November 2020

The Current State of Research Data Management in Canada

An Update to the LCDRI Data Management Position Paper

Authored by the Alliance Research Data Management Working Group:

Shahira Khair, Rozita Dara, Susan Haigh, Mark Leggott, Ian Milligan, Jeff Moon, Karen Payne, Elodie Portales-Casamar, Ghilaine Roquet, and Lee Wilson

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

Funded by the Government of Canada | Canada

Table of Contents

Acknowledgements

In May 2020, a Working Group on Research Data Management was struck to update the LCDRI position paper summarizing the current state of Research Data Management in Canada. The working group met bi-weekly to oversee the analysis and writing of the report.

The Working Group sought feedback from the community on its work as it evolved. We would like to thank the following groups for sharing their expertise: Alliance's Senior Analysts and Researcher Council, the First Nations Information Governance Centre, the Portage Coordinators, and the wider research data management community convened by Portage and Research Data Canada in their respective communities of practice.

1 Executive Summary

This report serves as an update to the 2017 Data Management Position Paper submitted to Innovation, Science and Economic Development Canada (ISED) by the Leadership Council for Digital Research Infrastructure (LCDRI), and as reflected in the public summary.[1] The report summarizes the Research Data Management (RDM) landscape in Canada, and documents challenges and opportunities for the current RDM ecosystem and Digital Research Alliance of Canada (the Alliance). The intent of this work is to position the Alliance to build on the current state and chart a path forward that advances RDM in coordination with other digital research infrastructure (DRI) elements to support research excellence.

The term "research data" refers to any information created or collected as evidence in the research process or commonly accepted in the research community as necessary to validate results and conclusions.[2] Research data are valuable assets, which when properly managed, have the potential to be reused and recombined in innovative ways to derive greater value and advance research and scholarship. The management of this data draws upon a range of infrastructures and skill sets to support its documentation, storage, access, and preservation over the course of a research investigation and following its conclusion. While benefiting the original research from which data are derived, the broader potential and objectives underlying RDM are rooted in the larger movement for Open Science that presents a vision for accelerated scientific discovery and advancement enabled by new information technologies, which will allow research and underlying data to be reviewed, communicated, shared, and reused more openly and accessibly.

Analyzing RDM as a unique pillar of DRI, in isolation from Advanced Research Computing (ARC) and Research Software (RS), is helpful for analyzing its functions, needs, and impact on digital research; however, in reality the DRI landscape is more complex. Rather, these pillars should be viewed as interconnected, enabling and deriving support from each other. As noted in the LCDRI's 2017 coordination paper, these individual components only function to their true potential when they are integrated to facilitate efficient and effective workflows for researchers. Effective management of digital research data relies on a robust array of supporting infrastructure that includes ARC and RS components. Inversely, effective use of ARC and RS requires that the research data they support be maintained over their entire lifecycle with effective management practices, supporting more complex investigations and deriving greater value and impact of investments in DRI.

The growth over time of initiatives, partnerships, networks, and supporting organizations has given rise to an increasingly mature, albeit complex, Canadian RDM landscape. Canadian researchers draw upon diverse resources throughout the lifecycle of a research project, which has led to a plurality of groups and organizations engaged in the stewardship of research data assets. The range of actors involved contribute to the evolution of the RDM ecosystem by

[1] Baker, D. et al. (2019). Research Data Management in Canada: A Backgrounder. Zenodo. http://doi.org/10.5281/zenodo.3341596

[2] "Research data", CASRAI Glossary. https://casrai.org/term/research-data (Retrieved November 2020)

advancing components through a range of platforms, services, guidance, and the research practice itself.

Since the work of the LCDRI, the efforts of the RDM community have continued to advance at a national level through the efforts of the Canadian Association of Research Libraries (CARL), Portage Network, CANARIE, and Research Data Canada (RDC), which have succeeded in advancing complementary agendas. The CARL Portage Network has coordinated efforts of institutional partners from across Canada, bringing together a network of experts to develop platforms, services, and guidance, providing practical support for RDM to Canadian higher education institutions. RDC has fostered opportunities for national dialogue, convening a wide range of stakeholders to discuss and coordinate a framework for national data services. Building upon the successes of these two organizations as they are integrated into the Alliance will require maintaining awareness of the diversity of existing actors and their roles.

Current State Summary

This report summarizes key entities in the Canadian landscape at local, national, and international scales, and organizes the multiple roles they play in the ecosystem into key components necessary to support RDM nationally.

Storage and Compute

From an RDM perspective, there are three distinct configurations of computing and storage infrastructure to support distinct stages of data throughout its lifecycle: Active, Repository, and Archival. These differ by purpose, practices, and the level of curation required to support containing data. In the active research phase, computing and storage infrastructure address the needs of data through the research process itself, while data are being collected, manipulated, or analyzed. Researchers will generally have available to them storage infrastructure provided by their home institutions, as well as that which they seek out from publicly funded or commercial service providers. Availability will greatly depend on institutional capacity. For this reason, institutions and governments have sought to fund research infrastructure regionally and nationally to provide more equitable access. Thus, researchers based at higher education institutions across Canada also have access to infrastructure supporting storage of data in the active research phase through Compute Canada, and the National Research Education Network.

Practices around data sharing vary greatly by research domain. In general, at the conclusion of an investigation or funded project, researchers make curatorial decisions around data retention and sharing to meet institutional, funder, or community expectations, support reproducibility of published findings, and advance future research needs in their fields. This stage of the research lifecycle relies on repository storage, which supports future discovery and appropriate access to research data. Some academic libraries across the country provide access to institutionally managed repositories designed to support many research data types from a wide user community. In recent years, libraries have begun pooling their resources through regional and national associations to develop shared repository platforms, improving access to infrastructure across academic institutions of varying local capacity (e.g., Scholars Portal Dataverse, the Federated Research Data Repository). As the diversity of available multidisciplinary repositories and related service models grows within and outside of institutions, clarifying this overwhelming complexity will be necessary to support both researchers and their data.

A range of research domain-specific data repositories operated by research groups and organizations located in Canada and internationally also play a significant role in the national RDM ecosystem. Many domain repositories are endorsed by their respective research communities, which means that the data they contain is more likely to be discovered and trusted by researchers in respective fields. Also, since the data these repositories contain reflects a narrow range of subject matter, domain-specific practices can be applied, increasing opportunities for interoperability between distinct datasets, as well as reusability, leading potentially to greater impact.

A significant gap in the Canadian RDM landscape is the availability and accessibility of archival storage to support the longer-term preservation of research data. Limited archival storage infrastructure for research data is currently supported at local and regional levels. A national strategy for archival storage that builds on existing initiatives can leverage the role of decentralization in risk mitigation as part of responsible long-term preservation. To achieve efficiencies in a decentralized model, national coordination among organizations would be necessary to oversee storage and preservation.

Interoperability

Achieving interoperability between components of the RDM ecosystem relies on common schemas, standards, and protocols for collecting, organizing and describing research data and supporting infrastructure. In order to maximize the potential of research data, it must be able to be exchanged securely and integrated between different systems, while being interpreted correctly and appropriately by different users. To support both semantic and technical interoperability, operating frameworks are required that define the procedures, terms, and relationships necessary to allow data to be exchanged unencumbered between digital research infrastructures. These provide an architecture to the ecosystem, which allows new data, software, and infrastructure to be developed and integrated by conforming to these existing frameworks.

Data Services

As the RDM needs of researchers have grown in scale and complexity in response to advances in technology and research practices, as well as new expectations from home institutions, funders, and publishers, a range of related support services have been developed, which span the research data lifecycle. A growing number of services are being offered within academic research institutions, through infrastructure providers in association with the specific platforms and tools they offer, via regional and national associations, and increasingly by commercial entities. These services are broadly organized around the research data lifecycle, including support for data management planning, curation, preservation, discovery, and exploration. As more research groups and organizations develop their own platforms supporting RDM, services supporting this infrastructure are also in development.

The digital shift in the research enterprise has resulted in significant needs for training in post-secondary institutions related to the adoption of good RDM practices. While the gap in RDM skills is reducing through the efforts of higher education institutions to invest in more training for researchers, capacity for this work is mostly concentrated in large universities. Funding agency mandates (e.g., the Tri-Agency draft RDM Policy) are one set of mechanisms to encourage growth among institutions. Research associations and academic societies also have important roles to play in encouraging skill development and in scaling the provision of training opportunities.

Governance

Many organizations have assumed roles supporting their communities of practice with RDM, through the development of guidance, policies, or funding opportunities. Coordination between these organizations is essential for fostering a diversity of successful approaches to supporting RDM. Consistent policies and requirements for research organizations, research infrastructures, and related services are necessary to ensure that researchers adopt common practices and frameworks, which in turn will enable systems supported nationally to respond to the necessary governance models. Any differences in institutional and regional requirements may contribute to challenges nationally. Within this current landscape, impacts of existing imbalances that exist locally and regionally should be considered in the national context. Harmonisation with international initiatives should also be considered to allow data to move across borders.

Key Challenges and Opportunities

The scale and growth at which research data are being generated, combined with the diversity of needs and interests poses significant challenges for sustainably supporting RDM at the national scale. Sustainable funding models that address the longer-term needs of RDM are currently limited across many disciplines. Before such funding models can be advanced, clarification regarding what is covered by ARC, DM, and RS envelopes within the Alliance is needed. In tandem, fundamental distinctions between DRI ecosystem components must also be clarified.

While this report presents a high-level overview of the range of ecosystem actors, infrastructures, and services supporting RDM in Canada, it is an incomplete picture that requires refinement. Many of the existing infrastructures, tools, and platforms operate in relative isolation from one another. Better integration among new and existing services and infrastructures requires the adoption of shared standards, schemas, and certifications for trusted interoperability. In parallel, many actors within the Canadian RDM ecosystem exist in relative isolation. Continued consultation and outreach efforts are necessary to understand their needs. Mechanisms are needed to ensure that both providers and users from all sectors and domains are represented and supported, with special consideration and accommodation to promote participation of under-represented voices and forms of non-western research. Alignment with changes in the policy landscape of research institutions, funders, and publishers is one mechanism to lead to greater adoption.

There is a need for further alignment and integration of organizations and services, not only Canadian entities supported by the Alliance, but also of their international counterparts. Determining how these partners in the RDM ecosystem fit together is an important step in furthering collaborative innovation, improving RDM support, and reducing overlap and duplication of efforts. An ideal system would consist of the provision of services at a range of levels, supported and structured through a national framework that is linked to and influenced by international standards and peer organizations.

2 Introduction

Many key milestones encompassing local, regional, national, and international efforts mark the advance in Canada towards recognition of research data management (RDM) as an essential component of Digital Research Infrastructure (DRI). Initiatives including Canada's Advisory Panel for the Review of Federal Support for Fundamental Science,[3] national funding policies such as the Tri-Agency Statement of Principles on Digital Data Management,[4] and groups such as the Leadership Council in Digital Research Infrastructure (LCDRI) point to the role of RDM in the DRI ecosystem and its importance in advancing Canadian research excellence in an increasingly data and computing-intensive research environment.

The growth over time of initiatives, partnerships, networks, and supporting organizations has given rise to an increasingly mature, albeit complex, Canadian RDM landscape. This progress came to a head in November 2016, when Innovation, Science, and Economic Development Canada (ISED) provided funding to the LCDRI to convene working groups composed of a broad set of stakeholders to undertake an analysis of the DRI landscape in Canada and produce position papers on the state of Data Management, Advanced Research Computing, and recommendations for national coordination. The LCDRI's 2017 Data Management Position Paper articulates a vision for national coordination and facilitation of RDM, based on an in-depth analysis of the current state of RDM in Canada. This assessment of the current state is updated annually through a joint effort of the national RDM community.[5]

Since the work of the LCDRI, the efforts of the RDM community have continued to advance at a national level through the efforts of the Canadian Association of Research Libraries (CARL) Portage network to develop shared resources, expertise, and training to support a national community of practice,[6] and through a series of National Data Services Framework Summits (NDSF) convened by Research Data Canada.[7] The Kanata Declaration, an integrating outcome of the NDSF, presents a community-based vision for a national RDM strategy and priorities for data services.[8]

Following the 2018 Federal Budget that committed $572.5M towards a national DRI strategy,[9] ISED announced a contribution program in Spring 2019 to fund a new not-for-profit DRI

---

[3] Naylor, C.D. et al. (2017). Investing in Canada's Future: Strengthening the Foundations of Canadian Research. http://www.sciencereview.ca/eic/site/059.nsf/eng/home

[4] Government of Canada. (2020). Tri-Agency Statement of Principles on Digital Data Management. https://www.ic.gc.ca/eic/site/063.nsf/eng/h_83F7624E.html (Retrieved November 2020)

[5] Baker, D. et al. (2019). Research Data Management in Canada: A Backgrounder. Zenodo. http://doi.org/10.5281/zenodo.3341596

[6] https://portagenetwork.ca

[7] https://www.rdc-drc.ca

[8] Attendees of the NDSF Summit. (2019). Kanata Declaration. Zenodo. http://doi.org/10.5281/zenodo.3234815

[9] Innovation, Science, and Economic Development Canada. (2019). Digital Research Infrastructure. https://www.ic.gc.ca/eic/site/136.nsf/eng/home (Retrieved November 2020)

organization to coordinate funding and strategic directions for national DRI activities related to Advanced Research Computing (ARC), Data Management (DM), and Research Software (RS). In response, members of several stakeholder organizations submitted a DM Roadmap for 2019-2024 to support ISED in this process, which proposed key functions and activities for a national RDM organization to advance in response to a shared vision.[10]

The national attention placed on RDM also reflects shifts in research culture and practices, led not by national organizations but through the efforts of researchers in a range of data intensive fields of study. In Canada, these mostly notably include research groups in astronomy and astrophysics, high-energy physics, earth and ocean sciences, the -omics branches of biosciences, and the digital humanities. Research groups and organizations across the country are advancing tools and platforms for managing, analyzing, and sharing data among academics within and across complex research projects. Advancing their work while finding ways of widely applying their successes to other research fields is necessary in providing truly national data services. This will also require connections with a range of international organizations with both broad and domain-specific focuses that are advancing practices within research domains, all with strong Canadian connections.

While many efforts have begun to coalesce at national levels, research is conducted at an international scale. Canada is not alone in recognizing the need for national support for DRI and RDM. Implementing common frameworks for platforms and services to ensure that research data can move across national and pan-national infrastructures will also be required to support researchers across Canada.

While national support for RDM in Canada is currently advanced by a core group of organizations, their efforts engage a range of collaborators and service providers who support researchers and their data throughout its lifecycle. The speed at which this landscape is evolving means that a refined understanding of existing actors, infrastructure, and services is crucial for future investment that benefits the research community. This report summarizes this complex landscape and describes current supports according to four key areas: Storage and Compute, Interoperability, Data Services, and Governance. From this review, key challenges and opportunities that must be addressed for national support for RDM in Canada to be successful are analyzed. In doing so, this report aims to position Digital Research Alliance of Canada (the Alliance) to build on preceding and current initiatives and chart a path forward that advances RDM in coordination with other DRI elements to support national research excellence.

Defining “Research Data”

This report adopts an inclusive definition of the term “research data” provided by CASRAI. *“Data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results. All other digital and non-digital content have the potential of becoming research data. Research data*

[10] Castle, D. et al. (2019). Position Paper – Data Management Roadmap: 2019-2024. https://www.rdc-drc.ca/download/position-paper-data-management-roadmap-2019-2024 (Retrieved November 2020)

may be experimental data, observational data, operational data, third party data, public sector data, monitoring data, processed data, or repurposed data."[11]

While many types of data can be considered as research data, the set of practices that can be applied to its management depend on a range of factors related to including how it was gathered and for what purpose, as well as the subject matter being described. For example, administrative data derived from the operation of administrative systems capturing data about programs, their operations and related subjects is heavily relied upon in various research fields, particularly the health sciences and social sciences. Alternatively, online services data covers a range of sources, from search engines to online transactions and communications, and can be used to answer a range of research questions. While researchers may be granted access to these sources of data for research purposes, their ability to curate, store, or preserve these sources is more limited compared to research data that is collected firsthand via observation or experimentation. Therefore, engagement and coordination of data producers and users across the research landscape in the adoption of strong data management practices is essential for sustaining the research enterprise.

The sources from which research data are derived also determine how they should be managed, in accordance with community expectations and any associated ethical, legal, or commercial obligations. In particular, data that relate to First Nations, Inuit, or Métis communities, including their peoples, territories, and cultures, whether generated directly through research activities or derived from secondary sources and used for research purposes, must be managed in accordance with the data management principles developed and approved by these communities, in respect of Indigenous data sovereignty.

As well, various research communities interpret the definition of research data differently, influencing their own practices and attitudes towards RDM. Fostering a research culture in Canada that values the management of research data in all forms will require an interdisciplinary approach that supports domains with both advanced and more limited experience working with digital research data.

[11] "Research data", CASRAI Glossary. https://casrai.org/term/research-data (Retrieved November 2020)

Impact and Value of RDM

Figure 1. Data management functions (inner circles) overlaying data lifecycle phases (outer circles). In *Data Management Position Paper: For Innovation, Science, and Economic Development Canada*. Leadership Council for Digital Research Infrastructure. Unpublished manuscript. August 31, 2017.

RDM refers to the documentation, storage, access and preservation of data produced over the course of a given investigation.[12] Data management practices cover the entire lifecycle of

[12] "Research data management", CASRAI Glossary. https://casrai.org/term/research-data-management (Retrieved November 2020)

research data, from planning the investigation to conducting it, from backing up data as it is created and used by its creators to documenting and describing data in preparation for sharing and reuse by collaborators, and finally the preservation of digital materials after the research investigation has concluded. See Appendix A for a more fulsome review of the RDM lifecycle.

The information age has brought on a flood of data – where all major research areas have become significant producers and consumers of research data. Therefore, increased capacity must be built to both manage what is currently being produced and to address future growth in data production. This includes support for the FAIR principles, in addition to considerations of longer-term preservation. Supporting researchers during the active research phase, when curation processes could reduce volume and improve quality is also an important consideration. For example, Calcul Québec roughly estimates that up to one eighth of their storage capacity is composed of redundant files (e.g., duplicate, temporary, or derivative files).[13] Well-funded initiatives that are supported by a nationally coordinated model will be essential to meeting the many challenges posed by increasingly data-driven research.

Growing efforts of academic research institutions across Canada are supporting researchers in improving their management of research data assets.[14] Organizations that support scholarship, including funders and scholarly publishers, are also recognizing the potential benefits that increased preservation, access, and openness of research data can produce, from improved reliability of published results to potential for reuse. For instance, a range of support and training programs are being developed by these institutions to support researchers with data management planning, publication, and preservation.[15, 16] Notwithstanding, significant barriers and challenges exist in translating this top-down support into real solutions for the complex and diverse range of issues researchers face in creating, processing, and analyzing their data in a way that enables sharing and reuse by their peers.

Putting aside for a moment the very real (and costly) gaps and challenges that make RDM so difficult, what do we stand to gain from efforts to overcome these hurdles? The “Tri-Agency Statement of Principles on Digital Data Management” underscores the importance of effective research data management in support of this goal:

> *Research data are gathered through a variety of methods, including experimentation, analysis, sampling, and repurposing of existing data. They are increasingly produced or translated into digital formats. When properly managed and responsibly shared, these digital resources enable researchers to ask new questions, pursue novel research programs, test alternative hypotheses, deploy innovative methodologies and collaborate across geographic and disciplinary boundaries. The ability to store, access, reuse and build upon digital research data*

[13] Talon, S., personal communication, Sept. 2020

[14] Cooper, A. et al. (2020). Institutional Research Data Management Services Capacity Survey. http://doi.org/10.14288/1.0388722

[15] Tenopir, C. et al. (2014). Research data management services in academic research libraries and perceptions of librarians. *Libr Inf Sci Res*, 36(2):84-90. https://doi.org/10.1016/j.lisr.2013.11.003

[16] Perrier, L. et al. (2017).Research data management in academic institutions: A scoping review. *PloS One*, 12(5), e0178261. https://doi.org/10.1371/journal.pone.0178261

has become critical to the advancement of science and scholarship, supports innovative solutions to economic and social challenges, and holds tremendous potential for Canada's productivity, competitiveness, and quality of life.[17]

Benefits of Data Management, Sharing, and Reuse

Accelerates scientific progress: *Data sharing allows researchers to access and understand others' data and re-use them for their own scientific purposes, thereby speeding up the rate of new discoveries, and preventing unnecessary expensive data collection.*

- Numerous research organizations in Canada have recognized this potential and have adopted open research practices (e.g., Tanenbaum Open Science Institute (McGill U.),[18] Centre for Biodiversity Genomics (U. Guelph)).[19]

- Milham et al. (2018) use the International Neuroimaging Data-Sharing Initiative to provide direct evidence for the impact of data sharing on the scale of related studies. They estimate that for the nearly 1,000 papers included in their analysis, the saved cost of de novo data generation are between $893M to $1,707M.[20]

- Figures published by the National Research Council estimate a greater number of articles are currently being published using preserved archival data from the Hubble Space Telescope than are published by new observations.[21]

Enhances collaboration: *Enables researchers to collaborate with each other by sharing data, research environments, and tools.*

- Open data enables recombination of data from heterogeneous sources spanning multiple times and places to ask new questions.[22]

- When data are created, organized, described, and preserved using the same standards, they become more interoperable and can be integrated into common tools.

    - For example, by taking advantage of common neuroscience data formats, the McGill Centre for Integrative Neuroscience is developing software tools and

[17] Government of Canada. (2020). Tri-Agency Statement of Principles on Digital Data Management. https://www.ic.gc.ca/eic/site/063.nsf/eng/h_83F7624E.html (Retrieved November 2020)

[18] https://www.mcgill.ca/neuro/open-science

[19] https://biodiversitygenomics.net

[20] Milham, M.P. et al. (2018). Assessment of the impact of shared brain imaging data on the scientific literature. *Nat. Commun.*, 9(1):1-7. https://doi.org/10.1038/s41467-018-04976-1

[21] Baker, D. et al. (2019). Research Data Management in Canada: A Backgrounder. Zenodo. http://doi.org/10.5281/zenodo.3341596

[22] Whitlock, M. C. (2011). Data archiving in ecology and evolution: best practices. *Trends Ecol. Evol.*, 26(2):61-65. https://doi.org/10.1016/j.tree.2010.11.006

platforms that are openly available for use by the Canadian and international research community.[23]

- A 2016 review of the open data made available by the European Bioinformatics Institute estimates a direct efficiency impact of between £1 billion and £5 billion per annum.[24]

Increases visibility and impact of research: *Data made discoverable and accessible through a data repository can dramatically increase the impact of that research.*

- Publishing research data has been associated with higher citation rates. For instance, publications from clinical trials that shared underlying data were found to be cited up to 70% more frequently than those that did not.[25]

- Initiatives like the Federated Research Data Repository (FRDR) drive increased visibility of repositories and their data, and present a national snapshot of Canadian data assets.[26]

Enables reproducibility of research results: *When data are archived and shared, results can be re-examined and data can be used for re-analysis, thereby improving reproducibility and trustworthiness of published results.*

- There are both tangible and intangible impacts of the current reproducibility crisis:

    - Freedman et al. (2015) estimate that the total prevalence of irreproducible biomedical research in the U.S. exceeds 50%, resulting in $28B annually spent on preclinical research that is not reproducible.[27]

    - Meanwhile this lack of scientific reproducibility plays a significant role in reducing public trust in science (G7, 2019).[28] Results of a 2019 Pew Research poll surveying the American public's trust in scientific experts highlighted open data as a factor

[23] https://mcin.ca

[24] Beagrie, J. & Houghton. (2016). The value and impact of the European Bioinformatics Institute: Full report. https://beagrie.com/static/resource/EBI-impact-report.pdf

[25] Piwowar, H.A. et al. (2007). Sharing Detailed Research Data is Associated with Increased Citation Rate. *PLoS One,* 2(3), e308. https://doi.org/10.1371/journal.pone.0000308

[26] https://www.frdr-dfdr.ca/repo/

[27] Freedman, L. P., Cockburn, I. M., & Simcoe, T. S. (2015). The economics of reproducibility in preclinical research. *PLoS Biol*, 13(6), e1002165. https://doi.org/10.1371/journal.pbio.1002165

[28] Summit of the G7 Science Academics. (2019). Science and trust. https://royalsociety.org/-/media/about-us/international/g-science-statements/2019-g7-declaration-science-and-trust.pdf (Retrieved November 2020)

that increased trust in research.[29] Among the funding public, sharing data was even slightly more impactful on trust than independent peer review.

