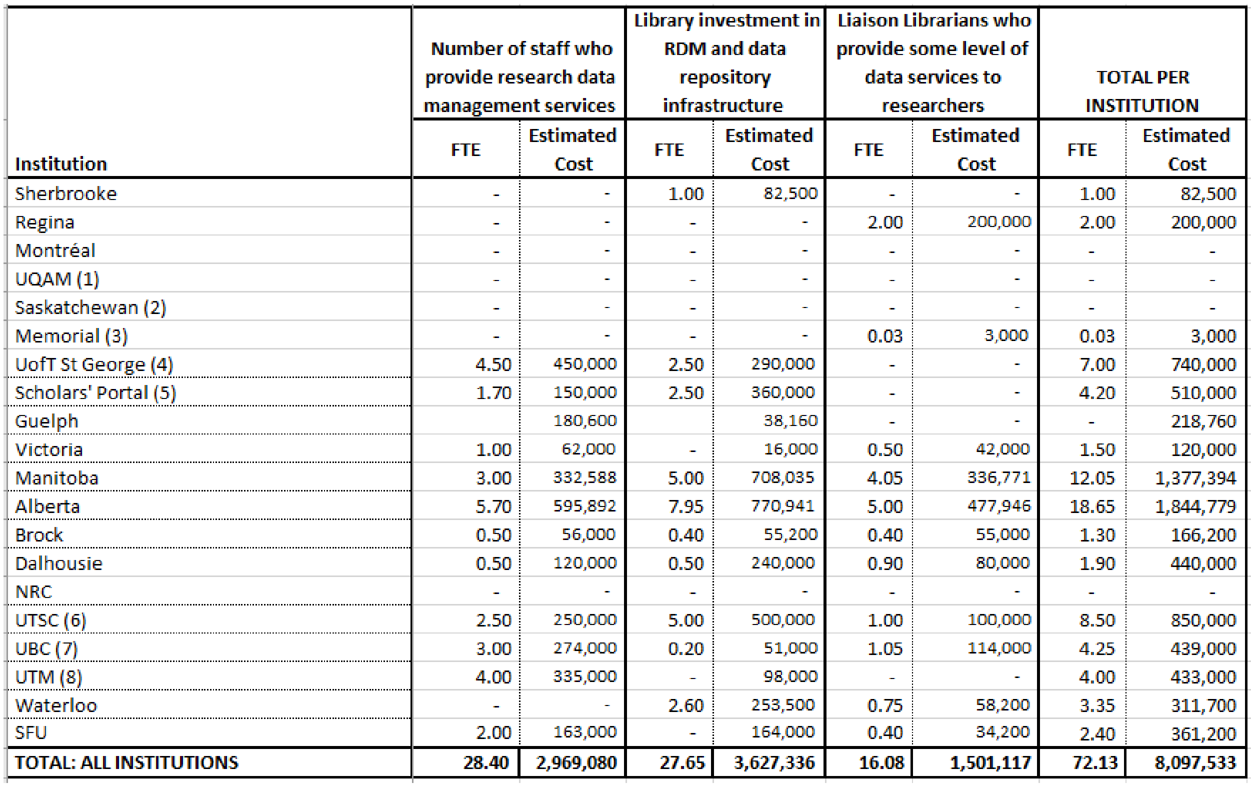
[11] http://theodi.org/research-economic-value-open-paid-data

Question 7: Current Regional/Institutional Investments

"It should be stressed that the majority of the investment that is required to support RDM in Canada would continue to be made by regional consortia and individual universities" (p. 36). What are the current total levels of investment from institutions and regions/provinces in RDM? And how much would they be expected to contribute under this proposal?

The current levels of investment are not easily accessible, partly due to the lack of national coordination of the ecosystem, but we can illustrate some of that investment through examples.

Universities and university libraries across Canada are actively supporting RDM on their campuses. A 2016 Survey of CARL libraries estimated that for the 19 institutions who responded, there were a total of over **70 FTEs** contributing, directly or indirectly, to RDM and significant investment in RDM and data repository infrastructure, combining to make an estimated expenditure of over $8 million annually. It is important to stress that these figures do not reflect the 10 CARL institutions who did not reply, or take into account spending by non-CARL institutions[12] (or other non-academic organizations) who were not polled by this survey, regional organizations, or discipline-specific communities



Institution	Number of staff who provide research data management services		Library investment in RDM and data repository infrastructure		Liaison Librarians who provide some level of data services to researchers		TOTAL PER INSTITUTION	
	FTE	Estimated Cost	FTE	Estimated Cost	FTE	Estimated Cost	FTE	Estimated Cost
Sherbrooke	-	-	1.00	82,500	-	-	1.00	82,500
Regina	-	-	-	-	2.00	200,000	2.00	200,000
Montréal	-	-	-	-	-	-	-	-
UQAM (1)	-	-	-	-	-	-	-	-
Saskatchewan (2)	-	-	-	-	-	-	-	-
Memorial (3)	-	-	-	-	0.03	3,000	0.03	3,000
UofT St George (4)	4.50	450,000	2.50	290,000	-	-	7.00	740,000
Scholars' Portal (5)	1.70	150,000	2.50	360,000	-	-	4.20	510,000
Guelph		180,600		38,160	-	-	-	218,760
Victoria	1.00	62,000	-	16,000	0.50	42,000	1.50	120,000
Manitoba	3.00	332,588	5.00	708,035	4.05	336,771	12.05	1,377,394
Alberta	5.70	595,892	7.95	770,941	5.00	477,946	18.65	1,844,779
Brock	0.50	56,000	0.40	55,200	0.40	55,000	1.30	166,200
Dalhousie	0.50	120,000	0.50	240,000	0.90	80,000	1.90	440,000
NRC	-	-	-	-	-	-	-	-
UTSC (6)	2.50	250,000	5.00	500,000	1.00	100,000	8.50	850,000
UBC (7)	3.00	274,000	0.20	51,000	1.05	114,000	4.25	439,000
UTM (8)	4.00	335,000	-	98,000	-	-	4.00	433,000
Waterloo	-	-	2.60	253,500	0.75	58,200	3.35	311,700
SFU	2.00	163,000	-	164,000	0.40	34,200	2.40	361,200
TOTAL: ALL INSTITUTIONS	**28.40**	**2,969,080**	**27.65**	**3,627,336**	**16.08**	**1,501,117**	**72.13**	**8,097,533**

Table 1. Results of CARL Portage Survey of CARL Libraries

[12] CARL members are from the 29 most research-intensive of Canada's 96 universities. But research generating takes place at all 96 universities, as well as at many of Canada's over 124 colleges and institutes, as well as by governments and the private sector.

It is understood that investments by RDMC will augment and complement, not replace, existing local commitments. In fact, it is expected that institutional, regional, and provincial investment in RDM will continue to grow under RDMC. We expect the RDM trends discussed in Question 5, including implementation of the anticipated Tri-Agency policy on RDM, will drive this investment, as researchers and institutions will be responsible for addressing these new requirements. Investments in RDMC will help facilitate and coordinate local, regional, provincial, and national RDM activities (through cooperative governance, services, tools, platforms, standards-development, etc.) to ensure a consistency of practice and the interoperability of infrastructure and services across Canada and internationally, but it will not replace the significant financial commitments and investments made by institutions, regional and discipline-specific organizations, or provincial governments.

Question 8: Breakdown of 5-Year Budget

Is a breakdown of the five-year budget available? Would be valuable to see more details on the funding for RDM tools, RDM platforms, and Archival Storage.

The structure of the five-year budget for RDMC was based on the costs associated with delivering the five core functions identified in the LCDRI DM report (see Table 2). Projections were made for items directly related to this goal, including RDMC-funded expert positions through the NOE, contractual positions, infrastructure leasing, platforms and tools development, archival storage costs, and funding to support the operations of the RDMC Secretariat, including associated HQP FTEs.

Table 3 provides a breakdown of the HQP that would form the core of the Network of Experts (see Question 2) and RDMC Secretariat personnel. Experts would be distributed across Canada’s four regions and local institutions, according to the approach outlined above. These positions are expected to increase over the five-year budget cycle as a culture of RDM and our understanding of researcher and front-line service needs grows.

Close to sixty percent of the proposed budget is directed toward facilitating the development and support of RDM tools and platforms. An RDM tool is an application that supports a particular task within a data workflow. Canadian examples of RDM tools include the proposed Format Policy Register, which will validate community-preferred formats in a preservation processing workflow or a metadata application that a researcher can use in her or his workflow. An RDM platform provides dedicated functions for one or more stages of the data lifecycle. Examples of such platforms in Canada include the DMP Assistant, which helps researchers with data management planning throughout the research lifecycle, and the Federated Research Data Repository, which provides a storage and dissemination option for researchers at later stages of the lifecycle, and enables discovery and reuse of data at a national (and international) scale. The funding for RDM platforms and tools is to cover the costs of developing, leasing or sharing, maintaining, and upgrading these essential resources. A poll of key infrastructure providers was used to determine the amounts for this line item.

The proposed budget also includes a line item for archival storage. These funds will supplement the investments that new and existing preservation storage providers (PSPs) are making to protect digital collections across Canada.[13] Budget figures for archival storage were estimated from a model developed with Compute Canada based on their active storage needs. This line item includes a one-time allocation to address the backlog of archival storage needs for RDM – we not only need to plan for future archival storage, but also to address the storage requirements of data already produced: RDMC will have a more accurate figure for the amount of retrospective data that need to be archived by year 3.