Research Data Management Through the Lens of COVID-19

On March 11, 2020 the coronavirus 2019 (COVID-19) acute respiratory disease was officially declared a global pandemic by the World Health Organization.[30] At the time of writing this report, the prevalence of digital data sources has allowed researchers to tackle this crisis from multiple angles, including: medical, social, and environmental. Compared to previous public health emergencies, such as the 2003 SARS outbreak, the contemporary digital research environment is enabled by greater connectivity and computing power, resulting in many advanced tools for sharing and analysis. There is also greater adoption of open science practices by researchers, enabled by open access policies such as the Tri-Agency Open Access Policy on Publications.[31] Taking full advantage of this landscape, the response of the international research community to help resolve this crisis provides a prime example for the impact of strong data management practices.

- International granting agencies have announced rapid response awards to fund COVID-19 related research, most with requirements for related data and publications to be shared openly.[32] In Canada, this includes rapid response funding programs from the Tri-Agency.[33]

- Research data repositories and curation services have joined forces to support research related to COVID-19. Most prominent is the Zenodo COVID-19 Community data repository, which is free and open to researchers worldwide to share research results that could be relevant for the scientific community.[34] Curation of these datasets is being supported by Europe's OpenAIRE program, who are also developing an online COVID-19 Gateway for connecting research data related to COVID-19 deposited across multiple repositories.[35] CARL-Portage have also brought together librarians and data curators to advance support for their institutions in managing research data related to COVID-19 by publishing a Guide to COVID-19 Rapid Response Data Sharing and Deposit for Canadian

---

[29] Funk C., et al. (2019). Trust and Mistrust in Americans' Views of Scientific Experts. Pew Research Centre. https://www.pewresearch.org/science/2019/08/02/trust-and-mistrust-in-americans-views-of-scientific-experts (Retrieved November 2020)

[30] World Health Organization. (2020). WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (Retrieved November 2020)

[31] Government of Canada. (2016). Tri-Agency Open Access Policy on Publications. https://www.ic.gc.ca/eic/site/063.nsf/eng/h_F6765465.html (Retrieved November 2020)

[32] Wellcome Trust. (2020). Sharing research data and findings relevant to the novel coronavirus (COVID-19) outbreak. https://wellcome.ac.uk/coronavirus-covid-19/open-data (Retrieved November 2020)

[33] Canadian Institutes of Health Research. (2020) Coronavirus: Canada's rapid research response. https://cihr-irsc.gc.ca/e/51890.html (Retrieved November 2020)

[34] https://zenodo.org/communities/covid-19

[35] https://www.openaire.eu/openaire-covid-19-gateway

Researchers.[36] Research communities have also responded to the crisis, for instance the iReceptor Project based at Simon Fraser University is making critical data available through new data repositories and an expedited curation pipeline.[37]

- Research communities are collaborating to develop best practices for the sharing and reuse of COVID-19 public health data. Canada's Chief Science Advisor mandated the creation of a national network to expedite communication and collaboration between the scientific, healthcare and policy communities during the COVID-19 crisis. The resulting CanCOVID is an expert network of Canadian COVID-19 researchers, clinical collaborators, and healthcare stakeholders from across the country.[38] Internationally, the Research Data Alliance's COVID-19 Fast track Working Group brought together international experts from across research domains to develop guidance for the management, sharing, and long-term preservation of COVID-19 related data.[39]

It is too soon to evaluate the full impact of these and related programs on the pandemic itself, however it marks a clear shift of expectations in research culture and the management of research data of national and international importance. While expectations can shift quickly, the ruling policy landscape developed by national stakeholders (e.g., TCPS 2),[40] to local institutions that influence how sensitive research data is managed, will take time to reflect, and there is an important opportunity for the Alliance to help coordinate and advance this consultation.

Open Science Movement

The values and objectives of RDM are rooted in the Open Science movement that presents a vision for accelerated scientific discovery and advancement enabled by new information technologies, which will allow research publications, results, and data to be shared openly and accessibly as part of a new social contract for science.[41] The Open Science movement is not limited to certain domains of research, but encompasses digital research and scholarship more generally, from the humanities to the physical sciences. Motivations for advancing Open Science range from value driven propositions of return on public investment in research and maximizing discoveries, to concerns about reproducibility and accountability, and to impacts of new

[36] Fry, J. et al. (2020). Guide to COVID-19 Rapid Response Data Sharing and Deposit for Canadian Researchers. Zenodo. http://doi.org/10.5281/zenodo.4270501

[37] https://cancovid.cahttp://ireceptor.irmacs.sfu.ca/repositories

[38] https://cancovid.ca

[39] RDA COVID-19 Working Group. (2020). RDA COVID-19 Recommendations and Guidelines on Data Sharing. https://doi.org/10.15497/rda00046

[40] Government of Canada, Panel on Research Ethics. (2018). Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS 2 (2018). https://ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2018.html

[41] "Open science", FOSTER Taxonomy. https://www.fosteropenscience.eu/taxonomy/term/7 (Retrieved November 2020)

collaborations and innovation.[42] Supporting this movement offers an opportunity to transform the entire research enterprise to further transparency, accountability, and public trust in science.

Irrespective of drivers and motivations, the realization of the Open Science vision relies on a foundation of robust and accessible DRI that enables the data underpinning research outputs to adhere to the FAIR principles of Findability, Accessibility, Interoperability and Reusability.[43] As the authors of the FAIR principles note, good data management is not an end goal in and of itself, but rather is a prerequisite that enables the various motivations driving the Open Science movement. For instance, for research data to be shared and reused widely by the academic community they should be described with rich metadata describing content, provenance, and limitations; their description, organization and structure should be harmonized and machine actionable; and they should be supported by systems that enable them to be uniquely discoverable, appropriately accessible, connected to related outputs, and preserved where appropriate for the long-term. Implementation of these practices is not a static procedure and can take many forms depending on the data itself and on the range of knowledge and skillsets among researchers, curators, system administrators and other stakeholders. Thus, good RDM should not simply be thought of as a box to be ticked or a secondary consideration, but rather as an integral part of conducting high quality research.

The relationship between RDM and Open Science is reflected in the federal government's recently published Roadmap for Open Science that provides a series of ten recommendations to guide Open Science activities in Canada.[44] Among these is the recognition that "open" data are not enough. Research data must be FAIR to maximize benefit; a prerequisite for which is strong management practices (see rec. 5). Considering the connections to DRI and RDM, a key leadership opportunity for the Alliance will be to support the Open Science movement on a national level, in partnership with relevant stakeholders (see rec. 9).

Indigenous Data Sovereignty

Indigenous data sovereignty recognizes the inherent rights of Indigenous communities to govern the collection, ownership, and use of their own data. This issue has a significant place in the national research data landscape and must be considered in the Alliance's efforts to advance RDM best practices in Canada. Acknowledging past malpractice concerning treatment of Indigenous communities in the research process, recognizing sovereignty of Indigenous communities over their own data, and advancing respect for the distinct data management practices of communities is an important component of RDM and supports reconciliation.

International Indigenous data sovereignty efforts recognize that Indigenous peoples must govern their own information, including research data, in alignment with their own interests. The CARE Principles for Indigenous Data Governance note that the current open science movement does

---

[42] Vicente-Saez, R. & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *J. Bus. Res*., 88:428-436 https://doi.org/10.1016/j.jbusres.2017.12.043

[43] Wilkinson, M.D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018. https://doi.org/10.1038/sdata.2016.18

[44] Office of the Chief Science Advisor. (2020). Roadmap for Open Science. http://science.gc.ca/eic/site/063.nsf/eng/h_97992.html (Retrieved November 2020)

not fully engage with the interests of Indigenous peoples and ignores power differentials and historical contexts, and as a result outline a set of principles to complement the FAIR Principles.[45] These principles speak to values of Collective benefit, Authority to control, Responsibility, and Ethics. The Principles are endorsed by the Global Indigenous Data Alliance (GIDA),[46] and complement the FAIR Principles, hence the phrase "*Be FAIR and CARE*".

Indigenous peoples in Canada also assert their own data governance principles:

First Nations

An important resource developed by First Nations communities, the OCAP® Principles are a tool to support strong information governance for First Nations data sovereignty.[47] OCAP® stands for Ownership, Control, Access, and Possession, and asserts that First Nations have control over data collection in their communities, that they own their data, and control how that information can be stored, interpreted, shared, and used. The Principles are expressed and asserted in line with a Nation's respective world view, traditional knowledge, and protocols.

Métis

The Manitoba Métis Federation subscribes to the OCAS principles of Ownership, Control, Access, and Stewardship.[48] The principles recognize the rights of Métis to own and make decisions over use of their data, and of ethical responsibility for proper planning and management of data resources.

Inuit

Inuit Qaujimajatuqangit (IQ) is a governing principle that applies to Inuit data. IQ translated literally means "*that which Inuit have always known to be true*" and is focused on integrating traditional Inuit culture into current governance structures and lessening the disempowerment of Inuit peoples.[49] Six guiding principles of IQ, as set out by Inuit Elders in a framework by the Government of Nunavut, include concepts of serving, consensus decision-making, skills and

---

[45] Research Data Alliance International Indigenous Data Sovereignty Interest Group. (2019). "CARE Principles for Indigenous Data Governance." *The Global Indigenous Data Alliance*. https://www.gida-global.org/care

[46] https://www.gida-global.org

[47] First Nations Information Governance Centre. (n.d.). Understanding OCAP®. https://fnigc.ca/ocap-training

[48] First Nations, Metis, and Inuit Health Research Strategic Planning Committee, University of Manitoba Faculty of Health Sciences. (2013). Framework for Research Engagementwith First Nation, Metis, and Inuit Peoples. https://umanitoba.ca/faculties/health_sciences/medicine/media/UofM_Framework_Report_web.pdf

[49] Tagalik, S. (2010). Inuit Qaujimajatuqangit: The role of Indigenous knowledge in supporting wellness in Inuit communities in Nunavut. https://www.ccnsa-nccah.ca/docs/health/FS-InuitQaujimajatuqangitWellnessNunavut-Tagalik-EN.pdf

knowledge acquisition, collaboration, environmental stewardship, and resourcefulness to solve problems.[50]

Research Data Management within Digital Research Infrastructure

Taking a reductive approach to analyzing DM as a unique pillar of DRI may be helpful for compartmentalizing and understanding its functions, needs, and impact on digital research; however, in reality the DRI landscape is more complex. Rather, components should be viewed as nested, enabling and deriving support from each other. As the LCDRI's 2017 coordination paper notes, these individual components only function to their true potential when they are integrated to facilitate efficient and effective workflows for researchers. Effective management of digital research data relies on a robust array of supporting digital research infrastructure that includes ARC and RS components. Inversely, effective use of ARC and RS requires that the research data they support be maintained over their lifecycle with effective management practices, in order to support complex investigations and derive value and greater impacts on research and society.

How does RDM support or rely on the other DRI elements?

Network

- Network supports digital research infrastructure, enabling common communication protocols to share digital resources between network nodes. This process relies on common architectures, standards, and procedures to reliably and securely share information.

- For research data to be shared and accessed over digital networks, it must be created and managed with these considerations in mind, which has led to the creation and adoption of a range of semantic and technical schemas and standards that enable data to be secured, shared, discovered, accessed, and preserved.

Advanced Research Computing (ARC)

- ARC is an umbrella term for the range of infrastructure and services required for data intensive research, including access to computing resources and active storage. Following best practices is crucial for hosting operations and management of the large data assets, while protecting information privacy and security.

- Notably core ARC infrastructure and services, as defined in the 2017 LCDRI ARC position paper, do not cover repository and archival storage infrastructure, which are critical components of DRI for long-term data stewardship and preservation.

[50] Nunavut Impact Review Board (n.d.) Inuit Qaujimajatuqangit. https://www.nirb.ca/inuit-qaujimajatuqangit (Retrieved November 2020)

Research Software

- Digital research relies on a range of software to help researchers to collect, shape, and analyze their data. Similarly, a range of software for curation, preservation, publishing, and discovery enables the management of that data over time.

- In return, in order to more widely apply research software to work with a range of inputs, data must adhere to requirements that should be considered in its ongoing management, including use of common data formats, ontologies, and standards.

National Vision for RDM Support

Previous efforts to develop a coordinated vision statement for national support for RDM in Canada have produced high-level statements that capture a range of collective ideals. For instance, the LCDRI's position paper on data management envisions "*An innovative and coordinated research data management community, providing responsive services and resources that support Canadian researchers in advancing the research that is critical to building and sustaining Canada's economic and social prosperity."* [51] As well, the 2019 Kanata Declaration presents a series of eighteen statements from members of the RDM community prioritizing requirements and actions for a national RDM organization.[52]

Recognizing the incredible diversity, scale, and breadth of the Canadian research community, synthesizing a single vision that speaks to the range of goals and capacities of research organizations across the country is a challenging task. Reducing previous statements to their core principles reveals that the Canadian research community in favour of national support that is innovative, researcher focused, inclusive, and sustainable. These four principles also reflect the Alliance's guiding principles,[53] which anchor the vision for the Alliance's future development.

Innovative

- As research becomes more data intensive, new approaches to supporting data curation and preservation should be encouraged to promote long-term availability. An innovative national service requires agility, the ability to identify and respond to new conditions, and the capacity to support initiatives from the ground-up that advance research and the broader DRI ecosystem.

Researcher-Focused

- The needs of researchers should be central to the design and delivery of national infrastructure and services. Resources that are invisible, inaccessible, cumbersome, and do not integrate with existing workflows and tools currently used by researchers will be

---

[51] Leadership Council for Digital Research Infrastructure. (2017). Data Management Position Paper for Innovation, Science and Economic Development Canada. p.26 (unpublished)

[52] Attendees of the NDSF Summit. (2019). Kanata Declaration. Zenodo. http://doi.org/10.5281/zenodo.3234815

[53] New Digital Research Infrastructure Organization (2020). Guiding Principles. https://engagedri.ca/about-engage-dri/guiding-principles (Retrieved November 2020)

poorly adopted. This requires communication and collaboration with researchers' home institutions and research organizations that provide the local support and infrastructure researchers rely upon.

Inclusive

- National support, funding, and training should be accessible by researchers across institutions and domains, throughout the research data lifecycle. Support should be provided equally for data that can be shared openly, as well for various forms of sensitive data that require additional privacy and security safeguards, or which abide by distinct governance systems.

- A national organization should encourage and support communication among the research community, including researchers, institutions, funders, and service providers to foster trust and consensus among RDM priorities, and should foster diversity within the organization.

Sustainable

- A mosaic of partnerships and funding models should be encouraged to foster resilience amongst components of the DRI ecosystem. This is particularly true for RDM services and infrastructure that support long-term availability of research data.

3 Current National RDM Support

In a distributed landscape, it is essential that practices, policies, and standards are coordinated. In the past decade, Research Data Canada, and the Canadian Association of Research Libraries' Portage Network have assumed leadership roles in facilitating support, coordination, and collaboration for RDM at a national level. Both organizations have succeeded in advancing complementary agendas. The CARL Portage Network has coordinated efforts of institutional partners from across Canada, bringing together a network of experts to develop platforms, services, and guidance, providing practical support to Canadian research institutions. RDC has fostered opportunities for national dialogue, convening a wide range of RDM stakeholders to discuss and coordinate a framework for national data services, giving rise to focused statements such as the Kanata Declaration. The work and progress made by these two organizations should continue uninterrupted while the Alliance brings together their distinct mandates into a cohesive platform for supporting and advancing the state of RDM in Canada.

Research Data Canada

(https://www.rdc-drc.ca/)

Research Data Canada (RDC) was established following a recommendation of the 2011 Canadian Research Data Summit,[54] to bring together key stakeholders to develop strategy, and

[54] Research Data Strategy Working Group (2011). The 2011 Canadian Research Data Summit: Mapping the Data Landscape. https://www.rdc-drc.ca/download/background-to-the-2011-canadian-research-data-summit

facilitate communication and partnerships to advance common objectives, education, and awareness for RDM nationally. RDC is a stakeholder-driven and supported organization. The RDC Steering Committee, which consists of representatives from stakeholder organizations, provides oversight and governance of RDC activities on behalf of the broader stakeholder community.

CANARIE has hosted and supported RDC's activities since 2014, and leadership for RDC has been provided by an Executive Director since 2016.

The work of RDC is advanced via a range of committees and working groups who have led initiatives to support standards for interoperable research data and RDM infrastructure, as well as to map the current landscape of stakeholders, services, and infrastructure.

Current work themes:

- Communications, Outreach, and Education
- Infrastructure
- Policy
- Standards & Interoperability

Since 2017, RDC has convened a series of NDSF Summits to bring together members of the RDM community to discuss and propose coordinated action for advancing the state of RDM in Canada. The summits have led to a number of key outputs, including the Kanata Declaration.

RDC also undertakes significant engagement with international initiatives. For instance, collaboration with the Research Data Alliance (RDA), of which RDC's current Executive Director is currently Co-Chair of the RDA Council. RDA is an influential platform where international research data experts meet to exchange views and advance topics related to best practices, standards, and protocols, leading to a range of outputs including RDA Recommendations.

During the transitionary period towards integration with the Alliance, RDC is focused on strengthening the RDC ecosystem through advancing actions suggested by recent NDSF summits, supporting collaboration between RDM and DRI funders, and facilitating international collaboration between Canadian and global Open Science efforts.

Portage Network, Canadian Association of Research Libraries

(https://portagenetwork.ca/)

Launched in 2015, the Portage Network is a national initiative of CARL, with the goal of promoting shared stewardship of research data and building RDM capacity in Canada through a network of over 150 community members in a growing community of practice. Portage began its work through engaging members of the university library community, and has since expanded its network to include support providers, administrators, and researchers from beyond academia. The central aim of Portage is to coordinate and expand expertise, services, tools, and platforms so that researchers across Canada have access to the support and tools necessary for RDM.

Oversight and direction for Portage is provided by a Steering Committee consisting of CARL Directors, and an Advisory Committee consisting of representatives from key external stakeholder groups.

Portage advances its work through a range of thematic expert groups and working groups, which span various stages of the research data lifecycle. These groups are responsible for conducting research, producing guidance and best practices, and overseeing Portage's development and support of RDM tools and services.

Current work themes:

- Curation
- Data Discovery
- Data Repositories
- Data Management Plans
- Institutional Strategies
- Preservation
- Research Intelligence
- Sensitive Data
- Training

Portage establishes partnerships with allied organizations to fund and support the development of its national platforms and services.

- DMP Assistant,[55] hosted by the University of Alberta, is a bilingual tool for preparing data management plans. The platform is free for researchers across Canada and permits research institutions to create custom guidance and templates for their members.

- Federated Research Data Repository,[56] co-developed with Compute Canada, is a scalable federated data repository platform for data management, publication, and discovery. The platform acts as a national discovery layer by harvesting metadata records from other Canadian repositories, exposing data assets and driving traffic to the host repositories. It also allows researchers to deposit datasets directly for curation, publication, and preservation, and is designed to accommodate datasets that are too large to host within most institutional or generalist repositories.

- The development of a national Dataverse repository instance, in partnership with OCUL Scholars Portal,[57] with more than 55 participating institutions, with guidance from the Portage Dataverse North Working Group.

[55] https://assistant.portagenetwork.ca

[56] https://www.frdr-dfdr.ca/repo

[57] https://dataverse.scholarsportal.info

- A Canadian DataCite Consortium,[58] led by CARL Portage and the Canadian Research Knowledge Network, that offers DOI minting services to Canadian institutions via DataCite.

To date, Portage has established a solid foundation for RDM services and platforms offered via research institutions and through direct support, and helped raise the national profile of RDM more generally. Until 2019, Portage was supported through investments and in-kind contributions made by CARL member institutions, which supported both Director and Service Manager positions to oversee the development of the network.

In 2019, Portage was successful in securing transition funding from ISED to expand capacity for DM activities in advance of the Alliance transition. CANARIE has administered this funding on their behalf. With the transition funding secured, Portage has been able to expand its operations through strategic hiring to support key expert groups and related service areas. Areas of focus include support for the Federated Research Data Repository to offer national data discovery, curation, and repository services, support for the development of a national network of curation support, service development for the DMP Assistant, supporting a national training program, and fostering support for a national preservation service. A second round of hiring anticipated in Fall 2020 will add new positions to provide additional support for data curation and data management planning, as well as new support for research intelligence and sensitive data.

4 National RDM Landscape Analysis

A range of other actors distributed throughout the landscape play a role in supporting national RDM. Canadian researchers draw upon support or services throughout the lifecycle of a research project, which has led to a plurality of groups and organizations supporting the long-term stewardship of many research data assets. Establishing a clear understanding of existing organizations, positioned from local to national scales with appropriate international coordination, will be important for the Alliance to successfully integrate into the RDM landscape and provide suitable national direction and coordination.

Given the scale and complexity of the ecosystem, it is not practical to provide a comprehensive listing of the hundreds of actors within the context of this report. The following section outlines stakeholders in this landscape that contribute to supporting RDM nationally, with some specific organizations listed for context where necessary. Key sectors are listed below, in order of proximity to researchers situated in academic institutions.

Higher Education

The distribution of responsibilities for Canadian higher education and research between individual institutions, the Provinces and Territories, and the Federal government has produced a stratified model of organizations, infrastructure, and services.

[58] https://www.crkn-rcdr.ca/en/datacite-canada-consortium

Higher education institutions have a mission to contribute to society through education, learning, and research. A range of parties work within higher education institutions to support research data created and managed by researchers within their community.

- The Departments and Faculties to which researchers belong play a role in supporting RDM. Through setting program requirements and course calendars, and offering seminars to students and faculty, they can directly influence RDM practices adopted by their research community, and more broadly contribute to fostering a culture of open science. In terms of infrastructure, they may also provide local network storage for research data, and institute specific policies governing storage, retention, and publication of research data.

- University Systems and IT Departments offer a catalogue of computing infrastructure and services that support RDM. This typically includes a range of storage and backup infrastructure for active research and longer-term storage whose timelines reflect institutional retention policies. They also manage identity management systems, which control access and security of research data holdings on local and integrated external systems.

- University Libraries and Archives are supporting RDM as an emerging service area, in step with the growing recognition that research data are important scholarly outputs that must be valued and managed in concert with more traditional scholarly objects. This support has evolved around hosting digital asset management systems, such as institutional repositories and digital collection platforms, as well as related scholarly communications and copyright services. Upstream services in the research lifecycle, such as support for data management planning and data curation, have arisen to improve the quality of data being archived downstream, and as more research data finds its way into their collections, these institutions are also supporting data downstream via digital preservation services.

- Research Offices support RDM by setting institutional policies and strategies related to research practices and administration. This includes supporting research ethics boards that ensure research involving human subjects abides by established institutional and national best practices. They also support researchers in complying with RDM policies and requirements set by relevant grant and award bodies.

National coordination and leadership for each these groups is led by national associations, including the Canadian Association of Research Administrators (CARA),[59] Canadian Association of Research Ethics Boards (CAREB),[60] CARL,[61] and the Canadian University Council of Chief Information Officers (CUCCIO).[62]

---

[59] https://cara-acaar.ca

[60] https://www.careb-accer.org

[61] https://www.carl-abrc.ca

[62] https://www.cuccio.net/en

The extent to which local parties are able to provide support varies greatly with the capacity and research intensity of an institution. To coordinate practices and standards across institutions, improve access to infrastructure and services, and derive benefits from economies of scale, regional and national service providers have emerged to support researchers.

- Compute Canada,[63] in partnership with regional organizations WestGrid,[64] Compute Ontario,[65] Calcul Québec,[66] and ACENET,[67] deploy storage infrastructure underpinning DM of many active research projects. Compute Canada is also a partner with CARL-Portage in the national Federated Research Data Repository platform.

- The National Research and Education Network (NREN) is an essential collective of infrastructure, tools, and people serving research and higher education. CANARIE and twelve partners form Canada's NREN.[68] These network resources enable connectivity, access, and security to many DM platforms and services. Several NREN partners also provide long term storage infrastructure to their member institutions.

- Regional university library consortiums, such as the Ontario Council of University Libraries (OCUL),[69] and the Council of Prairie and Pacific University Libraries (COPPUL),[70] provide access to shared services and infrastructure across member institutions. This support includes access to repository platforms (e.g., Scholars Portal Dataverse),[71] storage and preservation infrastructure (e.g., Ontario Library Research Cloud,[72] Permafrost,[73] WestVault[74]).

- The Canadian Research Knowledge Network supports increased access to scholarly resources by negotiating licenses on behalf of its members, [75] including access to

[63] https://www.computecanada.ca

[64] https://www.westgrid.ca

[65] https://computeontario.ca

[66] https://www.calculquebec.ca

[67] https://www.ace-net.ca

[68] https://www.canarie.ca/network/nren

[69] https://ocul.on.ca

[70] https://coppul.ca

[71] https://dataverse.scholarsportal.info

[72] https://cloud.scholarsportal.info

[73] https://permafrost.scholarsportal.info

[74] https://coppul.ca/westvault

[75] https://www.crkn-rcdr.ca/en/home

persistent identifier tools via the ORCID-CA Consortium,[76] and the DataCite Canada Consortium, [77] provided in partnership with CARL.

Research Organizations

Supporting access to leading edge research data presents a range of unique RDM considerations and challenges, typically requiring access to significant DRI resources. Across the country, research organizations responsible for the creation and stewardship of many national data assets relied upon by researchers as crucial data sources, have also been instrumental in developing innovative services and tools to advance RDM within their subject domains, benefiting researchers far afield. These organizations are typically hosted and supported by academic institutions and governmental research bodies, and absorb significant funding and DRI resources.

To date, a comprehensive inventory of these organizations, infrastructures, and data assets is lacking from an assessment of the Canadian landscape and should be prioritized to support the Alliance's future strategic planning efforts. Because many of the Canadian domain specific RDM initiatives work cooperatively with their counterparts in other countries, the inventory should include the ties between national initiatives and international federations and aggregators. Several existing efforts of Portage Network and RDC working groups should be leveraged to expedite this process.

The following section summarizes several existing nationally funded RDM initiatives according to key research areas. **This list is not intended to be exhaustive, but rather illustrative of the diversity and complexity of initiatives.** A collaborative relationship between the Alliance and these and similar organizations will be necessary to coordinate the national RDM ecosystem and build mutually beneficial partnerships.

Earth, Ocean, and Environment:

Canada's vast geography and range of climates contributes reams of data from observational research infrastructure with predictive modeling to support research in conservation, environmental management, and resource development. The range of organizations involved in collecting and managing this data rely on significant DRI resources and provide international leadership in the advancement of standards supporting interoperability across legal and spatial boundaries.

Canadian Integrated Ocean Observing System (CIOOS)[78]

In 2019, the Government of Canada announced an investment of $1.5 million per year in ongoing funding to support a Canadian Integrated Ocean Observing System to promote sharing of data and expertise to support research efforts to better understand, monitor and manage activities in

[76] https://orcid-ca.org/home

[77] https://www.crkn-rcdr.ca/en/datacite-canada-consortium

[78] https://cioos.ca

Canada's oceans.[79] The initiative is jointly funded by the Marine Environmental Observation, Prediction and Response Network (MEOPAR), which is providing $2 million over 4 years. Partners include research organizations across the country, including the Ocean Frontier Institute, Dalhousie University, the Coastal and Ocean Information Network Atlantic, the Fisheries and Marine Institute of Memorial University of Newfoundland, the Ocean Tracking Network, the St. Lawrence Global Observatory, the Tula Foundation, and Ocean Networks Canada (University of Victoria), who will implement the first phase of the system. This initiative represents a regional partnership with GOOS, the Global Ocean Observing System,[80] a programme executed by the Intergovernmental Oceanographic Commission (IOC) of the UNESCO.

Polar Data Catalogue[81]

The Canadian Cryospheric Information Network was established in the 1990's through a collaborative partnership between departments of the Canadian Government, the University of Waterloo, and the private sector to facilitate the exchange of information among researchers, northern communities, international programs, and the public. The Polar Data Catalogue is their repository of metadata and data that describes and provides access to datasets generated by Arctic and Antarctic researchers. Records cover many topics from natural sciences, to health and social sciences, to policy. The PDC is a member of The Scientific Committee on Antarctic Research (SCAR),[82] which is a thematic organisation of the International Science Council (ISC). SCAR recognizes the PDC as the National Antarctic Data Centre (NADC) for Canada as part of their obligations to make data available under The Antarctic Treaty (1959). In turn, SCAR is a partner of the International Arctic Science Committee.[83]

Astronomy and Astrophysics

The scale of producing and managing vast arrays of astronomical data requires significant DRI investment achievable through national and international collaborations. This reality has contributed to a collaborative data sharing culture among astronomy and astrophysics researchers resulting in several national and international organizations dedicated to resolving challenges in interoperability and long-term management of these data assets.

---

[79] Government of Canada, Department of Fisheries and Oceans. (2019). Government of Canada's Investment in ocean observation technology contributes to safer coastal waters and more resilient coastal communities. https://www.canada.ca/en/fisheries-oceans/news/2019/03/government-of-canadas-investment-in-ocean-observation-technology-contributes-to-safer-coastal-waters-and-more-resilient-coastal-communities.html (Retrieved November 2020)

[80] https://www.goosocean.org

[81] https://www.polardata.ca

[82] https://www.scar.org

[83] https://iasc.info

Canadian Astronomy Data Centre[84]

The Canadian Astronomy Data Centre (CADC) was established in 1986 by the National Research Council of Canada, through a grant provided by the Canadian Space Agency. The CADC, in partnership with Shared Services Canada, Compute Canada, CANARIE and the university community (funded through CFI), offers cloud computing, user-managed storage, group management, and data publication services, in addition to its ongoing mission to provide permanent storage for major data collections. In 2019, the CADC delivered over two petabytes of data to thousands of astronomers in Canada and in over 80 other countries. The CADC is a member of the IVOA, the International Virtual Observatory Alliance,[85] which has created a standardized framework for data centers to provide interoperable data services, analysis and visualization software in a user interface designed to support researchers globally.

Nuclear and Particle Physics

Research in high-energy physics requires very large infrastructure investments achievable only through international collaboration. Strategies for managing these large quantities of data rely on national investment and collaboration.

TRIUMF/ATLAS-T1[86]

TRIUMF is Canada's national particle accelerator centre, located at the University of British Columbia and governed by university members across Canada. TRIUMF is also home to the ATLAS-Canada Tier-1 Data Centre, funded by CFI. This centre is one of the main data centers of CERN, which distributes data from the Large Hadron Collider via an international computing grid for analysis by Canadian and international researchers.

Life Sciences

Technological advances in the life sciences, in particular the omics branches, have revolutionised approaches for researching living organisms. The generation of most biomedical data is highly distributed, however the technologies needed for acquiring, storing, and sharing digital information have required collective infrastructures open to the scientific community.

Barcode of Life Data System[87]

The Barcode of Life Data System (BOLD) is a data repository and online bioinformatics research environment for the investigation and use of DNA barcode data developed at the Center for Biodiversity Genomics at the University of Guelph. The platform provides access to more than 8

[84] http://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en

[85] http://ivoa.net

[86] https://www.triumf.ca/atlas-group

[87] http://v4.boldsystems.org

million barcodes from over 300 thousand species. Funders include Genome Canada through the Ontario Genomics Institute, Ontario Innovation Trust, and NSERC.[88]

BioGRID[89]

The Biological General Repository for Interaction Datasets (BioGRID) is a public database that archives and shares genetic and protein interaction data from model organisms and humans. The repository currently holds over 1,740,000 interactions curated from both high-throughput datasets and individual studies, derived from over 70,000+ publications in the primary literature. The repository integrates data from across dozens of model organism and interaction databases. Canadian host facilities include the Lunenfeld-Tanenbaum Research Institute at Sinai Hospital in Toronto, and the Université de Montréal, with funding provided by CIHR and the US National Institutes of Health.

Health Science and Medicine

A mosaic of research organizations, research hospitals, and governmental health authorities are advancing research into new therapeutics, clinical investigations, and public health and safety via new technologies. This work requires robust management practices that must balance responsible strategies for maintaining data security and public trust.

Brain-CODE[90]

The Brain-CODE neuroinformatics platform is led by the Ontario Brain Institute to support acquisition, storage, and access to multidimensional data collected from patients with a variety of brain disorders. The development of Brain-CODE is supported by a range of public sector and not for profit organizations, and is hosted by the Centre for Academic Computing.

iReceptor[91]

iReceptor is a science gateway that enables discovery, access, analysis, and sharing of antibody/B-cell and T-cell receptor repertoires (Adaptive Immune Receptor Repertoire or AIRR-seq data) from multiple labs and institutions. The gateway integrates large, distributed datasets following community standards for interoperability and sharing, allowing users to search across 2 billion sequences. The project is located at the IRMACS Centre at Simon Fraser University. It is funded by the CANARIE Network Enabled Platforms Program, CFI, the BC Knowledge Development Fund, CIHR, and the European Union's Horizon 2020 research and innovation programme.

---

[88] Ratnasingham, S. & Hebert, P.D.N. (2007). BARCODING: bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol. Ecol. Notes.,* 7(3):355-364. http://doi.org/10.1111/j.1471-8286.2007.01678.x

[89] https://thebiogrid.org

[90] https://www.braincode.ca

[91] https://gateway.ireceptor.org

CanDIG[92]

The Canadian Distributed Infrastructure for Genomics (CanDIG) is a project building a health genomics platform for national-scale, federated analyses over locally controlled private data sets. Their CHORD project aims to build a federated national data service for sensitive genomics and health related data. It is funded by the CFI Cyberinfrastructure program and connects sites at McGill University, Hospital for Sick Children, UHN Princess Margaret Cancer Centre, Canada's Michael Smith Genome Sciences Centre, Jewish General Hospital and Université de Sherbrooke. It is also a collaboration with Genome Canada, Compute Canada and CANARIE.

Social Sciences and Humanities (SSH)

Research infrastructures are enabling the creation, manipulation, and management of large and heterogeneous data sets in SSH fields. Research organizations are advancing work of digital humanists publishing and managing vast digital corpora and other cultural materials. While also supporting the work of humanists in non-digital mediums, who are faced with increasing expectations around data sharing. In the social sciences, well-established research organizations who produce, compile, and provide access to data produced from public statistics, scientific surveys, and opinion polls are essential data sources to the Canadian research community.

Canadian Research Data Centre Network[93]

The Canadian Research Data Centre Network (CRDCN) is a partnership between a consortium of Canadian universities and Statistics Canada to provide university, government and other approved researchers access to confidential social, economic and health microdata in secure computer facilities located on university campuses across the country. Headquartered at McMaster University since 2010, it comprises 32 Centres and Branches, and provides researchers with unique, secure access to Statistics Canada census and survey data, as well as to an increasing number of administrative data files. There is also a growing understanding that while secure microdata itself cannot be deposited/shared via open repositories, documentation and code can. The Network operates through a formal agreement with Statistics Canada. Funding sources include SSHRC, CIHR, and CFI, with cash and in-kind support also provided by host universities and Statistics Canada. As part of the CRDCN 2019-2024 strategic plan they have committed to strengthening strategic relationships with internationally aligned organizations.[94]

First Nations Information Governance Centre[95]

The First Nations Information Governance Centre (FNIGC) is a non-profit organization working to achieve data sovereignty for First Nations in Canada, and through collaboration with partners, oversees significant data gathering initiatives to survey the health and well-being of First Nations

---

[92] https://www.distributedgenomics.ca

[93] https://crdcn.org

[94] Canadian Research Data Centre Network. (n.d.) Strategic Plan 2019-2024. https://crdcn.org/sites/default/files/strategic_plan_0.pdf

[95] https://fnigc.ca

peoples and their communities. These include the First Nations Regional Health Survey and the First Nations Regional Early Childhood, Education and Employment Survey. The FNIGC is also responsible for the stewardship of the OCAP® Principles and undertakes related training and outreach.[96]

Coalition Publica[97] and CO.SHS[98]

Coalition Publica is a partnership formed between two Canadian open publishing platforms, Érudit,[99] and the Public Knowledge Project,[100] to advance the Canadian SSH scholarly journal community towards sustainable open access through shared technological development and support for research activities investigating the scholarly publishing ecosystem. An aligned initiative of Érudit is CO.SHS, an open research infrastructure project for SSH in Canada which aims to increase the discoverability of research disseminated on the Érudit platform and support exploration of this corpora with advanced analysis and visualisation tools.

Research Funding Agencies

Funding agency mandates requiring the sharing of research data strongly influence researcher behaviour and the demand for RDM infrastructure and services. Over the last ten years, funding agencies and governments around the world have recognized the need for the development of national RDM policies to support access to publicly funded data. In 2016, the Canadian Tri-Agencies released a Statement of Principles on Digital Data Management that outlines their expectations for RDM, and the responsibilities of various actors in meeting these expectations.[101] In 2018, a draft RDM Policy was announced, which would require all grant recipients… "*to deposit into a recognized digital repository all digital research data, metadata and code that directly support the research conclusions in … research outputs that arise from agency-supported research*", while respecting ethical, legal, and commercial requirements, and adhering to principles of Indigenous data sovereignty.[102]

Academic institutions across the country have already begun to respond to the Tri-Agency's draft RDM Policy to support local researchers.[103] This effort has received significant support from the

---

[96] First Nations Information Governance Centre. (n.d.). Understanding OCAP®. https://fnigc.ca/ocap-training

[97] https://www.coalition-publi.ca

[98] https://co-shs.ca

[99] https://www.erudit.org

[100] https://pkp.sfu.ca/ojs

[101] Government of Canada. (2020). Tri-Agency Statement of Principles on Digital Data Management. https://www.ic.gc.ca/eic/site/063.nsf/eng/h_83F7624E.html (Retrieved November 2020)

[102] Government of Canada. (2020). Draft Tri-Agency Research Data Management Policy. https://www.ic.gc.ca/eic/site/063.nsf/eng/h_97610.html (Retrieved November 2020).

[103] Cooper, A. et al. (2020). Institutional Research Data Management Services Capacity Survey. http://doi.org/10.14288/1.0388722

Portage Network, which has developed a range of platforms, tools, services, and training to support adherence by institutions and researchers to all three policy pillars; including a generic DMP tool and associated guidance, two general repository options that support data deposit, and templates and guidance for the development of institutional strategies. Portage has also monitored progress towards developing institutional capacity to respond to the Tri-Agency draft RDM policy through a number of targeted surveys. It remains to be seen how the Tri-Agency will support researchers with adherence to this policy, although a range of RDM-focused grants have been successfully deployed from SSHRC to help institutions and researchers build capacity in preparation.[104]

Provincial and territorial research funders (e.g., Fonds de recherche du Québec,[105] Research Manitoba,[106] and Ontario Research Fund[107]) do not yet have policies or requirements in place regarding RDM. A coordinated effort to engage provincial and territorial funders to inform convergence of requirements and policies would benefit pan-Canadian research. Models for collaboration can be found in existing research associations (e.g., National Alliance of Provincial Health Research Organizations[108]).

Internationally, U.S. and U.K public research funders,[109] in addition to other high-profile funding sources (e.g., Wellcome Trust,[110] Bill & Melinda Gates Foundation[111]) already require researchers to exercise good data management in order to share underlying datasets of published research, as a requirement of funding.

Moving forward, the Alliance will need to maintain an awareness of the funder landscape, to ensure that its own services and infrastructure can support researchers in complying with growing expectations and integrate with other supports that may be developed. The Alliance will also need to ensure that its own funding requirements support unique considerations of its research communities, including respect for Indigenous data sovereignty. Greater awareness should also support the development of funding streams and programs that fill gaps in the funding landscape.

---

[104] Social Sciences and Humanities Research Council. (2020). Research Data Management Capacity Building Initiative. https://www.sshrc-crsh.gc.ca/funding-financement/programs-programmes/data_management-gestion_des_donnees-eng.aspx (Retrieved November 2020)

[105] http://www.frqsc.gouv.qc.ca/en

[106] https://researchmanitoba.ca

[107] https://www.ontario.ca/page/ontario-research-fund

[108] https://www.naphro.ca/about

[109] Data Curation Centre (n.d.) Funders' data plan requirements. https://www.dcc.ac.uk/resources/data-management-plans/funders-requirements (Retrieved November 2020)

[110] Wellcome Trust. (2017). Data, software and materials management and sharing policy. https://wellcome.ac.uk/grant-funding/guidance/data-software-materials-management-and-sharing-policy (Retrieved November 2020)

[111] Bill & Melinda Gates Foundation. (2011). Bill & Melinda Gates Foundation's Data Access Principles: Frequently Asked Questions. https://docs.gatesfoundation.org/documents/faq.pdf (Retrieved November 2020)

Scholarly Publishers

Data sharing policies of scholarly publishers are another driver of research adoption of improved RDM practices. The number of journals and publishers introducing research data policies is on the rise. These policies cover a range of related issues, including data deposit into approved repositories, data availability statements, data citation, data standards and formats, and peer review of research data.[112]

In many instances, researchers, via their Academic Societies, have driven the open science mandate through their scholarly journals and related data sharing policies.[113] For instance, the American Geophysical Union,[114] Genetics Society of America,[115] and British Ecological Society,[116] offer some examples of early adopters.

While this shift reflects broader trends in the publishing landscape towards greater access to the scholarly record via open access models, for some publishers it also demonstrates moves towards greater influence and involvement throughout the scholarly lifecycle. Commercial publishers, including Springer-Nature,[117] and Elsevier,[118] recognizing the demand for RDM support from researchers have recently begun offering services directly supporting dataset curation and publication. While others have begun to form partnerships with third party repositories to support authors in publishing datasets underlying their publications (for e.g. the partnership between Wiley and Dryad).[119] The expansion of commercial publishers into the realm of data publishing is of some concern to the academic community, in light of the growth and related impacts of the monopolization on scholarly publishing by these corporations (see the works of Vincent Larivière for background).[120] Preliminary investigations led by Portage's Research Intelligence Expert Group reveal that while many publishers do permit researchers to

---

[112] Hrynaszkiewicz et al. (2020). Developing a Research Data Policy Framework for All Journals and Publishers. *Data Sci. J.*, 19(1): 5. http://doi.org/10.5334/dsj-2020-005

[113] Mitchener, W.K. (2015). Ecological data sharing. *Ecol Info*, 29(1):33-44. https://doi.org/10.1016/j.ecoinf.2015.06.010

[114] AGU (2019). Position Statement on Data. https://www.agu.org/Share-and-Advocate/Share/Policymakers/Position-Statements/Position_Data (Retrieved November 2020)

[115] McIntyre, L.M. (2010). Data: The foundation of science. *Genetics*, 184(1):1. https://www.genetics.org/content/184/1/1.full

[116] Norman, H. (2014). Mandating data archiving: experiences from the frontline. *Learn. Publ.,* 27(5):S35-S38. https://doi.org/10.1087/20140507

[117] https://www.springernature.com/gp/authors/research-data/research-data-support

[118] https://data.mendeley.com

[119] Wiley (n.d.) Data sharing: Data sharing with Dryad. https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/data-sharing-service.html (Retrieved November 2020)

[120] Larrivière, V. (n.d.) Articles de revues scientifiques ou professionnelles. https://crc.ebsi.umontreal.ca/publications/?person=71 (Retrieved November 2020)

publish data under open licenses, more restrictive models do exist (for e.g., IEEE Dataport, [121] which implements a subscription-based model for full access and requires depositors to pay a fee to publish their data under an open access license). Further investigation into licensing and service models developed by scholarly publishers is of growing importance as these services gain prominence, to ensure that community interests are upheld.

Scholarly publishers have also been an important driver in developing and implementing what is sometimes referred to as the "Science Graph" or "Research Graph": an online database linking researchers, primary data and derived scholarly literature. Research Graphs are an active area of development and are being expanded to include links to funding agencies and other scholarly outputs such as reproducible workflows and patents. Canadian researchers participate in Research Graph frameworks and implementations such as the Scholix Framework,[122] implemented in the EU OpenAIRE portal,[123] and the Web of Science.[124]

Academia-Adjacent Organizations

A number of research organizations situated outside of academia produce and manage significant data assets of value to the Canadian research community. While these organizations may rely upon their own DRI and support services, the research data they steward are often derived in partnership with or used directly by academic researchers. It is within the interests of the national research community to ensure that these data assets are accessible and interoperable with national DRI, while also respecting community needs and Indigenous data sovereignty. Notwithstanding unique aspects and considerations of these organizations, adopting common RDM frameworks and practices across the landscape can help support this achievement. Strengthening partnerships between academia and academia-adjacent organizations to support research data will also yield a more resilient and sustainable DRI landscape.

Government: Data-Rich Departments and Research Centers

Numerous government departments and agencies collect, manage, and publish data related to social and scientific aspects of Canadian life that presents a valuable resource for the Canadian research community. Many of these government organizations also have very close working relationships with external research organizations and academia. Thus, the separation between government and research organization is not always a clear distinction. For instance, the National Research Council of Canada, the primary research and technology organization of the Government of Canada, operates a number of research centres that collaborate closely with both academia and industry.[125] Statistics Canada is another national agency responsible for providing access to valuable data on social, economic and health issues to researchers in academia and

[121] https://ieee-dataport.org

[122] http://www.scholix.org

[123] https://www.openaire.eu

[124] https://clarivate.libguides.com/webofscienceplatform

[125] https://nrc.canada.ca

government through their CRDCN partnerships with a consortium of Canadian institutions.[126] Several examples of departments at all levels of government stewarding data sources of national importance are listed below.

- Research centers located within Federal departments and agencies lead scientific data collection, sharing outputs through both subject-specific repositories (e.g., Fisheries and Oceans Data Archive,[127] NRCan Earth Observation Data Management System,[128] StatsCan Public Use Microdata File Collection[129]) and the government of Canada's open data portal.[130]

- Provincial ministries collect and share data related to key sectors, including health, education, and natural resources, via databases, subject-specific data repositories and open data portals (e.g., Alberta Geological Survey,[131] Public Health Ontario,[132] BC Data Catalogue[133]).