This budget model provides costs for RDMC's contribution to the broader RDM community, and, as discussed elsewhere in this document, is built on the assumption that contributions to Canadian RDM initiatives will continue to be made by multiple stakeholders in addition to RDMC, including institutions, discipline-specific communities, regional organizations, and provincial governments - successful implementation of RDM in Canada will depend on leveraging existing investments and coordinating and incentivizing others. We did not include these non-RDMC contributions in the proposed budget because the data to calculate actual estimates for stakeholder investments outside of RDMC were not available; gathering this data would be a key activity for RDMC in the first 2 years of its mandate.

Table 2: A Five-Year Budget of Research Data Management Canada

		FY1	FY2	FY3	FY4	FY5	TOTAL
Highly Qualified RDM Personnel							
	Network of Expertise	$1,099,150	$2,085,493	$3,375,519	$3,666,427	$4,297,305	$14,523,893
	Contracted Experts	$295,000	$295,000	$295,000	$295,000	$295,000	$1,475,000
	SUBTOTAL	$1,394,150	$2,380,493	$3,670,519	$3,961,427	$4,592,305	$15,998,893
Tools and Platforms							
	RDM Tools	$730,000	$892,050	$1,098,129	$1,360,271	$1,693,822	$5,774,271
	RDM Platforms	$2,196,500	$2,248,999	$2,055,655	$2,002,312	$2,310,192	$10,813,658
	Archival Storage	$4,320,000	$3,720,000	$2,677,500	$1,792,500	$2,595,000	$15,105,000
	SUBTOTAL	$7,246,500	$6,861,049	$5,831,284	$5,155,083	$6,599,014	$31,692,929
Secretariat Office							
	HQP	$849,250	$946,413	$987,795	$1,031,377	$1,077,295	$4,892,130
	Operational Expenses	$344,000	$237,240	$251,348	$266,386	$282,418	$1,381,392
	SUBTOTAL	$1,193,250	$1,183,653	$1,239,144	$1,297,762	$1,359,713	$6,273,522

[13] Examples of PSPs are COPPUL's West Vault and Ontario Scholars Portal's Ontario Library Research Cloud.

TOTAL		$9,833,900	$10,425,195	$10,740,946	$10,414,272	$12,551,032	$53,965,344

Table 3: A Five-Year Projection of Highly Qualified RDM Personnel & Secretariat HQP

		FY1	FY2	FY3	FY4	FY5
Highly Qualified RDM personnel & Secretariat HQP						
	FTEs					
	Secretariat	7.5	8	8	8	8
	Network of Expertise	9.5	17.5	27.5	29	33
TOTAL		17	25.5	35.5	37	41

In addition, as noted in Question 3, we are suggesting that RDMC could play a role in administering funds and facilitating activities related to a number of key, federally-funded repositories that are at risk.

Question 9: RMI Functions and RDMC

Could RMI functions be delivered by RDMC?

While RMI and Research Data (RD) are both critical to the research lifecycle, they differ significantly in purpose and delivery requirements. RD directly supports researchers in undertaking scientific enquiry, scholarship, or artistic activity. RMI, on the other hand, enables administrative workflows and activities, such as funding applications, evaluation, reporting, compliance monitoring, and impact assessment. As a result, while there is overlap in their users – researchers, for instance, will use both RD and RMI over the research lifecycle - the approaches to collection, storage and sharing will differ in scope and required expertise.

Even with these differences, the intersection between RMI and RD is important. Only a subset of content in the overall RMI landscape is also applicable to RDM, but, when used to enrich RD metadata, that subset is powerful in achieving the aims of FAIR data. Examples of RMI that intersect with RD in this way include: Data Management Plans, some parts of project descriptions and CVs, and the unique identifiers like ORCIDs, DOIs and ISNIs that tie things together.

In answering this question, it is also important to note that there is a difference between "functions" and "facilitation and coordination". The "functions" required for successful RMI must be implemented and owned within and across many communities and types of organizations (including institutions and funders but also software suppliers and publishers) both domestically and internationally. The "facilitation and coordination" of so many diverse and independent stakeholders is most effectively accomplished through neutrally-governed and collaborative not-for-profit organizations with designated mandates for accomplishing this goal. In the case of RMI, these organizations would "coordinate and facilitate" the development and maintenance of shared standards (e.g. CASRAI) and shared software services (e.g. ORCID).

From the perspective of "functions", RDMC would have a very important role to play in the above-noted RMI ecosystem in terms of promotion of best practice for the 'enriched metadata' uses of RMI. RDMC would also work with the wider RMI community to ensure that RMI and RD infrastructures and services integrate easily with one another and provide seamless access points for researchers.

In closing, given the differences in: a) purpose; b) primary audiences (i.e. administrator vs. researcher); and c) that RDM-related functions only intersect with a small fraction of the complete RMI landscape, it would not be efficient or effective for RDMC to assume the role of delivering RMI functions.