- Municipal governments collect, aggregate, and release civic data through local open data portals (e.g., City of Ottawa[134]).

Regional Health Authorities and Health Research Agencies

Within hospitals across the country, clinicians and researchers work together to lead life-saving research to prevent, diagnose, and treat diseases through clinical trials, and gather evidence to improve health services. Securely managing this data, while providing access to authorized researchers relies on strong RDM practices. Regional health authorities may coordinate policies and practices across a region, and support access to data to external researchers (e.g., BC Cancer,[135] BC Health Authorities[136]).

Provincial health ministries may also provide funding to independent research associations that provide controlled access to administrative health services records for approved research

---

[126] https://crdcn.org/about-crdcn

[127] https://www.pac.dfo-mpo.gc.ca/science/oceans/data-donnees/index-eng.html

[128] https://www.eodms-sgdot.nrcan-rncan.gc.ca/index_en.jsp

[129] https://www150.statcan.gc.ca/n1/pub/11-625-x/11-625-x2010000-eng.htm

[130] https://open.canada.ca

[131] https://ags.aer.ca

[132] https://www.publichealthontario.ca/en/data-and-analysis/using-data/open-data

[133] https://data.gov.bc.ca

[134] https://open.ottawa.ca

[135] http://www.bccancer.bc.ca/health-professionals/professional-resources/bc-cancer-registry/request-registry-data

[136] https://www2.gov.bc.ca/gov/content/health/conducting-health-research-evaluation/data-access-health-data-central/health-authorities

requests and authorized researchers (e.g., Population Data BC,[137] ICES[138]). National health-related not-for-profit research agencies also lead the collection, management, and controlled access for health data, coordinated across provincial and municipal jurisdictions (e.g., Health Data Research Network,[139] Canadian Institute for Health Information[140]).

Cultural Institutions (galleries, libraries, archives, and museums (GLAM))

Cultural and memory institutions facilitate access to cultural heritage data, supporting exchanges between the GLAM sector and researchers, particularly in the humanities. As digitization technology improves and becomes more accessible, and as more materials are born in digital formats, improved digital curation and long-term management will be vital to effectively exploit these data assets as research data. GLAM institutions also have much to share with the academic research community to advance RDM, particularly concerning collection development (selection and appraisal) and preservation. Examples range from large institutions like Library and Archives Canada's digital collections,[141] to museum digital collections (e.g., Ingenium Open Data[142]), to smaller institutional archives.

Industry and Private Sector

Researchers across disciplines collaborate with industry and private sector organizations to conduct research and collect and manage data of commercial or strategic value. Canadian centers of academic-industry collaboration lead innovation and discovery in a range of fields (e.g., MaRS,[143] Quebec Consortium for Drug Discovery,[144] DWave's Leap Programme[145]). More researchers are also leveraging third-party data collected by commercial tech companies and service providers (e.g., social media, ISPs). Securely managing IP generated from these partnerships, while fulfilling new expectations for data access pose challenges in long-term management.

Third-Party Service Providers (commercial and non-profit)

Many academic research groups also rely on commercial digital infrastructure to support their research for their ease of use and familiarity, or when their DRI needs are too great to be filled by institutional service providers. This is particularly true of storage and computing infrastructure, with many research projects relying on services from Amazon, Google, and Microsoft. Long-term

[137] https://www.popdata.bc.ca

[138] https://www.ices.on.ca

[139] https://www.hdrn.ca

[140] https://www.cihi.ca

[141] https://www.bac-lac.gc.ca/eng/Pages/home.aspx

[142] https://ingeniumcanada.org/collection-research/open-data

[143] https://www.marsdd.com

[144] https://cqdm.org

[145] https://www.dwavesys.com/take-leap

management and sharing data stored on commercial infrastructure pose significant challenges from a DM perspective. These challenges include, but are not limited to: the heavy lift of migrating existing data resources to cloud infrastructures and subsequent development of new workflows, institutional policies associated with procurement, and ensuring commercial providers comply with institutional security and privacy policies and regulations. In the future the Alliance should have a role identifying challenges and benefits of working with commercial digital infrastructure providers.

A number of commercial and not for profit organizations supporting the open science movement have also gained prominence in various research communities in the last decade through the development of virtual research environments. For example, collaboration and sharing platforms like Open Science Framework,[146] repository platforms like Dryad and Figshare,[147,148] computational research platforms like Code Ocean,[149] and the many electronic lab notebook providers. Partnerships and integrations between publicly funded infrastructure and services with widely adopted open science tools should be investigated to support national research.

International Organizations

A number of significant national or pan-national initiatives exist to coordinate the open science landscape and provide foundational services and infrastructure to support researchers with the management of their research data. As the Alliance matures, engagement with these initiatives, which is to an extent already in play through RDC and Portage, should be sought to ensure alignment and benefit from commonalities.

National and Pan-National Government Initiatives

A number of national or pan-national initiatives exist in various phases of development in jurisdictions around the world, which have comparable mandates to Canada's Alliance to provide foundational services and infrastructure to support researchers with the management of their research data. The range of approaches taken by these initiatives with respect to organization, services, and business models offer potential learning opportunities for the Alliance.

A federated approach to advancing open science is taking shape across Europe. In 2016, the European Commission allocated funding for the federation of scientific data infrastructures through a new entity known as the European Open Science Cloud (EOSC).[150] EOSC will foster a network of organisations and infrastructures from various countries and communities that supports the open creation and dissemination of knowledge and scientific data. The objective of EOSC is to support RDM across Europe via interoperable data, services, and infrastructures via

[146] https://osf.io

[147] https://datadryad.org

[148] https://figshare.com

[149] https://codeocean.com

[150] https://www.eosc-portal.eu

the formation of a “minimal viable platform” consisting of rules for participation to guide the provision of interconnected services and interoperable data.

Another federated approach is taking shape in Africa through the African Open Science Platform (AOSP), which is an initiative launched in 2016 by the South African Department of Science and Technology with the objective of developing connections between open science activities underway across Africa via mechanisms for collaboration and coordination, and the exchange of best practices.[151]

Many existing national initiatives also consider the importance of RDM in supporting national DRI. The Australian Research Data Commons (ARDC) has a mandate to provide national coherence to data and e-research platform capability, and to accelerate Australian research by developing, testing, and supporting platforms where investigators can store, discover, share, access, and interact with digital objects (data, software, etc.). [152] While in Germany, the National Research Data Infrastructure takes a different approach by bringing collaborators together via a coordinated network of consortia tasked with providing data services to research communities.[153] Consortia are generally organised by research domain or method and their aim is to improve and safeguard access to and use of research data in their relevant areas.

For a more complete analysis of comparable national and pan-national initiatives, see Appendix B.

International Associations

Numerous international organizations with ties to Canada are advancing RDM best practice through the development and coordination of communities of practice. The most prominent of these initiatives is the Research Data Alliance (RDA) launched in 2013.[154] As of March 2020, it had almost 10,000 members based in 144 countries[155]. RDA provides a platform where international research data experts meet to exchange views and advance topics related to best practices, standards, and protocols. Work is conducted through working groups with functional (e.g., ID Management, Vocabulary services, Virtual Research Environments) or domain-based focuses. Outputs include RDA Recommendations (documents that may include specifications, ontologies, workflows, data models, etc.), which are officially endorsed by RDA.

Meanwhile, other international organizations focus on providing services and support to member organizations. For instance, the World Data System is an interdisciplinary body of the International Science Council, with the mission of supporting access and stewardship of trusted scientific data and data services, products, and information.[156] Members include 125 scientific data repositories,

---

[151] http://africanopenscience.org.za

[152] https://ardc.edu.au

[153] https://www.dfg.de/en/research_funding/programmes/nfdi/index.html

[154] https://rd-alliance.org

[155] WDS-ITO. (2019). RDA Membership Worldwide. Tableau Public. https://public.tableau.com/profile/littlehelper#!/vizhome/RDAMembershipWorldwide/RDAMembershipWorldwide (Retrieved November 2020)

[156] https://www.icsu-wds.org

scientific societies, data services, and related organizations located around the world. Their International Technology Office (ITO) is based at the University of Victoria and supported by three Canadian organizations: Ocean Networks Canada, Polar Data Catalogue, and Canadian Astronomy Data Centre.[157] The ITO supports member organizations and partners via technical infrastructure, expert advice, and services to support access to scientific data.

Canada has been a member of The ISC's Committee on Data (CODATA) since the 1960's, and two Canadians currently serve on the CODATA Executive Committee.[158] CODATA is represented and coordinated in Canada via the Canadian National Committee for CODATA (CNC/CODATA).[159] The CODATA International Policy Committee provides expert input on the development and implementation of data policies to a range of international initiatives,[160] including the joint CODATA-OECD Principles and Guidelines for Access to Research Data from Public Funding.[161]

As both CODATA and WDS are international bodies of the International Science Council, both are supporting the Decadal program, which focuses on Interoperability of data across disciples.[162] More broadly, in 2020, four large international data organizations (WDS, the Research Data Alliance (RDA), CODATA and GO FAIR), formally stated their joint commitment to work together to optimize the global research data ecosystem. Collectively referred to as "Data Together," this document obligates all signatories to work together to "optimise the global research data ecosystem and to identify the opportunities and needs that will trigger federated infrastructures to service the new reality of data-driven science."[163]

The Global Indigenous Data Alliance (GIDA) is an international network of Indigenous researchers, data experts, and policy makers devoted to advancing Indigenous control over Indigenous data through advocating for data sovereignty and Indigenous data governance at the international level and within nation-states. The Alliance has generated resources that support Indigenous data sovereignty, including the CARE principles, and the recent COVID-19 guidelines for data sharing respecting Indigenous sovereignty.[164]

---

[157] https://wds-ito.org

[158] https://codata.org

[159] https://codata.org/canada

[160] CODATA. (n.d.). International Data Policy Committee. https://codata.org/initiatives/strategic-programme/international-data-policy-committee (Retrieved November 2020).

[161] OECD. (2007). OECD Principles and Guidelines for Access to Research Data from Public Funding. https://www.oecd.org/sti/inno/38500813.pdf

[162] CODATA. (2019). Decadal Programme: Making data work for cross-domain grand challenges. https://codata.org/initiatives/strategic-programme/decadal-programme (Retrieved November 2020)

[163] GoFAIR. (2020). Data Together Statement. https://www.go-fair.org/2020/03/30/data-together-statement (Retrieved November 2020)

[164] RDA COVID-19 Working Group. (2020). RDA COVID-19 Recommendations and Guidelines on Data Sharing. http://doi.org/10.15497/rda00052

A growing number of international organizations with domain focuses are also advancing and supporting RDM in their communities to coordinate global research efforts. For example:

- Global Open Data for Agriculture and Nutrition (GODAN) comprises a network of over 1,000 members from national governments, NGOs, international and private sector organizations.[165] GODAN and partners aim to improve food security and improve livelihoods of farming communities by combining open data advocacy and consultancy with innovative solutions, GODAN's secretariat is hosted by McGill University.

- Global Alliance for Genomics and Health (GA4GH) is an international Alliance that brings together 500+ organizations in healthcare, research, patient advocacy, life science, and information technology.[166] The GA4GH community advances frameworks and standards to enable the responsible, voluntary, and secure sharing of genomic and health-related data. GA4GH is headquartered in Toronto, in the MaRS Discovery District, and funded in part by the pan Canadian International Data Sharing Initiative (Can-SHARE).[167]

For a more detailed list of allied international organizations supporting RDM, see Appendix C.

5 Current State Assessment

The RDM "ecosystem" consists of an evolving series of interconnected components that interact with and support one another. The concept of an ecosystem provides a good analogy, where components are connected and interdependent, and which evolve with research practices and technology forming a growing network organized around supporting underlying research data flowing through it, and where a range of "niches" reflect domain-specific focuses or requirements. The relationships found in this ecosystem go beyond that of a data producer and a given dataset, to include many related aspects that exert impacts over the data's lifecycle. For instance, the standards and protocols applied to the data, the infrastructure it is created, analyzed and stored upon, training received by creators and managers, and related governing policies. The range of organizations outlined in the landscape outlined above, benefit and contribute to the evolution of the RDM ecosystem by advancing components through a range of platforms, services, guidance, and the research practice itself.

[165] https://www.godan.info

[166] https://www.ga4gh.org

[167] https://www.genomebc.ca/projects/canadian-international-data-sharing-initiative-can-share

The RDM Current State Map (Figure 2) organizes the ecosystem into discrete components needed to support RDM nationally: Storage and Compute, Interoperability, Data Services, and Governance. The following section of this report delves into each of these categories to examine how they contribute to supporting RDM and summarizes existing available support.

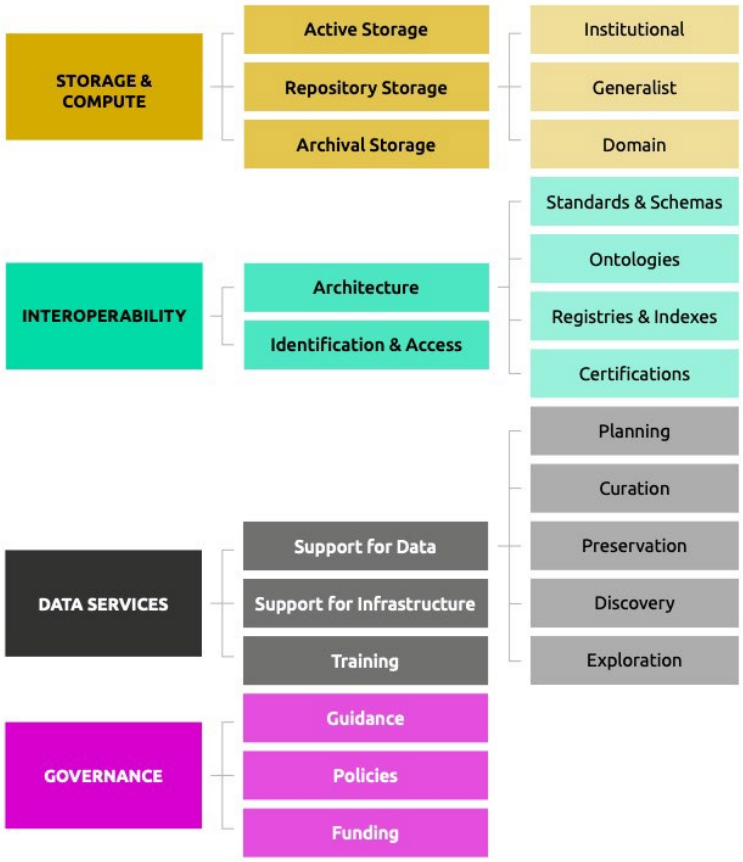


Figure 2. Map of key infrastructure and service categories used to describe the current state of the RDM ecosystem in Canada

Storage and Compute

From the perspective of RDM, there are three distinct configurations of computing and storage infrastructure to support distinct stages of data throughout its lifecycle: active, repository and archival. It is important to note that the key difference between storage types that this model aims to emphasize is not technical. For example, archival storage infrastructure is provided for some active research projects (e.g., high-capacity tape subsystems for infrequently accessed datasets that are part of an active research project),[168] and there are repositories and archival storage services that employ the same storage infrastructure used by active research projects, while at

[168] See “Nearline storage”, Compute Canada Technical Glossary. (Retrieved November 2020). https://docs.computecanada.ca/wiki/Technical_glossary_for_the_resource_allocation_competitions

the same time back-up to archival tape systems. Rather, the continuum model emphasizes the adoption of practices, resources, and policies suitable to these distinct stages in the data lifecycle.

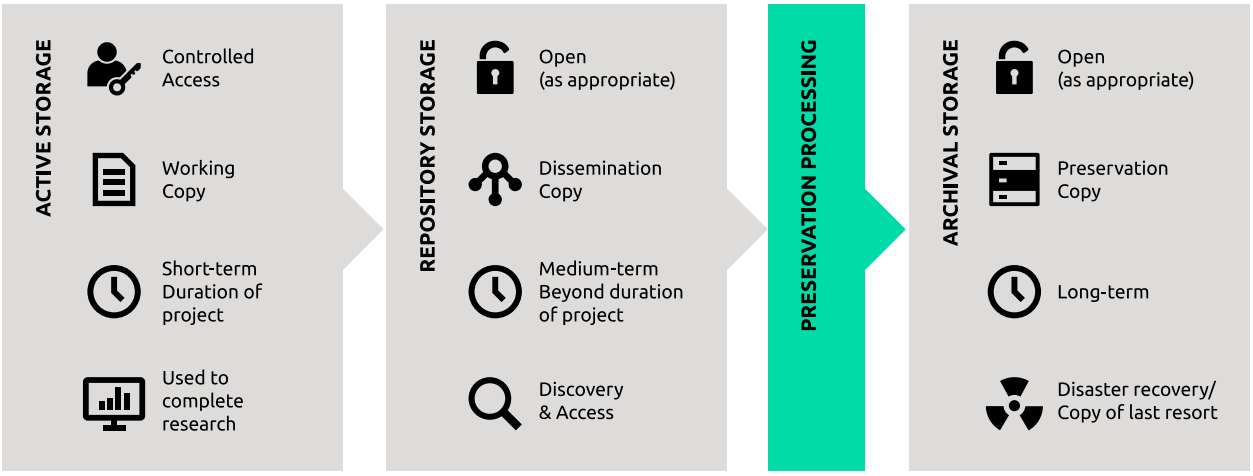


Figure 3. Description of research data storage spectrum (credit: CARL-Portage)

Active Storage (research phase)

In the active research phase, computing and storage infrastructure address the needs of researchers and the data during the research process itself, when data are being collected, modified, or analyzed. It includes high performance computing (HPC), massive storage, cloud computing/storage, distributed computing/storage, personal computing/storage, among other types. Requirements vary depending on the length of the research project and the amount of data processed and stored. While active research infrastructure, and the role of institutional, regional, and national long-term storage solutions will require ongoing discussions within the ARC community, for the purposes of this report, it is important to specify the connections to RDM, for which storage infrastructure is particularly relevant.

Researchers will generally have available to them storage infrastructure provided by their home institutions, as well as that which they seek out from publicly funded or commercial service providers. For example, within an academic institution, a researcher's home faculty or department may provide networked file storage, and the university systems department may offer researchers across campus access to HPC and cloud storage. Availability will greatly depend on institutional capacity. For this reason, institutions and governments have sought to fund active research infrastructure regionally and nationally, to benefit from economies of scale. Thus, researchers based at higher education institutions across Canada also have access to infrastructure supporting management of data in the active research phase through Compute Canada, CANARIE and the National Research Education Network.

More recently, the prevalence of cloud computing/storage and convenience of access has resulted in many researchers funding their active storage needs through commercial providers like Amazon, Microsoft, and Google. While reliance on these services poses data management challenges around long-term preservation and risk considerations around what data can or should be kept on these servers, partnerships with commercial providers have proven an effective means for certain research groups to lower economic and technological barriers to access ARC resources to meet their data needs. It bears consideration how to engage with commercial storage

providers and NREN partners to provide national support. For instance, the US National Institutes of Health (NIH) launched their STRIDES initiative in 2018 to support funded researchers in accessing commercial cloud services to accelerate biomedical discoveries.[169] Service agreements with commercial service providers raise a number of concerns that were investigated in the NOAA Big Data Project between 2015 and 2019.[170] NOAA signed five identical Cooperative Research and Development Agreements (CRADAs) with Amazon Web Services (AWS), Google Cloud Platform (GCP), IBM, Microsoft Azure, and the Open Commons Consortium (OCC) and concluded that cloud provider platforms offer both technical advantages as well as potential pitfalls for data managers and researchers. While there are too many issues to be considered in the context of this report, it should be highlighted that the ease of scalability of commercial systems comes at the potential cost of vendor lock-in, and may actually create silos of data between competing clouds, potentially reducing the ability to analyze data across providers. In all cases, it is clear that skilled personnel are required to manage the policies, governance, and procurement of the service level agreements, and that supported infrastructure can integrate effectively with repository and preservation storage.

Repository Storage (access and publishing phase)

At the conclusion of an investigation or funded project, often in line with a publication, researchers make curatorial decisions around what data to retain and share to meet institutional, publisher, and funder requirements, support reproducibility of published findings, and advance future research needs. This stage of the research lifecycle relies on repository storage, which supports future discovery and appropriate access.[171] The primary purpose of repository storage is to ensure that digital data of research value are stored securely and can be discovered and accessed appropriately. Therefore, they must support sufficient controls for the data to be reliable, accessible, and usable ongoing. The recently published TRUST principles (Transparency, Responsibility, User focus, Sustainability, and Technology) reflect community expectations for responsible management and administration of data repositories.[172] While the FAIR principles provide a framework for discussions around best practices in data management, TRUST reflects principles for reliable digital repositories. While implementation of principle-based requirements is challenging to operationalize, objective assessments of TRUST may include certification of repositories by authoritative organizations (discussed further under the interoperability section). In an increasingly complex data repository landscape, assessment of quality is important to prevent long-term consequences of data loss.

Institutional and Generalist Repositories

To support researchers in sharing their research outputs in line with community, funder, and publisher expectations around open science, academic libraries have been supporting access to

---

[169] https://datascience.nih.gov/strides

[170] Vance, T.C. et al. (2019). From the oceans to the cloud: Opportunities and challenges for data, models, computation and workflows. *Front. Mar. Sci.*, 6:211 https://doi.org/10.3389/fmars.2019.00211

[171] Castle, D. et al. (2019). Position Paper – Data Management Roadmap: 2019-2024. https://www.rdc-drc.ca/download/position-paper-data-management-roadmap-2019-2024 (Retrieved November 2020)

[172] Lin, D. et al. (2020). The TRUST Principles for digital repositories. *Sci Data* 7, 144. https://doi.org/10.1038/s41597-020-0486-7

institutionally managed digital repositories for nearly two decades. More recently, this work has grown to include institutional data repositories, designed specifically to support the needs of research data (in contrast to other digital objects that synthesize data, like publications). As these repositories are designed to support all research data types from a wide user community, they are often referred to as "multi-disciplinary" repositories, which collect high-level metadata about their collections at the "dataset" level (rather than the datum-level) and accept a range of file formats.

In 2017, CARL-Portage convened an expert group to investigate opportunities for a national institutional repository service using the Dataverse repository platform, which would allow for coordination and platform development across member institutions who would benefit from economies of scale around repository management and operations. As a result of that work, Scholars Portal, the service arm of the Ontario Council of University Libraries, has since undertaken a national service model for their Dataverse repository and has partnerships with 55 academic institutions across Canada.[173] Costs and responsibilities are shared among members. Coordination, development, and institutional support is provided by Scholars Portal, while local services and researcher-facing support are provided by member libraries. Currently, data deposited into Scholars Portal Dataverse is stored on the University of Toronto Data Centre.

Scaling institutional repositories to accommodate very large datasets (100s GB to TBs) is particularly challenging given the infrastructure required and the high costs institutions would need to absorb over the long term, as more and more data is stored. With these constraints in mind, CARL-Portage and Compute Canada have led the development of the Federated Research Data Repository,[174] a national multi-disciplinary repository based on Compute Canada infrastructure, which would accommodate very large datasets and would be made available nationally to researchers from academic institutions across the country. A framework governing access and storage allocations is currently in development.

In addition to publicly funded generalist repositories, a range of multi-disciplinary repositories operated by not-for-profit research organizations have gained a significant place in the landscape due to adoption by many researchers. One clear benefit of these repositories is they support international collaboration and deposit regardless of institutional or national affiliation. Most notably, these include Zenodo, run by the European Organization for Nuclear Research (CERN),[175] and Dryad, a non-for-profit repository organization formed through partnerships between scientific societies and scholarly publishers.[176]

Domain Repositories

A range of domain-specific data repositories operated by research groups and organizations located in Canada and internationally also play a key role in the national RDM ecosystem. Many domain repositories are endorsed by their respective research communities, which means that the data they contain is more likely to be discovered and trusted by researchers in respective

[173] https://dataverse.scholarsportal.info/#partners

[174] https://www.frdr-dfdr.ca/repo

[175] https://home.cern

[176] https://datadryad.org/stash/our_membership

fields. Also, since the data these repositories contains reflect a narrow range of subject matter, richer metadata can be applied, as well as domain-specific practices in data standards and file types, in comparison to generalist repositories, increasing opportunities for interoperability between distinct datasets, as well as reusability.

Domain repositories can assume a range of forms to suit function (Table 1). For examples of Canadian research organizations operating domain-specific repositories, see the Landscape section above.

Type	Attributes	Examples
Research project / group repositories	• Shares data collected during span of funded research project(s) • Repository itself may be funded research project • Does not support open deposit • May provide application for data manipulation, visualization, or other data visitation features • Managed by individual labs or teams of curators	• Mountain Legacy Project • Canadian Epigenetics, Environment and Health Research Consortium Network • Primate Cell Type DB
Research organization repositories	• Serves out data collected under multiple funded research projects • Aggregates data across research projects • Does not support open deposit of data from external researchers • May provide application for data manipulation, visualization, or other data visitation features	• Ocean Networks Canada • First Nations Data Centre
Government repositories	• Serves out data collected or compiled by government departments • Specifically domain-focused (i.e. not generic open data sites)	• BC Data Conservation Centre • World Ozone and Ultraviolet Radiation Data Centre • National Climate Data Archive • NRCan Earth Observation Data Management System

Domain repositories	• Collects data from multiple researchers, projects, organizations, related to specific domain • May or may not provide curation services • May or may not aggregate data across datasets	• Polar Data Catalogue • Barcode of Life Data System
Knowledgebases	• Extract, gather, and curate data in a subject area • Relies on core datasets to link together a growing body of information	• Avibase • DrugBank • BioGRID
Federated repository infrastructure	• Software or platforms to support federated search across data repositories	• Open Data Canada • iReceptor Commons • Global Biodiversity Information Facility • Canadian Open Neuroscience Platform

Table 1. Categories of Domain Repositories in Canada

While domain repositories have a vital role to play in supporting RDM, their administration by a diverse set of research groups and organizations poses a particular challenge regarding long-term sustainability. While some organizations operating domain data repositories have business models that bring in revenue or have predictable institutional funding, most rely on short-term (3-5 year) project-based funding from research granting agencies, which reflects the typical life spans of research projects. Reliance on this type of funding source is incompatible for the long-term missions of data repositories, thus other funding sources should be created to sustain the RDM functions provided by repositories.

Project-based funding for domain data repositories has also resulted in the proliferation of many specialized repositories suited to outputs of specific research projects. These repositories in particular may be vulnerable if their administrators do not have strong backgrounds in RDM best practices to ensure implementation of established standards, documentation practices, and backup procedures. The proliferation of many small, project-based repositories is unsustainable from the objective of long-term preservation. Initiatives that establish pipelines for data to migrate from repositories to more stable, shared platforms are also currently being investigated by CARL-Portage's Preservation Expert Group.[177] The growing number of repositories also has implications on sustainable funding models. A recent OECD report on data repository business models found that many are largely dependent on public funding, most combining structural funding with other

[177] https://portagenetwork.ca/network-of-experts/portage-preservation-expert-group

streams of revenue including value added services and deposit side fees.[178] The authors note that as this space develops there will be increased opportunities to source infrastructure, platforms, and services through specialist service providers. Currently, the DataONE program provides an example of how diverse repositories can derive benefits from a range of shared services and infrastructure to suit their needs, while achieving scales of economy through a federated model.[179]

More broadly, efforts to bring together domain and multidisciplinary repositories would benefit the RDM ecosystem. These repositories exist but operate in relative isolation from one another. Current efforts of the CARL-Portage and RDC's joint Data Repository Expert Group to provide high-level coordination and a cohesive approach to repository development in Canada are working to bridge this divide.[180] There is also an opportunity for synergies with research software programs to develop and promote tools and platforms that support cohesion and collaboration between repositories.

Archival Storage (preservation phase)

Archival storage, also known as preservation storage, supports long-term care and accessibility of digital objects of research value. The Open Archival Information System reference model offers a coherent framework of principles and terminology for management and preservation practices of a digital archive.[181] Archival storage is intended to preserve a copy for the long term that is independently verifiable, trustworthy, and not software or hardware dependant, and therefore has many more considerations regarding stewardship of its contents. Archival storage is therefore one consideration in the broader discipline of digital preservation, which is concerned with ensuring that digital information of continuing value remains accessible and usable.[182] In addition to storage, regular maintenance activities such as migrations, media refreshment, error checking, and disaster recovery plans play an important role in enabling long-term preservation of access.

Archival storage infrastructure and digital preservation support for research data is currently supported at local and regional levels. Within higher education institutions, Systems and IT departments may provide options for secure storage that reflect institutional or granting agency retention policies (typically 7-10 years). The prevalence and capacity of this practice bears further investigation. Any preservation practices will vary depending on the type of storage used and administration priorities. Regional computing service providers also provide archival storage

---

[178] OECD. (2017). Business models for sustainable research data repositories. *OECD Science, Technology and Industry Policy Papers*. https://doi.org/10.1787/302b12bb-en

[179] https://www.dataone.org

[180] https://portagenetwork.ca/network-of-experts/data-repository-expert-group

[181] http://www.oais.info/

[182] UVic Libraries. (2017) Digital preservation framework. https://www.uvic.ca/library/featured/digitalpreservation/dp-framework-FINAL.pdf (Retrieved November 2020)

infrastructure (for e.g., BCNet's EduCloud Backup service,[183] and SciNet's High Performance Storage System[184]).

At the regional level, university library consortia have led initiatives to provide archival storage infrastructure to member institutions to preserve their digital collections which represent the collective memory of their institutions and communities. For instance, the Council of Prairie and Pacific University Libraries' WestVault distributed digital preservation storage network provides a high-redundancy peer storage network across all four Western provinces.[185] As well, OCUL Scholars Portal's Permafrost digital preservation service provides integration between Archivematica software,[186] and the Ontario Library Research Cloud for secure, long-term preservation storage. Supported by this infrastructure, academic libraries and archives may form partnerships with researchers to accept research data collections they deem of value to their respective organizational missions and mandates, to be managed under their long-term digital preservation strategies.

A 2015 report by RDC found that the biggest gap in the Canadian RDM landscape was the availability of archival storage.[187] A primary cause being that no single organization has the mandate to fund and support the provision of archival storage. A national strategy for archival storage and preservation of research data must recognize the role of decentralization in risk mitigation as part of responsible stewardship. The "lots of copies keep stuff safe" (or LOCKSS) rule and practice of using geographically distributed storage locations helps ensure data recovery in the event of disaster. This model also favours forming partnerships between organizations from institutional, to regional, to national levels, to ensure resilience and long-term sustainability. To achieve efficiencies in a decentralized model, national coordination among new and existing organizations would be necessary to oversee storage and preservation.

The CARL-Portage Preservation Expert Group's White Paper on a national research data preservation model notes that overseeing the provision of active, repository, and archival storage as part of a federated national storage strategy will introduce substantial efficiencies.[188] Their proposed model, argues against imposing homogeneity in archival storage architecture in favour of supporting the coordination of 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 for maintenance and security, while leaving the operation and maintenance of individual PSPs to their host institutions. This approach would leverage existing institutional and organizational capacity, expertise, and investment in support of the broader archival storage strategy for the

---

[183] https://www.bc.net/service-catalogue/educloud-backup

[184] https://www.scinethpc.ca/high-performance-storage-system-hpss/

[185] https://coppul.ca/westvault

[186] https://www.archivematica.org/en

[187] Baker, D. et al. (2019). Research Data Management in Canada: A Backgrounder. Zenodo. http://doi.org/10.5281/zenodo.3341596

[188] Qasim, U. et al. (2018). Research Data Preservation in Canada: A White Paper. https://dx.doi.org/10.14288/1.0371946

country. It would also help recognize PSPs that represent best practices in specific domains, which may provide the same services to an international community of researchers. This recognition may also speak to the need for a more sustainable level of support for all Canadian PSPs, both national and international.

Preservation storage, and the role of institutional, regional, and national long-term storage solutions, will need to be developed and deployed, requiring ongoing discussions with the relevant communities. As the set of partners and capacity in the current RDM ecosystem becomes clearer, it will be possible to coordinate repository and preservation storage needs. These efforts would not only improve stewardship of and access to research data but would also achieve economies of scale that would help build a sustainable network of research data storage.

Interoperability

Achieving interoperability between components of the RDM ecosystem relies on common schemas, standards, and protocols for organizing and describing research data and supporting infrastructure. The FAIR principles describe interoperability in terms of combining data from distinct datasets, as well as integrating with platforms and applications for analysis, storage, and processing.[189] This concept of interoperability should also be extended to include other artefacts created during the research process, including software code, lab protocols, and scientific workflows, as well as the overarching policies and administration practices governing their creation and management. In order to maximize the potential of research data, it must be able to be exchanged securely and integrated between different systems, while being interpreted correctly and appropriately by different users.

The draft EOSC Interoperability Framework (EIF) describes four types of interoperability necessary for governance of national and pan-national data services: Technical, Semantic, Organizational, and Legal.[190] (For a discussion on aspects of Organizational and Legal interoperability, see the Governance section below).

- Technical interoperability describes the ability of applications and infrastructures to exchange data and complete necessary tasks without operator intervention. According to the EIF, this may include interface specifications, data integration services, data presentation and exchange, and secure communication protocols.

- Semantic interoperability describes “the ability of computer systems to transmit data with unambiguous shared meaning”.[191] This relies on common research artefacts being adopted across entire research communities, including metadata schemas and

[189] Wilkinson, M.D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018. https://doi.org/10.1038/sdata.2016.18

[190] Corcho, O. et al. (2020). EOSC Interoperability Framework (v1.0). https://www.eoscsecretariat.eu/sites/default/files/eosc-interoperability-framework-v1.0.pdf

[191] Heikki, L. et al. (2019). D2.1 Report on FAIR requirements for persistence and interoperability 2019. Zenodo. https://zenodo.org/record/3557381

ontologies. This also requires maintenance of registries of these artefacts within research communities to support their discovery and use.

For the purposes of describing the RDM Ecosystem, elements of technical and semantic interoperability are described according to categories of Architecture and Access.

Architecture

To support semantic and technical interoperability between components of the DRI ecosystem, operating frameworks are required that define the procedures, terms, and relationships necessary to allow data to be exchanged unencumbered between digital research infrastructures. These provide an architecture to the ecosystem, which allows new data, software, and infrastructure to be developed and integrated by conforming to these existing frameworks.

Standards and Schemas

The term schema describes a common framework modelling the structures and relationships between a series of related information elements. Once a schema is formalized by a recognized organization, it becomes a standard: a document that applies collectively to codes, specifications, recommended practices, classifications, test methods, and guides, which have been prepared by a standards developing organization or group, and published in accordance with established procedures.[192] Promoting the adoption or advancement of existing standards, rather than contributing to the expansion of custom or novel schemas, is an important consideration for national service development. Research communities that are characterized as being primarily served by long tail data or who don't have a rich culture of data management may be well served by looking for proven best practices within the big data scientific communities of physics, astronomy, earth observation and 'omics, who have by necessity created mature standards and systems to manage complex data and associated services.

- At the research data-level, standard file formats and metadata schemas widely adopted within research communities support data to be collected in consistent ways, recombined across datasets, and allow for shared curation and preservation best practices to be developed. There are many domain specific metadata standards in use by research communities, which are often developed and advanced by related domain associations (e.g., GA4GH,[193] DDI Alliance[194]). Mapping these standards to conceptual metadata frameworks or standards (e.g., ISO 11179[195]) is a key priority for increased interoperability.

[192] "Standard", CASRAI Glossary. (Retrieved November 2020). https://casrai-test.evision.ca/?page_id=485

[193] https://www.ga4gh.org

[194] https://ddialliance.org

[195] https://www.iso.org/standard/68766.html

- Standards for research software are also needed to support interoperability of research data, across software platforms and stacks (e.g., REFI-QDA Standard[196]).

- At the infrastructure level, standard communications protocols are needed for exchanging data and metadata (e.g., OAI-PMH[197]) between systems.

Throughout the ecosystem, developers and infrastructure managers adopt and advance frameworks through national and international standards bodies (e.g., International Organization for Standardization (ISO), Standards Council of Canada (SCC), National Information Standards Organization (NISO)). Adoption of these elements are advanced organically within research communities, as the best-suited frameworks or related tools rise through the ranks or are purposefully advanced by research organizations. Increased awareness of standards and schemas is lacking in many domain communities and should be advanced nationally in collaboration with domain associations to support greater adoption.

Work is currently underway to propose coordinated standardization activity across stakeholder groups in Canada (e.g., government, academia, industry) by the Canadian Data Governance Standardization Collaborative, established in May 2019.[198] Its role will be to produce a roadmap articulating gaps and needs in the landscape and identify priority areas to address, where standards and conformity are needed.

Ontologies

Ontologies represent, name, and define the categories, properties, and relationships between entities in a given subject domain.[199] Greater use of ontologies to formally represent categories, properties and relationships between concepts, data and entities in a dataset is a key component in advancing semantic interoperability. To interpret and use appropriately within and across domains, research data and metadata must have clear meanings that are expressed in machine-readable ways to abide by the FAIR principles. This can be accomplished via artifacts such as controlled vocabularies and thesauri, and related taxonomies.

A range of ontologies can exist within a given domain, as becomes clear when exploring a resource such as BioPortal, maintained by the U.S. National Centre for Biomedical Ontology.[200] Ontological artefacts arise from a range of sources, related to research areas, use of instruments, systems, or methodologies, and are advanced by national and international associations and shared community initiatives. For example, the Open Biomedical and Biological Ontologies (OBO) Foundry was initiated in 2007 by ontology developers committed to open collaboration and

[196] https://www.qdasoftware.org

[197] https://www.openarchives.org/pmh

[198] https://www.scc.ca/en/flagships/data-governance

[199] "Ontology (information science)", Wikipedia. (Retrieved November 2020). https://en.wikipedia.org/wiki/Ontology_(information_science)

[200] https://bioportal.bioontology.org

adherence to shared principles.[201] Domain-focused examples include the MMI Ontology Registry and Repository for marine sciences,[202] and the ESIP Community Ontology Repository for earth science.[203] In response to the COVID-19 pandemic, CIDO,[204] a community-driven open-source biomedical ontology in the area of coronavirus infectious disease, has also been launched.[205]

Like standards and schemas, awareness of the existence and importance of these semantic artefacts for interpretation and reuse must be advanced across domains to support greater interoperability. Common languages and crosswalks across domains to express semantic differences between shared concepts is also needed to support wider interoperability. Utilizing AI and machine learning tools for ontology development and alignment is an active area of research.[206]

Registries and Indexes

Registries are resources containing identifiers assigned to items with descriptions of the associated items and play an important role in connecting and directing users and machines throughout the RDM ecosystem.[207] They benefit not only researchers searching for tools and resources for their work, but also developers and administrators searching to integrate components into systems and infrastructures. Persistent identifiers (PIDs), a core enabler of these resources, are further discussed under the Access category.

Indexes are similar to registries without registered identifiers associated with listed items. Both registries and indexes exist or are in development for a range of components in the RDM ecosystem. As mentioned above, BioPortal is an example of a resource indexing ontologies in the biomedical field. Fairsharing.org is another curated index of metadata standards, inter-related to databases and data policies.[208] PRONOM is a registry of file formats and related technical information.[209]

The Interoperability Service Reference Framework developed by the EOSC-supported ELIXIR (the European life-sciences Infrastructure for biological Information) bioscience RDM

---

[201] http://www.obofoundry.org

[202] https://mmisw.org

[203] http://cor.esipfed.org

[204] CIDO: Coronavirus Infectious Disease Ontology, GitHub. https://github.com/CIDO-ontology/cido (Retrieved November 2020)

[205] He, Y. et al. (2020). CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. *Sci Data* 7, 181. https://doi.org/10.1038/s41597-020-0523-6

[206] Gromann, D., Espinosa Anke, L., & Declerck, T. (2019). Special Issue on Semantic Deep Learning. 815 – 822. https://content.iospress.com/articles/semantic-web/sw190364

[207] "Registry", CASRAI Glossary. https://casrai-test.evision.ca/glossary-term/registry/ (Retrieved November 2020)

[208] https://fairsharing.org/

[209] https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx

organization, provides a model for developing a registry service that supports people and machines to discover, access, integrate and analyse research data.[210] Similar resources should be supported at both national and domain levels, in order to support a range of DRI ecosystem components.

Certifications

Capabilities and levels of trust in DRI, as well as the research data it supports, vary according to requirements of both systems and users. Certifications provided by authorities trusted by a community provide a warranty of conformance with provisions of referenced standards, codes, or other requirements.[211] One certification of quality assurance is ISO certification, which relies on the availability of underlying ISO standards against which to evaluate.[212]

In the data repository community, the CoreTrustSeal has advanced as the leading certification for trusted repositories.[213] The CoreTrustSeal offers to any interested data repository a core certification based on the Core Trustworthy Data Repositories Requirements catalogue,[214] developed between the Data Seal of Approval and the World Data System under the scope of the Research Data Alliance to harmonize their own data repository certifications. Supporting Canadian research data repositories to achieve the high standards of the CoreTrustSeal is an anticipated initiative of the Portage Network, which will help to improve the quality and standards of Canadian repositories nationally.

At the level of research data and research methods, new certification models are being considered to assert quality and reproducibility in published research. For instance, the Certification Agency for Scientific Code and Data (CASCAD) supported by the French National Science Foundation and a consortium of French research institutions.[215] Assessing reproducibility and quality of confidential data and related analyses is particularly challenging for sensitive data. For this reason, CASCAD and the Centre d'Accès Sécurisé aux Données (a counterpart to Canada's CRDCN) have collaborated to design a reproducibility certification process for confidential data.[216] Similar entities and relationships could be explored in the Canadian context.

---

[210] https://elixir-europe.org/platforms/interoperability

[211] "Certified product", CASRAI Glossary. https://casrai-test.evision.ca/glossary-term/certified-product (Retrieved November 2020)

[212] https://www.isoqsltd.com/faq

[213] https://www.coretrustseal.org

[214] ICSU World Data System. (2016). Core Trustworthy Data Repositories Requirements v01.00. https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf

[215] https://www.cascad.tech

[216] Pérignon et al. (2019). Certify reproducibility with confidential data. *Science*, 365(6449):127-128. https://science.sciencemag.org/content/365/6449/127

Identification and Access

Due to the scale and complexity of national and international research, controls that define relationships between entities and access permissions also play an important role in connecting the RDM ecosystem.

PIDs are emerging as important elements supporting RDM. PIDs are long-lasting references to unique objects that give information about that object independent of individual institutions or infrastructure implementation.[217] PID systems are thus becoming the preferred way for referring to and accessing entities within the DRI ecosystem unambiguously and sustainably. The range of objects PID systems could be developed for is vast, encompassing researchers, datasets, scientific instruments, and other elements.

While PID systems can be both implemented and managed locally within research institutions, the implementation of widely adopted, third-party administered systems presents a more unified approach for connecting data and infrastructure across the ecosystem. In Canada, two PID systems have gained national endorsement.

- ORCID (Open Researcher and Contributor ID) is a PID system for uniquely identifying researchers, which has been adopted by major publishers, funders, and research institutions globally.[218] The ORCID PID is thereby able to connect researchers to the research ecosystem, via their publication record, employment history, awards, collaborations, and other outputs. The ORCID researcher ID system is endorsed by the ORCID Canada Consortium, who provide research organizations with an ORCID membership at a reduced cost, and also provide community support and engagement.[219] It currently has 37 members.

- DataCite is global provider of a Digital Object Identifier PID system for datasets, which has been implemented across many digital repositories.[220] The DataCite Canada Consortium, a recent initiative led by the Canadian Research Knowledge Network and CARL Portage, share the costs associated with Datacite membership among 45 member organizations, while providing central administration and community support.[221]

---

[217] Koster, L. (2020). Persistent identifiers for heritage objects. *Code4Lib*, Issue 47. https://journal.code4lib.org/articles/14978

[218] https://orcid.org

[219] https://orcid-ca.org/about

[220] https://datacite.org

[221] https://www.crkn-rcdr.ca/en/datacite-canada-consortium

Other types of emerging international PID systems include Research Activity IDs (RAID),[222] Scientific Instrument IDs (PIDINST),[223] and Research Organization IDs (ROR).[224]

To be effective, PIDs must be available to be integrated in all systems and interfaces used for describing the objects in question, both for human end users and machines. However, even the most widely adopted PID is only as persistent as its system's administration. Consciously advancing the adoption of recommended PID systems and finding sustainable cost models, such as the consortium approaches described above, are key to long term availability of PID systems.

When PIDs are interoperable with identity systems, they can also support systems of authentication and authorization. Within Canadian higher education institutions, network resources provided by CANARIE and Canada's NREN support identity and access management solutions for Canadian research and education institutions and content providers. For instance, CANARIE's Canadian Access Federation is a trusted access management environment that provides researchers across Canada connectivity to third-party platforms using the identity provider credentials of their home research institutions.[225] Canada's federated identity service is expanded internationally by their participation in eduGAIN, an interfederation service that connects identity federations globally.[226] Cybersecurity is a related element of the RDM ecosystem, including mechanisms and strategies that prevent unauthorized access to digital assets, from the level of data to networks. Cybersecurity is being advanced jointly by the Alliance and CANARIE, and will be discussed in other outputs.

Data Services

As the needs of researchers have grown in response to advances in technology and research practices, as well as new expectations from host institutions, funders, and journal publishers, a range of related services have been developed to support them in accessing DRI to improve RDM through adopting new practices and tools into their research workflows. Services are offered within academic research institutions, by infrastructure providers in association with specific platforms and tools they offer, via regional and national associations and service providers, and increasingly by commercial entities. The range of available services span the research data lifecycle, targeting the management of research data in all of its forms.

As described in the above Storage and Compute section, many research groups are also operators of RDM infrastructure, particularly project-based or domain-based repositories, making their research data discoverable and accessible to the wider research community. Thus, models are also being developed to support research organizations with developing and operating DRI for RDM.

---

222 https://www.raid.org.au

223 https://www.rd-alliance.org/group/persistent-identification-instruments/case-statement/persistent-identification-instruments

224 https://ror.org

225 https://www.canarie.ca/identity/caf

226 https://technical.edugain.org/status

Support for Research Data

Planning

A Data Management Plan (DMP) is a formal statement describing how research data will be managed and documented throughout a research project and the terms regarding deposit of the data with a data repository for long-term management, sharing, and preservation.[227] The development of a DMP from the outset of a research project can help identify and thereby mitigate issues in the management of research data generated throughout a project, and is therefore considered a best practice of RDM. DMPs are expected by research grant and awards agencies and academic institutions in the US, UK, and Australia.[228] It is anticipated to be a funding requirement as part of the Tri-Agency's forthcoming RDM policy.

In 2015, CARL-Portage led the creation of an online platform to assist in the creation of DMPs, called the DMP Assistant, which is bilingual and freely available to researchers across the country.[229] While the University of Alberta Libraries hosts and maintains the platform, individual institutions and research groups manage local templates for their respective research communities. As of July 07, there were 52 institutional accounts, and 12,489 researchers with individual accounts. Responsibilities for outreach, support, and engagement are shared between the Portage secretariat and the institutions who have locally implemented the tool. Within academic institutions, local support for researchers using the DMP Assistant and developing DMPs is often shared between libraries and research offices.

While DMPs have been an established practice for more than a decade, they are primarily human readable documents, which limits their usefulness beyond the authoring research group. The next generation of DMPs are expected to be machine actionable, which will allow them to better support the research enterprise through interoperability with other research systems.[230] For instance, integrating with Research Ethics Board applications, notifying storage providers of capacity needs, and predicting preservation processes, are just some of the expected benefits.

An update to the DMP Assistant is expected in Fall 2020 using the merged code bases of other existing international DMP platforms (known as the DMP roadmap) that will include API functionality, setting the stage for greater machine readability and improved interoperability.[231]

---

[227] "Data Management Plan", CASRAI Glossary. (Retrieved November 2020). https://casrai.org/term/data-management-plan

[228] Unsworth K., & Smale, N. (2017) Overview of Australian institution and UK/USA funding body data management plan mandates. University of Melbourne. Dataset. https://doi.org/10.4225/49/5986bde74f8f5

[229] https://assistant.portagenetwork.ca

[230] Simms, S. et al. (2017). Machine-actionable data management plans (maDMPs). *Research Ideas and Outcomes 3*: e13086. https://doi.org/10.3897/rio.3.e13086

[231] DMPRoadmap, GitHub repository. https://github.com/DMPRoadmap (Retrieved November 2020)

Curation

A range of data curation services have emerged across the RDM ecosystem in response to both advances and availability of infrastructure, as well as changes in researcher attitudes and requirements towards long term data management and sharing. Data curation is best thought of as the iterative process of optimizing datasets for current use as well as future discovery and reuse, guided by the conceptual framework of the FAIR Principles.[232] Good curation practices support the management of data throughout its lifecycle, as data and collections are cleaned, documented, standardized, inter-related, stored, and shared. Many skill sets are drawn upon in supporting this process, including disciplinary knowledge, familiarity with the research process, knowledge of metadata standards and best practices in data management, and abilities for working with various technologies.

Researchers are encouraged to document data collection, manipulation, and analysis processes during the active part of their research process, to ensure accurate and full metadata is collected, in order to support later management actions, like publication, preservation, and reuse. However, currently many published datasets are inconsistently documented, and curation is only thought of at the end of a research project. Given the scope of data curation, researchers may derive support for data curation from numerous sources. The process is typically informed by the need to make a given dataset functional with another element of the RDM or broader DRI ecosystems (e.g., integrating with another dataset, adhering to a metadata standard, operating with research software, depositing to a repository, or sharing with colleagues). While some data curation can be accomplished through use of research software (which should continue to be investigated as a scalable means to support researchers) given the complexity of the process, it usually requires some form of human intervention.

Within higher education institutions, support for curating research data can be found internal to research groups. For instance, many groups with significant data assets employ data managers to support curation. Support may also be sought from Systems/IT departments, to curate data to integrate with available research computing infrastructure. Libraries and archives also provide curation support to researchers depositing data to institutional repositories or other hosted digital asset management systems, as well as external data repositories. A recent survey of the current capacity of Canadian higher education institutions to provide RDM services revealed slightly more than half of the 77 respondents provided some form of curation support via their libraries.[233] The same survey revealed between one quarter and one third of respondents provide data computing services and/or technical support (e.g., data encryption/anonymization) from System or IT departments. An important gap apparent across many research institutions is their limited support for curation and storage of sensitive data, particularly relevant in light of the recent funding programs for COVID-19, which require open sharing of outputs and data.[234]

---

[232] Portage Network. (2019). Primer: Data Curation. https://portagenetwork.ca/wp-content/uploads/2019/09/Curation_Primer_Aug2019_EN.pdf

[233] Cooper, A. et al. (2020). Institutional Research Data Management Services Capacity Survey. http://doi.org/10.14288/1.0388722

[234] Tayler, F. & Ripp, C. (2020). FAQ: COVID-19 Rapid Response Data Sharing and Deposit Support. Scholars Portal Dataverse. https://doi.org/10.5683/SP2/522KV2

Discussions among participants of the 2019 Canadian Data Curation Forum highlighted the need for a clear understanding of what level of curation institutions are prepared to offer.[235] One model of scaling the capacity and specialization needed to support data curation across numerous domains represented within higher education institutions is offered by the Data Curation Network (DCN).[236] Institutions that are members of the DCN share human resources with other member institutions, effectively pooling their time and diversity of expertise. This enables all member data repositories to collectively, and more effectively, curate a wider variety of data types (e.g., discipline, file types, software, etc.) that expands beyond what most individual institutions could offer alone.

Perhaps the most common type of curation support that researchers experience is when they decide to publish their data in a repository. Most data repositories, whether general or domain-specific, provide some level of curation support to researchers relying on their infrastructure. There are many different software platforms employed by data repositories, which enable or automate some elements of curation during the deposit process. Authors of a recent study break down some of these functions according to the RDA Repository Interest Group model.[237]

Data repositories also provide more direct support to researchers through their data curators and support teams. Support provided may range from instruction on using their platforms and interpreting supported metadata schemas, to more heavy lifting involving curation of the dataset files, as well as related code and documentation. For example, a recent partnership in the U.S. between the National Institutes of Health (NIH) and figshare to pilot a new generalist repository for research funded by NIH grants without a designated domain-specific repository for their data.[238] In this pilot, depositors are paired with trained data librarians who review metadata and licensing.[239] This level of curation contrasts more hands-on support, typically provided by more domain-focused repositories. For instance, researchers seeking to deposit data to the International Neuroimaging Data-Sharing Initiative (INDI) are instructed to contact the repository first to receive detailed instructions for contributing data.[240] In their model, curators are available to help troubleshoot during the deposit process, develop customized scripts for researchers to

[235] Brodeur, J., Sawchuk, S., & Newson, K. (2020). Materials from the 2019 Canadian Data Curation Forum. Zenodo. http://doi.org/10.5281/zenodo.3899401

[236] https://datacurationnetwork.org

[237] Kim, S. (2018). Functional requirements for research data repositories. *IJKCDT*, 8(1), 25-36. http://ijkcdt.net/xml/13281/13281.pdf

[238] https://nih.figshare.com

[239] Hyndman, A. (2019). Figshare announces data repository partnership with the National Institutes of Health to store and reuse research data. https://figshare.com/blog/Figshare_announces_data_repository_artnership_with_the_National_Institutes_of_Health_to_store_and_reuse_research_data/518 (Retrieved November 2020).

[240] http://fcon_1000.projects.nitrc.org

automate data preparation, or can carry out data preparation for the researchers, if approved by the researcher's ethics board.[241]

Models for when curation intervention takes place also vary. For instance, datasets deposited to FRDR or Dryad, both generalist repositories, are queued for review by a Curator before final publication. Curators review the dataset for completeness of the metadata, organization, and reusability of the data files, ensure that files are not corrupt and do not contain sensitive information, and otherwise adheres to terms of use.[242] Meanwhile, researchers depositing data to the Qualitative Data Repository (QDR) are suggested to make an initial deposit with just the bare minimum information early on in their project to trigger an ongoing consultation process.[243]

Data repositories that have the infrastructure and experience necessary to handle datasets that may contain sensitive information play an important role in curating data in accordance with established standards, laws, and ethics boards requirements. For instance, when submitting data to the Cancer Imaging Archive, researchers are instructed to anonymize and encrypt data prior to submission. All deposited datasets are treated as though they may contain sensitive information and are first captured in a secure system, where they are then reviewed by curators trained in health information and privacy regulations.[244] Other repositories that can handle sensitive data perform similar disclosure risk reviews. For instance, curators at ICPSR will suggest methods to modify the data to limit risk, or suggest sharing the data at a higher level of restriction within their repository.[245]

Greater coordination of data curation efforts across organizations and the DRI ecosystem would lead to many benefits for the research data, related platforms, and skills of curators themselves. In Canada, Portage is leading national coordination efforts for data curation via a model Canadian Data Curators Network. Opportunities for national coordination were discussed at the 2019 Canadian Data Curation Forum, leading to the publication of a summary report with recommendations directed at the Alliance for advancing national support for data curation. These recommendations include investing in the development of human capacity within research

---

[241] INDI. (n.d.). Data Contribution Guide v3.1. http://fcon_1000.projects.nitrc.org/indi/indi_data_contribution_guide.pdf

[242] FRDR (n.d.). Documentation: After Depositing. https://www.frdr-dfdr.ca/docs/en/after_depositing (Retrieved November 2020); and
Dryad (n.d.) Dryad Submission and Publication Process. https://datadryad.org/stash/submission_process#curation (Retrieved November 2020)

[243] https://qdr.syr.edu

[244] Cancer Imaging Archive. (2020). Submission and De-identification Overview. https://wiki.cancerimagingarchive.net/display/Public/Submission+and+De-identification+Overview (Retrieved November 2020).

[245] ICPSR. (n.d.). Data Confidentiality. https://www.icpsr.umich.edu/web/pages/datamanagement/confidentiality/index.html (Retrieved November 2020)

organizations, as well as shared computing infrastructure, from which national coordination, communication, and representation could be advanced via a national network approach.[246]

Preservation

In contrast to curation, digital preservation relies on a range of strategies to support the long-term maintenance of access to materials beyond the limits of media failure or technological change.[247] There is no end-state at which one may claim that a digital object is finally preserved. A range of preservation activities are necessary to manage the variety of digital objects associated with a given dataset or wider research project. Analyzing these activities in isolation from other elements of the DRI landscape is challenging, as the act of digital preservation relies on many components working together to ensure that digital information can be successfully preserved.

Whether through safeguarding integrity of files, functionality of software components, or against obsolescence of storage infrastructure, digital preservation is a spectrum of work. The National Digital Stewardship Alliance (NDSA) presents their Levels of Preservation matrix as a series of functional areas with successive phases of preservation actions, from knowing the digital content (Level 1), to protecting it, monitoring it, through to sustaining it (Level 4).[248]

The development and integration of software into DRI can enable preservation services. For instance, Archivematica is a suite of open-source tools that implement discrete tasks on ingested digital objects to produce archival-ready outputs.[249] This software is currently integrated into both of Canada's national data repository options (FRDR and Scholars Portal Dataverse), allowing curators to apply preservation processing to datasets and migrate outputs to archival storage.

However, while software can enable preservation processes, it is ultimately a human resource intensive practise requiring attention, verification, and maintenance. Currently, services supporting preservation of research data are limited within academic research institutions. Across campuses, relatively few libraries and archives in Canada staff positions to support research data preservation.[250] However, regionally, academic library consortia have advanced shared support models. COPPUL's Digital Stewardship Network and OCUL's Permafrost service both provide tools, training, and support to member institutions to preserve their digital collections.

While research organizations operating domain-specific data repositories are primarily concerned with the long-term preservation of their own digital objects, interesting service models are being advanced to support the needs of their communities. For instance, Population Data BC provides

---

[246] Clary, E., et al. (2020). Conceptualizing a National Approach to Data Curation Services in Canada. Zenodo. http://doi.org/10.5281/zenodo.3894935

[247] "Digital Preservation", CASRAI Glossary. https://casrai.org/term/digital-preservation (Retrieved November 2020)

[248] NDSA. (2019). Levels of Digital Preservation. https://ndsa.org/publications/levels-of-digital-preservation (Retrieved November 2020)

[249] https://www.archivematica.org/en

[250] Cooper, A. et al. (2020). Institutional Research Data Management Services Capacity Survey. http://doi.org/10.14288/1.0388722

archival storage of older versions of data for historical reference.[251] They also are able to store archival copies of data extracts in order to support researchers while meeting Research Ethics Board requirements for approved retention periods.[252] The Canadian Astronomy Data Centre supports complex workflows to archive both the raw data obtained from instruments as well as data products that result from processing through science pipelines to transform massive amounts of data into usable products by researchers.[253]

Extending existing services to integrate into a nationally supported model is the rationale behind Portage's distributed Preservation Service Provider model,[254] which proposes shared infrastructure at scale while enabling local curation and decision making around individual collections. This would allow for economies of scale around development and maintenance of physical infrastructure, without disregarding unique knowledge and objectives about collections held locally. While still in the proof-of-concept stage, a key requirement to realizing sustained national preservation efforts will be enabling the interoperability of existing systems and practices, which will rely upon the creation and adoption of frameworks and standards agreed upon by the communities involved.

Discovery

A primary motivation behind RDM is to enable sharing and reuse of existing datasets, and an important yet often overlooked service is support for researchers and organizations to discover research data of interest. The concept of discovery covers support for searching, identifying, interpreting, and accessing published datasets of interest. Within institutions of higher education, research libraries have played an important role in supporting researchers to identify and use publicly accessible data of interest for many years. An important milestone in Canada was the formation of Data Liberation Initiative,[255] a partnership between Statistics Canada and higher education institutions to support access to STC published datasets.

The growing number of data repositories requires a concerted effort to inventory repositories and their data assets across Canada and internationally. In Canada, the Federated Research Data Repository (FRDR) platform is designed to operate as a national data discovery layer, by harvesting metadata from identified Canadian data repositories into a national search engine. In this way, FRDR provides exposure for large and small data repositories alike and drives user traffic to hosts. The Portage Data Discovery Expert Group is leading an effort to identify data

---

[251] https://www.popdata.bc.ca

[252] Population Data BC. (n.d.). Archival back up and storage of data. https://www.popdata.bc.ca/dataproviders/services/archivalstorage (Retrieved November 2020)

[253] Canadian Astronomy Data Centre. (2020). Archive as a Service. https://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/doc/AaaS/ (Retrieved November 2020)

[254] Qasim, U. et al. (2018). Research Data Preservation in Canada: A White Paper. https://dx.doi.org/10.14288/1.0371946

[255] https://www.statcan.gc.ca/eng/dli/dli

repositories located across Canada, and to support improved awareness though integration into the FRDR discovery service.[256]

Internationally, efforts to index published datasets can be found in the Data Citation Index, a commercial tool from Clarivate,[257] and Google's Dataset Search tool,[258] which harvests research data (among other types) from across the web that are described with the schema.org standard. OpenAIRE Explore research graph tool goes one step beyond by linking together metadata from a massive source of scientific outputs, including datasets, software, articles, with information about research organizations, funders, and partners.[259] Portage's Federated Research Data Repository is integrating with both OpenAIRE and the Data Citation Index to ensure that Canadian research data are more discoverable internationally.

One of the primary issues associated with these efforts to index and make data widely discoverable is that they, by necessity, tend to rely on simplified metadata standards to index content. In contrast, domain specific metadata is much more detailed and richer, making it easier for researchers to assess the relevance of data and determine if it is fit for purpose. Going forward there will likely be an increased emphasis on either indexing richer metadata content or extending simple metadata standards. For example, there are several community groups creating extensions to schema.org to increase usability in the biosciences,[260] and earth sciences.[261]

Exploration

Rather than simply providing data files to researchers to download and reuse locally, software integrations into data repositories are enabling users to display, manipulate, and interpret data within the existing platform. This concept of "data visitation" has many benefits for long-term curation and preservation. For example, in response to the COVID-19 crisis, a set of Fair Data Points (FDPs) were created under the umbrella of the Virus Outbreak Data Network (VODAN) initiative.[262] FDPs are FAIR data repositories with 'docking' capabilities that accept virtual machines that come to 'visit' the data locally, with a specific question or processing task to execute.[263] The result of the task, not the data, is returned to the initiating client. Because research data does not leave the repository, the proliferation of derivative datasets that also need to be curated and preserved over time is prevented. Rather, data configurations or analyses can be reproduced within the repository platform. Other examples of repositories that support data

---

[256] https://portagenetwork.ca/network-of-experts/data-discovery-expert-group

[257] https://clarivate.com/webofsciencegroup/solutions/webofscience-data-citation-index

[258] https://datasetsearch.research.google.com

[259] https://explore.openaire.eu

[260] https://bioschemas.org

[261] https://github.com/ESIPFed/science-on-schema.org

[262] Mons, B. (2020). The VODAN IN: support of a FAIR-based instrument for COVID-19. Eur.J. Hum. Genet 28, 724–727. https://doi.org/10.1038/s41431-020-0635-7

[263] FAIR Data Point, GitHub. https://github.com/FAIRDataTeam/FAIRDataPoint (Retrieved November 2020).

visitation include Stats Can's RTRA program,[264] Barcode of Life Data System,[265] OpenNeuro,[266] ICPSR's Virtual Data enclave,[267] and the Data and Information Access Services (DIAS) hubs for the EU's Copernicus Earth Observation Programme.[268] In fields that produce massive amounts of data that require specialized training to manipulate into meaningful outputs before being analyzed, data visitation also presents opportunities to increase accessibility of the data and increase scientific outputs.[269]

The growing adoption of programs like Jupyter and R markdown, which produce dynamic computing documents containing live code and descriptive text, combined with container tools like Docker,[270] and Binder,[271] are also allowing data re-users to interact with published datasets directly. Support for these software tools may be increasingly provided by repositories to support computational reproducibility (for e.g., the Git - Zenodo - Binder integration).[272] Support for a national Jupyter interactive computing environment is currently being provided by Compute Canada and Cybera,[273] and is integrated with more than 20 member higher-education institutions.[274]

Infrastructure Support

A range of computing infrastructure and software tools enable RDM. The RDM initiatives described above also require support for their own development. This element of the RDM ecosystem has significant overlap with ARC and RS portfolios of the Alliance. Within academic institutions, Systems and IT departments offer support initiatives hosted locally, while national service providers Compute Canada and CANARIE offer support for nationally hosted initiatives. RDM-specific development opportunities are relatively infrequent. One recent example is CANARIE's Research Data Management Program,[275] which funds the development of new software tools to support researchers with integrating or adopting RDM best practices into their workflows.

---

264 https://www.statcan.gc.ca/eng/rtra/rtra

265 https://www.boldsystems.org/index.php

266 https://openneuro.org

267 https://www.icpsr.umich.edu/web/pages/NACJD/virtual-data-enclave.html

268 https://www.copernicus.eu/en/access-data

269 Kern, J., Glendenning, B., & Robnett, J. (2019). The Science Ready Data Products Revolution at the NRAO. https://science.nrao.edu/science/astro2020/apc-white-papers/136-78e9d3c08d7e3e149ce77c6fdd6b9366_KernJeffreyS.pdf

270 https://www.docker.com

271 https://mybinder.org

272 https://blog.jupyter.org/binder-with-zenodo-af68ed6648a6

273 https://www.cybera.ca

274 https://syzygy.ca/

275 https://www.canarie.ca/rdm/

Models for national support can also be gathered from international organizations. For example, the World Data System International Technology Office supports members with infrastructure and services to support RDM and to examine data holdings.[276] Identifying opportunities to support research organizations in advancing their own RDM tools and platforms while finding ways of widely applying their successes more broadly across the ecosystem will be necessary in providing truly national data services.

Training

The digital shift in the research enterprise has yielded significant needs for training and upskilling in both researchers and research support professions. The range of digital skills needed encompass both data science and data stewardship aspects. The European Commission's report on Turning FAIR into Reality defines data science as "the ability to handle, process and analyze data to draw insights from it",[277] drawing on skills in computer science, software development, and statistics. Whereas data stewardship is defined as "skills to ensure data are properly managed, shared, and preserved throughout the research lifecycle." While all researchers require competencies in these skill sets, the inclusion of specialist positions into research projects is increasingly recognized as a valuable mechanism for supporting digital research. For instance, it was recently estimated that 1 in 20 members of a research workforce should be digital support specialists.[278]

The growth in data-intensive research across domains has revealed a significant gap in training in post-secondary institutions related to the adoption of good RDM practices, among other digital skill sets.[279] The recent OECD report on building skills capacity for digital science argues that universities have key roles to play as the primary providers of research training.[280] In particular, university libraries are best placed to support the development of data stewardship skills, while computing departments have much to contribute to software and computing skills underlying data science. While the gap in RDM skills is reducing through the efforts of higher education institutions to invest in more training for researchers, capacity for this work is mostly concentrated in large universities.[281] Government mandates and incentives are one set of mechanisms to encourage growth among institutions.

---

[276] https://wds-ito.org/

[277] European Commission. (2018). Turning FAIR into reality: Final report and action plan from the European Commission Expert Group on FAIR Data. https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf

[278] Mons, B. (2020). Invest 5% of research funds in ensuring data are reusable. *Nature,* 578, 491. https://www.nature.com/articles/d41586-020-00505-7

[279] The Leadership Council. (2014). Canadian DI Environmental Scan: A Supplement to the Background Précis Document Provided to DI Summit 2012 (as cited in Baker et al. (2019), accessible at https://zenodo.org/record/3574685)

[280] OECD. (2020). Building digital workforce capacity and skills for data-intensive science. *OECD Science, Technology and Industry Policy Papers.* https://doi.org/10.1787/e08aa3bb-en

[281] Cooper, A. et al. (2020). Institutional Research Data Management Services Capacity Survey. http://doi.org/10.14288/1.0388722

Research associations and academic societies have important roles to play in building communities and in scaling the provision of training opportunities. For instance, opportunities to bridge gaps in access across institutions have been undertaken by Compute Canada's regional partners (e.g., WestGrid Summer/Winter School series,[282] ACENET Training Catalogue[283]) and by CARL Portage's training program.[284] National or regional approaches to training can assume a range of models. For instance, Portage provides training opportunities to researchers and data professionals via in-person and online mediums to increase reach. They also produce training materials for data professionals to deliver to local research communities in a "train the trainer" model. A number of research organizations operating data services also offer specialized training for RDM targeting researchers, data managers, and administrators. For instance, the First Nations Information Governance Centre's Fundamentals of OCAP® course,[285] and Population Data BC's Education and Training series.[286] Internationally, a number of RDM training models could be adopted to target key demographics. For example, ELIXIR's Training Registry for the life-sciences,[287] CODATA-RDA's School of Research Data Science targeting early career researchers,[288] or the Research Data Management Librarian Academy.[289] Community initiatives that arise from gaps in mainstream programming should also be recognized as important elements of the training landscape. RDC and CARL-Portage are both collaborating on developing a national approach to RDM training. Coordinated action across the ecosystem will be needed to build and maintain the workforce of highly qualified personnel necessary to advance digital research and support open science objectives.

Governance

Many organizations have assumed roles supporting communities of practice with RDM, through the development of guidance, policies, or funding opportunities. Coordination between these organizations is essential for fostering a diversity of successful approaches to RDM. Within this current landscape, impacts of existing imbalances that exist locally and regionally should be considered in the national context. As well, harmonisation with international initiatives should also be considered to allow data to move across frontiers.

Numerous organizations have provided recommendations, advice, or other information responding to arising issues of importance in the management of research data. Internationally, a number of organizations with both general and more domain-focused research mandates have produced important resources adopted by research communities (see Appendix C). In Canada,

---

[282] https://www.westgrid.ca/support/training

[283] https://www.ace-net.ca/training/

[284] https://portagenetwork.ca/training-resources/

[285] https://fnigc.ca/training/fundamentals-ocap.html

[286] https://www.popdata.bc.ca/etu

[287] https://tess.elixir-europe.org/

[288] https://codata.org/initiatives/strategic-programme/research-data-science-summer-schools

[289] https://rdmla.github.io/

this has largely been led by national associations representing research support entities, in particular, CARL's Portage Network of Experts have produced both reports and more applied outputs for the benefit of the research institutions and the wider RDM community.[290] Within research communities, relatively few Canadian research domain associations or societies have produced resources for their members to guide improved management of research data – especially in less computationally intensive fields of research. RDM practices and considerations vary greatly by domain, so the greater involvement of these groups should be encouraged. The efforts of Research Data Canada to bring together stakeholders from across the landscape in National Data Framework Summits is one such example of encouraging greater involvement with research associations.

As a result of this growing body of intelligence on the value of RDM, researchers and their organizations find themselves faced with a growing set of related policies; for instance, the Tri-Agency's influential draft RDM policy, or those from scholarly publishers related to data sharing and access. In response, research institutions across Canada are gradually developing policies to address the management data resulting from funded research.[291] Within research institutions, existing policies related to research practice, ethics, and intellectual property already have implications on how research data must be managed. Efforts to inventory existing institutional RDM policies are underway in Canada, led by Portage's Research Intelligence Expert Group, as are international efforts convened by RDA.[292]

Consistent policies and requirements for research organizations, research infrastructures and related services are necessary to ensure that researchers adopt common practices and frameworks. Differences in institutional and regional requirements, as well as the need to respect Indigenous rights, contribute to challenges nationally. For example, differences between provinces in how personal information is managed may affect cross-border sharing and collaboration. Both British Columbia and Nova Scotia limit public organizations and their service providers from moving personal information outside of Canada, while other provinces do not,[293] impacting health and social sciences fields, in particular.

The management of research data during and following a given research project draws upon substantial financial resources to support human and technical infrastructure. However, there is limited directed funding for researchers to draw upon in funded projects, as RDM is often not viewed as part of the standard research process, nor part of the normal research budget.[294] More

---

[290] https://zenodo.org/communities/portage-network

[291] Cooper, A. et al. (2020). Institutional Research Data Management Services Capacity Survey. http://doi.org/10.14288/1.0388722

[292] https://www.rd-alliance.org/group/research-funders-and-stakeholders-open-research-and-data-management-policies-and-practices-ig

[293] Office of the Privacy Commissioner of Canada. (2018). Summary of privacy laws in Canada. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/02_05_d_15 (Retrieved November 2020)

[294] Knowledge Exchange Research Data Expert Group and Science Europe Working Group on Research Data. (2016). Funding research data management and related infrastructures. https://www.scienceeurope.org/media/uuqf0i03/se-ke_briefing_paper_funding_rdm.pdf

recently, RDM-specific funding opportunities have arisen to encourage researchers to improve the management of their research data and solve related challenges through capacity building events. For example, SSHRC's Connections Grants for RDM Capacity Building,[295] or IDRC's Open Research Data Initiative.[296]

For some grant-funded research projects, data curation costs incurred during the research period may be included in the project budget. However, ongoing management past the life of a research project ought to be included in indirect costs programs (e.g., Tri-Agency Research Support Fund[297]). Eligible research institutions apply to indirect costs programs to offset costs incurred by managing those awards, and the amounts received are determined by a yearly calculation exercise. However, much of those awards are already spoken for as other areas of a research institution depend on them, and cost calculations can be slow to change relative to the increasing support needed for RDM.[298]

While the policy landscape facing researchers continues to grow, those policies do not specify who is responsible for ongoing management (e.g., researcher, institution, funder), nor do they specify who is to cover costs. Before those questions can be addressed, wider considerations and implications around responsibilities for RDM must be clarified. For instance, what outputs of a research project must be stored and preserved over time? What are the criteria differentiating the storage spectrum (active, repository, archival)? What are the terms around preservation of research outputs (e.g., retention length, responsibilities and criteria for selection, maintenance, and deselection)?

Costs of RDM vary according to specifics of the project (e.g. storage required, sensitivity of data, preservation length),[299] and curation needs through the research lifecycle.[300] The High Level Expert Group on the European Open Science Cloud estimates 5% of a research project's total expenditures should be used for RDM, as a general rule of thumb.[301] However, funding for the computing, storage, and software infrastructure that RDM relies on, where it exists, is primarily intended for development (e.g. CFI Innovation Fund,[302] Compute Canada Resource

---

[295] https://www.sshrc-crsh.gc.ca/funding-financement/programs-programmes/data_management-gestion_des_donnees-eng.aspx

[296] https://www.idrc.ca/en/funding/open-research-data-initiative

[297] https://www.rsf-fsr.gc.ca/administer-administrer/expenditures-depenses-eng.aspx

[298] Erway, R. & Rinehart, A. (2016). If You Build It, Will They Fund? Making Research Data Management Sustainable. *OCLC Research.* https://www.oclc.org/content/dam/research/publications/2016/oclcresearch-making-research-data-management-sustainable-2016.pdf

[299] OpenAIRE. (n.d.). What will it cost to manage and share my data? https://www.openaire.eu/rdm-researcher-costs-infographic/view-document (Retrieved November 2020)

[300] Westerhof, A. et al. (n.d.). Research Data Management: more than just storage. https://www.lcrdm.nl/files/lcrdm/2020-04/RDM%20and%20Costs_v20160218_EN.pdf

[301] LCRDM. (n.d.). xFinancials. https://www.lcrdm.nl/en/xfinancials (Retrieved November 2020)

[302] https://www.innovation.ca/awards/innovation-fund

Allocations,[303] CANARIE Software Development Calls[304]) rather than operations and maintenance, leading to brief funding periods that are incompatible with RDM. Even CFI's Major Science Initiatives Fund, which contributes to ongoing operating and maintenance needs of national research facilities are awarded on a three-to-five-year basis.[305] New models for sustainable funding infrastructure supporting RDM are needed to ensure longevity.

The impact of sustained funding for data repositories on the data management practices of the researchers they support cannot be understated. This effect is evident in research communities with robust, nationally funded RDM platforms, that have become integral resources for research to those communities (e.g., the Canadian astronomy community and the Canadian Astronomy Data Centre, funded via the National Research Council of Canada and the Canadian Space Agency).[306]

In their examination of funding sources for institutional data repositories run by academic libraries, the Online Computer Library Center (OCLC) found a mix of cost models; however, almost half of those contacted relied solely on base funding.[307] A mix of funding models (e.g., structural funding, value added services, pay per use) are necessary for resilience to changes in the governing landscape. In Canada, efforts of CARL-Portage to develop shared repository infrastructure via Dataverse North and FRDR are examples of national support mechanisms to offset cost barriers and support more equitable access to institutions across Canada. Funding through partnerships with private sector organizations may provide significant opportunities but must take into consideration means of providing data access, particularly if doing so creates no real commercial disadvantage.[308]

Funding for long-term preservation and archival storage is particularly challenging and bears further consideration to derive sustainable funding mechanisms. From the outset of datasets being deposited in a repository, one cannot predict with certainty which datasets will have long term value, or which organizations will benefit most and so should share more in costs. Meanwhile, reliance on demand driven metrics of citation counts or downloads to determine value and impact of datasets contravenes the motivating open science principles. The experience of the archival community with these challenging questions should be considered in view of long-term support needed for RDM.

---

[303] https://www.computecanada.ca/research-portal/accessing-resources/resource-allocation-competitions/

[304] https://www.canarie.ca/software/

[305] https://www.innovation.ca/awards/major-science-initiatives-fund

[306] http://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/

[307] Erway, R. & Rinehart, A. (2016). If You Build It, Will They Fund? Making Research Data Management Sustainable. *OCLC Research.* https://www.oclc.org/content/dam/research/publications/2016/oclcresearch-making-research-data-management-sustainable-2016.pdf

[308] Knowledge Exchange Research Data Expert Group and Science Europe Working Group on Research Data. (2016). Funding research data management and related infrastructures. https://www.scienceeurope.org/media/uuqf0i03/se-ke_briefing_paper_funding_rdm.pdf

6 Key Challenges and Opportunities

Based on the above assessment of the current state of RDM in Canada, the following section presents a series of challenges in the landscape facing the Alliance as it assumes the national leadership role in supporting and advancing RDM in Canada. The issues identified are non-exhaustive and intentionally presented at a high-level to promote awareness and discussion, as well as highlight opportunities and current efforts to solve them. Many of the recommendations and suggestions that follow also emerged from facilitated conversations with the broader community, materialized through the Kanata Declaration and NDSF Summit reports.

This understanding of current challenges and opportunities will be augmented by the Alliance's upcoming researcher needs assessment process. As well, the later integration of the ARC and RS current state reports will support a more strategic assessment of how the Alliance will engage and prioritize challenges, both as Canada's DRI leader and in collaboration with partners across Canada and internationally.

Coordination

The importance of better coordination and communication emerges as a top priority in conversations with the RDM community, and it presents a challenge and an opportunity for the Alliance from the beginning. While this report presents a high-level overview of the range of actors, infrastructures, and services supporting RDM in Canada, it is an incomplete picture that requires refinement. One challenge is that many of the existing infrastructures, tools, platforms operate in relative isolation from one another. Better integration among new and existing services and infrastructures requires the adoption of shared standards, schemas, and certifications for trusted interoperability. This is particularly challenging across domains, where there can be vast differences in practices and tools available for managing data throughout its lifecycle. Existing initiatives referenced throughout this report can serve as models for supporting improved interoperability among ecosystem components.

In parallel, many actors within the Canadian RDM ecosystem exist in relative isolation. Continued consultation and outreach efforts are necessary to understand their needs. Ongoing engagement and discussion with the RDM community is also needed in order to build credibility and trust, support greater coordination and integration, and to communicate developments in the RDM ecosystem to researchers, allied organizations, and the general public. The current state could be improved through the development of bi-directional channels of communication among organizations, as well through advancing opportunities for mutually beneficial collaboration. Within the higher education sector, the work of the Portage Network serves as an example for the benefits and effectiveness of a community of practice model. Diverse communities described in this report should be brought together to build trust, understanding, and consensus. For instance, representatives from key organizations could be brought together to develop and maintain a high-level RDM Roadmap to help prioritize investment and development. As well, the establishment of an RDM Advisory Committee composed of key representatives could support the Alliance with guidance and oversight.

Canada must continue to engage internationally, working collaboratively to advance and adopt common practices and standards to further the development of the international RDM ecosystem. This engagement must also be effectively resourced to ensure Canada has a strong voice on the

international stage. As both the Canadian and broader international RDM ecosystems continue to mature, there may be opportunities to adapt tools developed elsewhere (as has been the case the DMP Assistant) or to develop consistent metrics for assessment and benchmarking should be supported for greater integration.

Representation and Inclusion

Among the general research population, awareness of existing RDM services and infrastructures available at local, regional, or national levels trends low, acknowledging of course that there is variation among domains and organizations. To promote greater awareness, a comprehensive index of services and infrastructures available at local, national, and international levels should be prioritized. Clear mechanisms to represent and refer to the range of ecosystem components are also needed. This could include support for expanded use of PIDs or the development of registries of ecosystem components, sustained through a globally coordinated effort. Of course, registries and indexes alone are not enough to serve awareness, so mechanisms of active promotion must be considered. National organizations are not always well placed to undertake this work, so a range of collaborators (e.g., academic societies, research organizations, and higher education institutions) with more direct access to researchers should be engaged in this process.

Beyond simply awareness, a key issue is promoting uptake of existing best practices, infrastructures, and services. Alignment with changes in the policy landscape of research institutions, funders, and publishers is one strategy to promote greater adoption. The Tri-Agency's draft RDM policy, in particular, is an important opportunity in Canada to support the intersection of institutional data governance models with available support through its requirement for institutional RDM strategies. The work of Portage to develop platforms that broadly support institutions and researchers to implement policy elements, such as data management plans and data deposit, is foundational, but should be expanded to also highlight intersections with domain specific and international infrastructures.

Any discussion about representation or promotion of ecosystem components must be led by a deep consideration of inclusion practices. Mechanisms are needed to ensure that both providers and users from all sectors and domains are represented and supported, with special consideration and accommodation to promote participation of under-represented voices, including those of women, racialized communities, and researchers from smaller and/or less technically-focus research domains. Forms of non-western research must also be included. As part of Canada's responsibility to advance reconciliation, First Nations, Métis, and Inuit communities must be full participants in this process.

As part of a broader effort, the Alliance has a role to play in supporting mechanisms for recognition and reward for researcher adoption of good RDM practices, which can promote wider cultural changes in attitudes towards RDM. Training and support for early career researchers is particularly important in facilitating this change.

Sustainability

The vision of a coordinated and scalable national RDM ecosystem where data can move with minimal restriction between component systems from its creation to long term preservation, and

which delivers responsive services and resources to meet the needs of researchers, requires a mosaic of partnerships and funding models to foster resiliency. Sustainable RDM requires suitable storage infrastructures across active, repository, and archival phases that are not only connected, but which are coordinated and anticipate future demand. It also requires that data be created and curated over its lifecycle with the FAIR and CARE principles in mind, requiring sufficient support be provided to researchers in terms of tools, training, and personnel, from both within and outside their home institutions, to support long term preservation and usability of their data as it moves throughout the ecosystem during its lifecycle.

The scale and growth at which research data are being generated, combined with the diversity of needs and interests, poses numerous challenges for sustainably supporting RDM at the national scale. With relative stability in the Alliance's funding envelope, there must be proportionate allocation for RDM infrastructure and services to support the broader aspirations of DRI. For many research communities, there is currently a lack of sustainable funding mechanisms that address their longer-term needs for RDM. Frameworks to sustainably fund core RDM infrastructure that are widely accessible and trusted by research communities (both supporting middleware, and repository and archival storage systems) are especially needed for decision making.

Before such funding models can be advanced, clarification regarding what is covered by ARC, DM, and RS envelopes within the Alliance is needed. In relation, fundamental distinctions between DRI ecosystem components must be clarified. For instance, both ARC and RDM have some overlapping infrastructure requirements while serving different objectives. While both repositories and archives support data beyond research project lifespans, related timeframes and mechanisms are not well established. Within this storage spectrum, the development of a shared understanding between researchers, service providers, policy makers, and other stakeholders is needed for consistency, with important impacts on policy and funding decisions. Some activities suggested by the Kanata Declaration include developing roadmaps for integrating ARC and DM components (e.g., common workflow language, harvesting or visitation protocols, etc.), and roadmaps for funding both mechanisms and providers of both repository and archival storage systems.

There is the risk with any new organization that rather than building on existing models and programs, they will choose to rather 'reinvent the wheel' at the expense of continuity and sustainability. As this report outlines, significant progress has been made in the RDM ecosystem, even in the short time since the publication of the last LCDRI papers. The Alliance should leverage the capacity existing in this ecosystem to optimize the benefits experienced by Canadian researchers. There is a need for further alignment and integration of organizations and services, not only Canadian entities supported by the Alliance, but also of their international counterparts. Determining how these partners in the RDM ecosystem fit together at all levels - local, regional, national, international - is an important step in furthering collaborative innovation, improving RDM support, and reducing overlap and duplication of efforts. An ideal system would consist of the provision of services at a range of levels, supported and structured through a national framework that is linked to and influenced by international standards and peer organizations.

7 Next Steps

Readiness to respond to current challenges and opportunities requires a thorough understanding of the landscape, in addition to clear targets for long term objectives. This report summarizes the current state of the RDM landscape in Canada to support a common understanding among the Alliance's members of the breadth and complexity of engagement in this field and serves as a basis from which the Alliance can set a path forward for national support for RDM in Canada. En route to developing the Alliance's strategic plan are several milestones that will contribute to defining and clarifying the Alliance's role in the ecosystem.

Researcher Needs Assessment (October - May 2021)

The Alliance will consult the research community to assess current services and identify priority-based needs for computing, data, and software. The newly formed Researcher Council will oversee the researcher needs assessment process that will engage a wide range of researcher communities and disciplinary associations and provide independent advice to the Alliance on matters related to the delivery of services and programs for the research community.

Following the needs assessment process, outcomes will be integrated with findings from the ARC, RS, and DM current state reports into a cohesive and directed DRI position paper that makes recommendations to support the Alliance's strategic planning effort.

Service Delivery Model (May 2021)

In collaboration with DRI partners, the Alliance will refine a new service delivery model that defines national, regional, and local services for DM, ARC, and RS, including expected service levels, new funding models, and roles and responsibilities.

Strategic Plan for 2021-24 (September 2021)

The Alliance will present a national strategy and vision for ARC, RS, and DM, integrating findings from its assessment and outreach activities. The strategic plan will include a roadmap for transforming the ecosystem from the pillarized current state, where ARC, RS, and DM are treated as separate entities, into a more integrated desired future state where research is supported by robust DRI across its lifecycle.

Appendix A. The Research Lifecycle and RDM Functions

This excerpt is reproduced with permission from Baker, D., Bourne-Tyson, D., Gerlitz, L., Haigh, S., Khair, S., Leggott, M., Moon, J., Tourangeau, R., and Whitehead, M. (2019, July 18). Research Data Management in Canada: A Backgrounder. Zenodo. http://doi.org/10.5281/zenodo.3341596

Figure A1. Data-Related Activities During the Research Process. (Created by the Leadership Council for Digital Research Infrastructure. In *Advanced Research Computing (ARC) Position Paper: For Innovation, Science, and Economic Development Canada*. Leadership Council for Digital Research Infrastructure. Unpublished manuscript. August 31, 2017, 5.)

The Research Lifecycle serves as a roadmap for researchers to understand what considerations they need to make for their data at every stage. Also, at each stage are five high-level points in which DM actions fall: Policies, Standards and Protocols, Policies and Procedures, Leadership, Advice, Support and Training, and Tools and Platforms.

Plan

The Plan phase of the lifecycle is the stage at which the researcher organizes themselves and their data for discovery, reuse and archiving further along in time. Ideally, they should acquaint themselves with DM guidelines and mandates relevant to their funding or postsecondary institution or other organization and identify appropriate standards and protocols that follow best practices for DM in their organization or domain. This includes the creation of a data management plan (DMP) and determining an appropriate repository for data storage and archiving.

This is the perfect time for the researcher to seek the assistance of RDM experts who can provide guidance in making these decisions. These experts can offer support in the form of DM training, clarify university (or other organization) processes that may intersect with national or domain mandates, and offer guidance on planning for the depositing, sharing and reuse of data.

Create

The Create stage involves the identification, acquisition and creation of research data and metadata. Unsurprisingly, it is at this stage that researchers must be aware of any institutional or domain-specific policies that define procedure for data collection. DM enters this phase in regard to best practices for data quality and integrity, versioning, and provenance; said practices at the local level often intersect with the international level. Best practices for metadata also must be observed at this stage to ensure interoperability and discovery, and occurs through the use of schemas and protocols. DM personnel and institutions can offer training and events in the realms of data quality and integrity and focus on domain-specific approaches. Data can be shared and transformed during its creation using research software platforms such as Virtual Research Environments, Science Gateways and e-Science platforms.

Process

In the Process stage, data is prepared for analysis (checking, validating cleaning, describing, and so on); part of this process involves ensuring domain-specific ontologies are followed. Code also must be managed at this stage so that it may be discoverable and reused, often through platforms like GitHub and Jupyter. The standardization of workflows is another consideration, as well as documenting every process. DM experts can assist in identifying tools that can assist with all of these tasks (for example, tools that can be used for the reuse of workflows like Taverna, Galaxy, and Kepler), and help with domain-specific policies and procedures in (meta)data wrangling.

Analyze

Analysis of data naturally follows preparation and processing. At this stage, code and workflow management and process documentation is still important, along with the creation and promotion of domain policies that facilitate analysis, outputs, data linking, reproducibility and privacy. DM experts, again, can offer guidance in these areas, as well as training on the use of data software and modeling. Specialized computing resources such as high-performance computing and cloud

services offered by Compute Canada, Amazon and Microsoft Azure may also be required by the researchers.

Disseminate

The Dissemination stage is the stage in which most of the sharing occurs. Before data can be transferred to a repository, deposit agreements, licensing, conditions for reuse (or access), and methods of discovery and preservation are considered. National policy framework is reflected in university policies (e.g. regarding ethics and privacy) and typically intersects with publisher policies and greater university strategies. Software code and codebooks, and other kinds of system details also need to be made available so that research may be reproduced.

Best practices to ensure sustainability, interoperability, discoverability and reuse must be followed. This includes the use of persistent identifiers for data, appropriate file formatting, and complying with international practices. Repositories and other data sharing platforms, such as Federated Research Data Repository (FRDR), Scholars Portal Dataverse and the Canadian Astronomy Data Centre are useful resources as they ensure metadata creation and quality assurance. DM experts may offer support to researchers by both creating and promoting best practices for sharing and reproducibility and offering consultation on (meta)data curation.

Preserve

Preservation is the second last stage in the Research Lifecycle and involves the process of moving data from an active to an archival state. In order to protect data, long-term university and national preservation policies (that often reflect or intersect with international guidelines) must be implemented. Data, metadata, documentation, coding and all back-up copies must be prepared for long-term access and reuse; in some cases, data needs to be migrated to more preservation-friendly formats. Other kinds of digital preservation processing include file characterization and normalization. Trusted Data Repositories are excellent services that have undergone certification to prove to the research community the repository's digital infrastructure is trustworthy and sustainable, offering a platform in which data can be stored and accessed long term. At this level DM experts may offer training on best practices for archiving and digital preservation, and review and implement data deposit agreements or mandates.

Reuse

Reuse of research is the final stage in the Research Lifecycle, and, regarding data, involves ensuring discoverability and access to data so that it may be combined to form new datasets, and referenced or analyzed by other researchers. At the highest level are national, international and domain policies and legislative frameworks focused on sharing and deposit. Following the FAIR Principles at this stage allow for ease of data reuse, as do tools that allow the reuse of documentation (Lab Books) and software or coding (GitHub). DM Experts can continue to support with data wrangling and understanding policies surrounding attribution, provenance and licensing, as well as searching and secondary analysis.

Store

The storage of data differs depending on the Research Lifecycle, and whether it is in an active state. Regardless of the state, the data, or at the very least the metadata, should be made to some extent accessible to other researchers. Therefore, storage also includes considerations into the deposit and retrieval of data (often facilitated by open standards such as SWIFT and ORE) into online storage platforms (and when appropriate, physical media). Archival use requires thought towards long term access and protecting the digital integrity of the content, as well as dissemination. University, national and domain policies regarding privacy and security and data-sharing must be taken into account when determining access.

In order for (meta)data to be accessible long term storage platforms should be open and sustainable, and there currently exists a number of options: OpenStack, FRDR Globus, Centre for Open Science Open Science Framework, and other domain services). DM experts can provide knowledge into domain-specific services, and in defining appropriate storage timelines as per local, national, international and domain data-governance practices. Experts should also consider the integration of desktop environments and processes like file synchronization to ease the task of data storage for researchers.

Discover

Researchers should strive to make their data discoverable at all stages of the Research Lifecycle, with ‘discover’ in this context relating not only to searching for data, but to the mobilization, location, interpretation, and assessment of it. This, in turn, allows fellow researchers to compile and create new (meta)data. There are a number of best practices that should be considered that will lead to high quality discoverability later in the Research Lifecycle: deciding upon appropriate metadata schemas and ontologies and understanding potential cross walks, considering relevant harvesting protocols for all types of metadata (OAI-ORE), and adopting PIDs (DOI, ORCID). Understanding the repository in which the data will be deposited is also an important step, as this will enable the researcher to be prepared to follow the repository’s standards (such as the SHARE data model), or take into account specialized discovery layers, such as registries that facilitate a federated approach to discover.

Following the FAIR Principles reinforces accessibility and discoverability to metadata from all stages of the Research Lifecycle, including that derived from research data and information. DM experts can provide guidance on FAIR as well as training in different kinds of discovery services and approaches. Experts should also consider the development of services based on broad use cases.

Document and Curate

Documentation and curation of (meta)data should be planned out early and occur throughout the Research Lifecycle for maximum interoperability and discoverability. This involves describing the context and workflow surrounding the data - coding and other materials, for example - and using appropriate metadata standards (found through resources such as FAIRsharing.org) to provide rich descriptions at appropriate levels. Of course, the data itself should also be described, identified and explained for preservation purposes.

FAIR and appropriate national data management policies, as well as journal and domain-specific policies should be understood by the researcher, but if they are unfamiliar they can seek out their institution's DM experts or external resources such as the CASRAI RDC RDM glossary to facilitate training. Format policy registries (e.g. PRONOM, RDA registries) also provide valuable standardization services.

Secure

Consent around sharing of data is another aspect of DM that needs to be considered; whether consent or anonymization of data is required, how much or how little of the data can be shared, and ensuring the legal and ethical conditions on the use of the data are followed and integrity and provenance are maintained. Researchers should be prepared to guard against unintended disclosure while also allowing appropriate access to data. Understanding ethics policies at all levels is necessary, and, in many cases, a researcher has domain-specific or international best practices (e.g. Health Insurance Portability and Accountability Act, Federal Information Security Act) that can offer direction. Other standards such as W3C security standards can be helpful in guiding the researcher to success in securing their data. Proven security platforms like RedCap and DataSHIELD, or use of secure facilities like Canadian Research Data Centres Network (CRDCN), are resources that should be highlighted at this stage. It is particularly important to understand security changes with the Research Lifecycle, and that the researcher adopt best practices and procedures that reflect these changes. Privacy offices, university IT security services, and communities in which privacy and access are of particular importance can provide advice and support in the security of research data.

Appendix B - Environmental Scan of National and Pan-National Digital Research Infrastructure Initiatives Supporting Research Data Management

WORKING PAPER

Introduction

A number of significant national or pan-national initiatives are in development to coordinate the open science landscape and provide foundational services and infrastructure to support researchers with the management of their research data. In Canada, the Digital Research Alliance of Canada (the Alliance) will play a critical role in advancing and coordinating DRI for Canadian research. In collaboration with partners and stakeholders across the country, this new organization will enable Canadian researchers with digital tools, services and infrastructure needed to support research excellence, innovation and advancement across disciplines.

The following report profiles a number of initiatives underway in other jurisdictions to develop common infrastructures and services to support and advance the state of research data management, for the purposes of informing and benchmarking against Canada's Alliance.

Review Criteria

Initiatives are reviewed using a combination of the following elements:[309]

Element	Description
Mandate	Mission and direction of the organization
Administration	Organization of governance structure
Implementation	Form, function, and/or structure of the organization • Discovery and access (Findable, Accessible) • Data services (Interoperable, Reusable + Quality, Preservation) • Skills and training

[309] Adapted from CODATA. (2019). Coordinating Global Open Science Commons Initiatives. https://conference.codata.org/CODATA_2019/sessions/155 (Retrieved May 2020)

Pan-National Initiatives

European Open Science Cloud [310]

Mandate

In Europe, a federated approach to advancing open science is taking shape. In 2016, the European Commission allocated €260 million for the federation of scientific data infrastructures through a new entity known as the European Open Science Cloud (EOSC). EOSC will foster a network of organisations and infrastructures from various countries and communities that supports the open creation and dissemination of knowledge and scientific data. The objective of EOSC is to give the EU a lead in research data management via "*a virtual environment with free at the point of use, open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines*".[311]

While EOSC's vision of interoperable data, services, and infrastructures will take time to realise, the initial steps include the formation of a "minimal viable platform" consisting of rules for participation to guide service provision and an action plan for data interoperability to operationalise the FAIR principles.

Administration

For the period 2019-2020, the EOSC governance model includes a Governance Board, an Executive Board, and a Stakeholders Forum.

- The Governance Board is composed of representatives from Member and Associate States, and chaired by representatives from the European Commission.
- The Executive Board is composed of 11 members chosen from a call for applications.
- The Stakeholders Forum will comprise users, EU-level and national projects, service providers, public sector, SMEs, Industry, etc.
- Five Working Groups coordinate progress on the priorities chosen by the Governance Board after proposal from the Executive Board. They are composed of representatives from EOSC stakeholders.
    - Landscape: Mapping of the existing research infrastructures which are candidates to be part of the EOSC federation;

[310] https://www.eosc-portal.eu

[311] European Commission. (2018). Implementation Roadmap for the European Open Science Cloud, SWD(2018)83. https://ec.europa.eu/transparency/documents-register/detail?ref=SWD(2018)83&lang=en

- FAIR: Implementing the FAIR data principles by defining the corresponding requirements for the development of EOSC services, in order to foster cross-disciplinary interoperability;
- Architecture: Defining the technical framework required to enable and sustain an evolving EOSC federation of systems;
- Rules of Participation: Designing the Rules of Participation that shall define the rights, obligations governing EOSC transactions between EOSC users, providers and operators;
- Sustainability: Providing a set of recommendations concerning the implementation of an operational, scalable and sustainable EOSC federation after 2020.

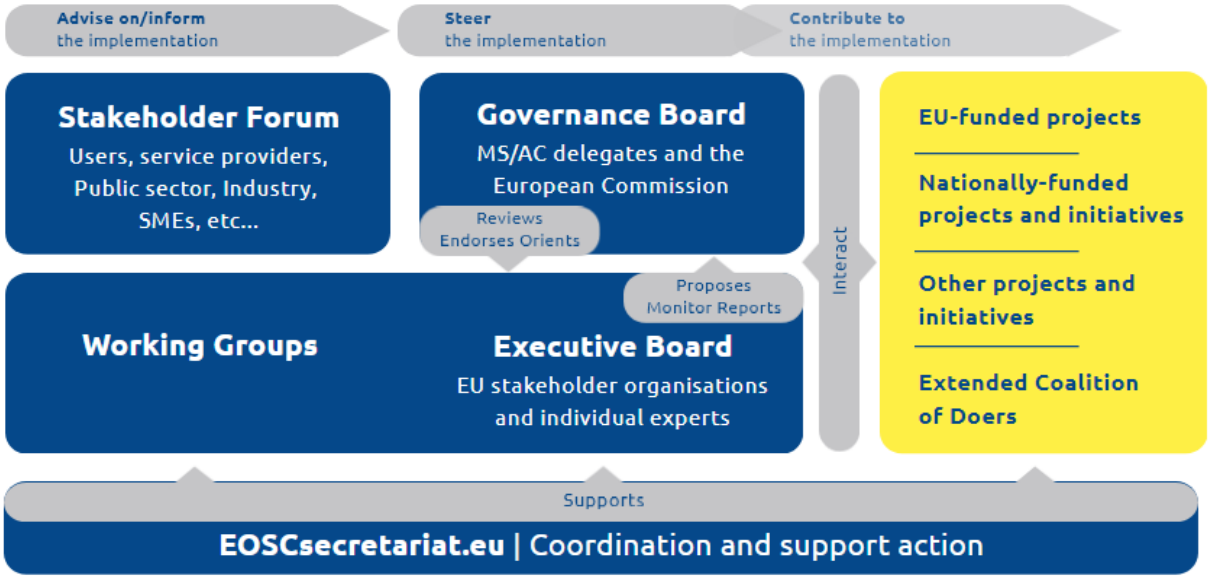


Figure B1. Governance model of EOSC.[312]

Implementation

An implementation roadmap describes six actions for the implementation of EOSC:[313]

- Architecture
    - EOSC would comprise a federating core and a variety of federated research data infrastructures committed to providing services as part of the EOSC. The EOSC

[312] Figure reproduced from https://www.eoscsecretariat.eu

[313] European Commission. (2018). Implementation Roadmap for the European Open Science Cloud, SWD(2018)83. https://ec.europa.eu/transparency/documents-register/detail?ref=SWD(2018)83&lang=en

federating core is to be constituted by EOSC shared resources and by a compliance framework, including the Rules of Participation.

- The process of federation entails two inter-related activities:
    - To develop shared resources as part of the federating core
    - To connect to the core a large number of research data infrastructures

- Data
    - Fostering the development of professional practices of research data management and stewardship in Europe by:
        - developing a better culture of research data management and practical skills among EU researchers;
        - developing FAIR data tools, specifications, catalogues and standards, and supply-side services to support researchers; and
        - encouraging consistent mandates and incentives for FAIR data from research funders and institutions across Europe.

- Services
    - EOSC plans to offer five main services to European researchers, regardless of disciplinary affiliation or national boundaries.
        1. A unique identification and authentication service and an access point and routing system towards the resources of the EOSC;
        2. A protected and personalised work environment/space;
        3. Access to relevant service information (e.g. list of federated data infrastructures, policy-related information) and to specific guidelines (guidelines for FAIR data, repository certification);
        4. Services to find, access, re-use and analyse research data generated by others, accessible through catalogues and data services (e.g. analytics, fusion, mining, processing); and
        5. Services to make their own data FAIR, to store them and ensure long-term preservation.

- Access & Interface
    - Multiple points of entry for accessing EOSC services are encouraged to support smooth transition from legacy systems, rather than forcing a single access point.

    - Entry points would consist of a front end that can be tailored to the specific needs of user communities, which would sit on top of a common platform providing access to shared EOSC resources.

- Rules
    - Shared rules between participating stakeholders will set out rights, obligations and accountability.
    - Rules ought to address use of EOSC shared resources (tools, specifications, standards, catalogues), regulation of transactions in the EOSC, and applicable legal frameworks.
    - Compliance with rules may vary with role, location, organizational maturity, and disciplinary context.
- Governance
    - An operational framework for overall governance and coordination with relevant national initiatives.
    - A funding framework to support sustainability.

The first projects developing outputs that will act as the initial base layer of the EOSC have been funded via the Horizon 2020 programme.[314] For example:

- EOSCPilot established the governance framework and initial rules of participation.[315]
- EOSC-Hub is supporting the federation of core eInfrastructures by creating an integration and management system that will act as a single point of contact for researchers to discover, access, and use DRI resources.[316]
- eInfraCentral and EOSC-Hub collaborated on the discover portal for EOSC resources.[317]
- HNSciCloud developed a hybrid cloud platform connecting commercial cloud service providers and publicly funded research organisations.[318]

[314] For a more complete list of funded projects contributing to EOSC: https://www.eosc-portal.eu/about/eosc-projects

[315] https://cordis.europa.eu/project/id/739563

[316] https://cordis.europa.eu/project/id/777536

[317] https://cordis.europa.eu/project/id/731049

[318] https://cordis.europa.eu/project/id/687614

- GÉANT is the network provider for the EOSC delivering appropriate access to cloud services, data, research infrastructures and the many other resources and services of the EOSC.[319]

- OpenAIRE is supporting a range of initiatives to advance the uptake of open science by the research community.[320]

EOSC's current services catalogue provides access to integrated resources, which are discoverable by scientific domain, service category or provider. See https://www.eosc-portal.eu/services-resources

African Open Science Platform [321]

Mandate

The African Open Science Platform (AOSP) is an initiative launched in 2016 by the South African Department of Science and Technology, with the objective of developing connections between open science activities underway across Africa via mechanisms for collaboration and coordination, and the exchange of best practices. The AOSP aims to support advanced open science research across Africa, and ensure alignment with existing programmes in regional and national research systems.

The vision for is AOSP is:

1. A federated system that provides researchers and other societal actors the means to find, deposit, manage, share, and reuse data, software, and metadata in pursuing their work

2. A network connecting dispersed actors, enabling adoption of digital tools, and developing capacity of individuals and institutions

Administration

The 2017-19 pilot phase for AOSP was supported by the South African Department of Science and Innovation, and managed by the National Research Foundation of South Africa (NRF) and Academy of Science of South Africa, with partners from the International Science Council and CODATA. In April 2020, it was announced that the AOSP Project Office would be hosted by the NRF.[322]

---

[319] https://ec.europa.eu/digital-single-market/en/geant-project-european-success-story

[320] https://cordis.europa.eu/project/id/731011

[321] http://africanopenscience.org.za

[322] https://www.nrf.ac.za/media-room/news/nrf-south-africa-host-aosp-project-office

Implementation

The next phase of the AOSP will include the development of a governance framework, sustainable long-term funding model, and formalisation of the AOSP Operating Model.

The initial management team will include: A Director, four Platform Officers (Data science, Data stewardship, Training and skills, and Network building, communications and outreach) and an Administrative Officer.

Members of the AOSP will include universities and representative bodies, science academies, granting councils, and services providers.

The pilot phase of the AOSP saw the launch of the platform through stakeholder workshops, meetings and presentations to create awareness of the need to curate scientific data in Africa in a trusted way. Deliverables included a landscape study to map data intensive research initiatives, as well as exploratory work towards frameworks and roadmaps for open science policy, infrastructure, and capacity building.

Enabling activities outlined in the AOSP strategy paper include:[323]

- Providing cloud computing facilities for networked computation, data access and analysis tools
- Providing software and experience-based advice on RDM, and open science policies and practice
- Creating and sustaining competitive research capacity in data analytics and AI
- Creating programmes of data-intensive research through the application of data technologies to major research domains

The Arab States Research and Education Network (ASREN) [324]

Mandate

ASREN aims to implement, manage and extend sustainable Pan-Arab e-Infrastructures dedicated for the research and education communities and to boost scientific research and cooperation in member countries through the provision of world-class e-Infrastructures and e-services.

Administration

ASREN is a nonprofit international organization, registered in Dusseldorf, Germany, on 3rd of June, 2011, under the umbrella of the League of Arab States. It is composed of the association

[323] The African Open Science Platform (n.d.). The future of science and science for the future. https://www.nrf.ac.za/sites/default/files/documents/AOSP%20Strategy%20Final%20HR.pdf

[324] http://asrenorg.net

of the Arab region National Research and Education Networks (NRENs), as well as their strategic partners.

Implementation

The initial focus of ASREN has been on creating science gateway communities that provide high speed access to scientific applications and compute resources. It has done this by:

- developing high-speed data-communications networks, including
    - a Point of Presence (PoP) in London providing EU termination and peering at its PoP to Arab NREN links, and enabling interconnection with the GÉANT network in Europe, Internet2 in the US and with other regional networks across the world.[325]
    - contributing to a regional building on the EUMEDCONNECT and AFRICACONNECT network infrastructure, cofunded by the European Commission.
- developing grid e-Infrastructure EUMEDGRID, and promoting the porting of new applications on the grid platform

ASREN provides high speed internet services and authentication services via eduroam and eduGAIN, respectively.

Collaborations

ASREN is currently working in collaboration with EGI.eu to coordinate and harmonize their e-Infrastructures by defining an operational and organizational model that is interoperable with e-infrastructures within EU countries and as bridges to other regions.[326]

ASREN is also working in conjunction with LIBSENSE and WACREN to create an open access journal and data repository serving north Africa and the Middle East.[327]

ASREN's MAGIC Project (Middleware for collaborative Applications and Global Virtual Communities) is establishing agreements for Europe, Latin America and other participating regions to create a marketplace of services and real-time applications for international research groups.

---

[325] http://asrenorg.net/?q=content/london-pop

[326] http://asrenorg.net/?q=content/chain-reds

[327] http://asrenorg.net/?q=content/libsense-iii-workshop-agenda

Nordic e-Infrastructure Collaboration (NeIC) [328]

Mandate

NeIC was established in 2012 with a vision modeling cross-border distributed and sustainable e-infrastructure collaborations. NeIC collaboratively explores, evaluates, develops and deploys innovative infrastructure services in response to the needs of the national e-infrastructure providers, their users, and projects of joint Nordic interest.

NeIC projects are in place in the areas of Physics and Engineering Sciences, Environmental Sciences, Humanities, Culture and Society, Life Science and e-Sciences.

Administration

NeIC is hosted by NordForsk, the research and research-infrastructure funding arm of the Nordic Council, which is the official body for inter-parliamentary cooperation among Nordic countries.

The NeIC Board consists of one representative of each member's national e-infrastructure provider. These include CSC (Finland), SNIC (Sweden), UNINETT Sigma2 (Norway), DeIC (Denmark), RH Net (Iceland) and ETAIS (Estonia).

The NeIC Board has the authority to make strategic decisions regarding computing and data-storage infrastructure. The NeIC Board recommends the NeIC Director to be appointed by NordForsk.

NeIC is managed by an Executive Team chaired by the NeIC Director. The Executive Team coordinates activities and participate in project steering groups as project owners.

In collaboration with national eInfrastructure providers and user-community representatives, NeIC engages experts to participate in both projects and operational activities.

Funding of NeIC's activities is provided through national funding agencies, NordForsk and participating project partners.

Implementation

NeIC provides researchers in member countries access to a common support center, common tools for data sharing and analysis, sharing mechanisms between high-performance computing resources, and workshops for researchers in data analysis, software development, and research data management.

NeIC funds collaborative e-infrastructure projects with Nordic partners.

NeIC also provides a range of related services, which are registered with a DataCite PID (10.23673/kpyv-1k13).[329]

---

[328] https://neic.no

[329] Nordic e-Infrastructure Collaboration (2020) "Nordic e-Infrastructure Services." EOSC-Nordic. https://search.datacite.org/works/10.23673/kpyv-1k13

Services related to data management include an online platform for creating data management plans, repository platforms for data storage, sharing and management, a notebook environment for working with data and programming, a platform for collecting, storing, analyzing, and sharing sensitive data, and workshops on FAIR data and research software.

NeIC also builds professional networks to join experts who work on similar challenges in different Nordic organizations. NeIC does this through workshops and building virtual spaces for distributed teamwork.

National Initiatives

Netherlands

Data Archiving and Networked Services (DANS) [330]

Mandate

To promote sustainable access to digital research data, DANS provides expert advice and services to support researchers in making their digital research data “FAIR”. This includes services for long-term archiving and reuse of research data from completed research, and support for data management during active projects.

Administration

DANS is a joint initiative of the Royal Netherlands Academy of Arts and Sciences, and the Netherlands Organization for Scientific Research. A Steering Committee is responsible for supervising daily business, management and policies. Three advisory boards guide platforms and services: Scientific Advisory Committee, NARCIS Advisory Board, and the DataverseNL Advisory Board. A User Panel consisting of researchers and employees of Dutch universities provide feedback and input on services.

A central secretariat is responsible for developing infrastructure, services and policies, as well as operations.

[330] https://dans.knaw.nl/

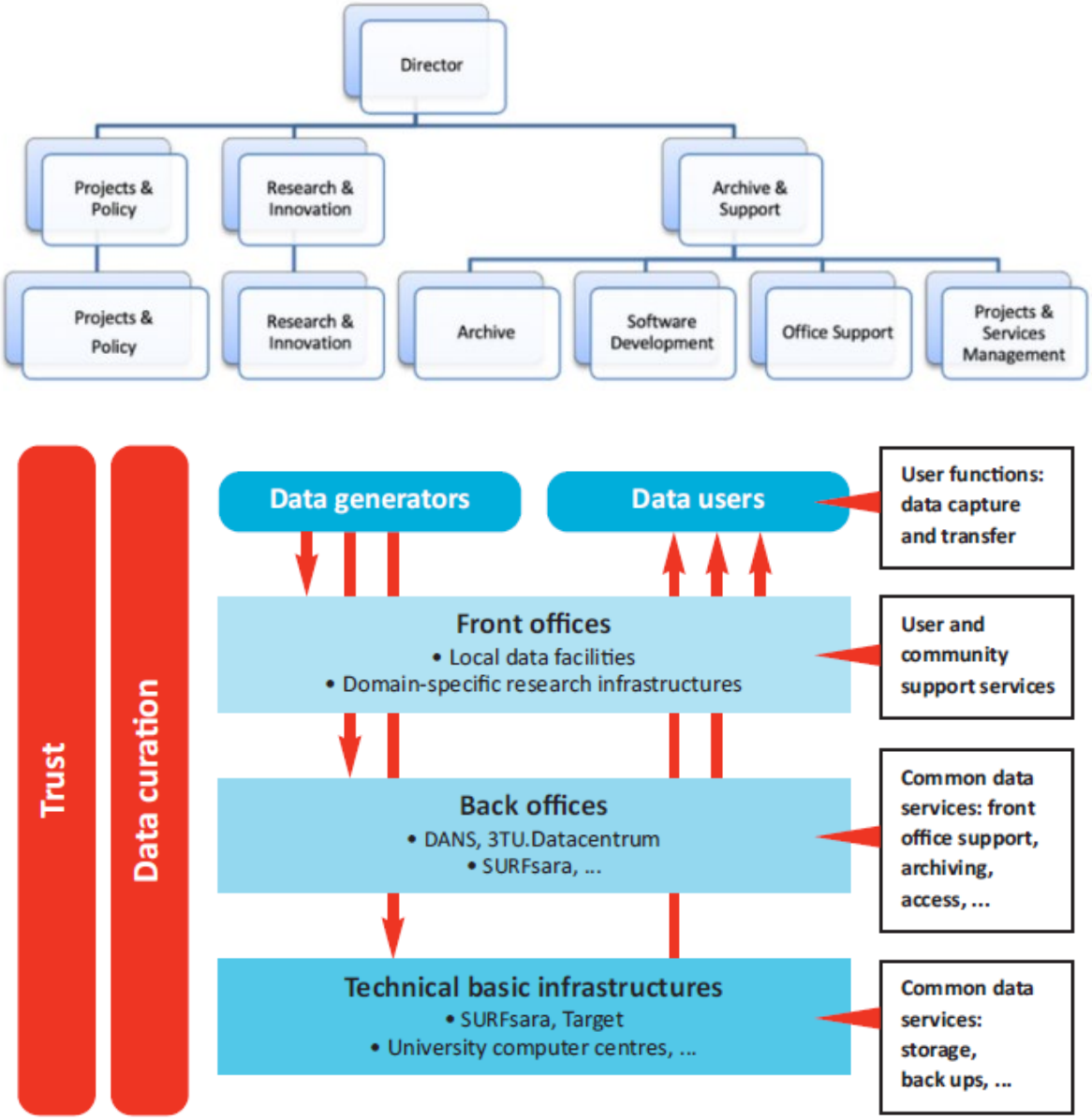


Figure B2 & B3. Organizational structure and workflow of DANS.[331]

[331] Figures reproduced from https://dans.knaw.nl/en/about/organisation-and-policy/organigram; and https://dans.knaw.nl/en/about/organisation-and-policy/information-material/DANSstrategienota20152020UK.pdf

Implementation

DANS is building federated infrastructure and services to support their research community. DANS offers the following services:[332]

- DataverseNL: research data can be stored, shared and published via DataverseNL during research. DANS manages the repository network, while participating institutes are responsible for managing deposited data in their local nodes. All institutions participate in an Advisory Board, which determines policies for the service. Institutions pay a fixed amount for participation, with additional storage charges.
- EASY: after research, research data can be permanently stored and shared via EASY, their online archiving system. EASY also offers access to the secure micro data of Statistics Netherlands. EASY is certified by the CoreTrustSeal and Nestor Seal.
- NARCIS: information about research projects, open publications, and research software can be shared via the NARCIS science portal.
- Training & consultancy: DANS offers training and consultancy in the field of digital sustainability, software sustainability, data management, FAIR data, and Research Data Management.

Within the Netherlands, DANS collaborates with DRI stakeholders and a range of domain repositories. They are also involved in the international RDM community through a range of networks and infrastructure projects.[333]

In collaboration with Netherlands eScience Centre, DANS has launched “fair-software.nl” to support research software sustainability training.[334]

Germany

National Research Data Infrastructure (NFDI) [335]

Mandate

The aim of the National Research Data Infrastructure (NFDI) is to manage scientific and research data, provide long-term data storage, backup and accessibility, and network the data both nationally and internationally. The NFDI will bring multiple stakeholders together via a coordinated network of consortia tasked with providing science-driven data services to research communities.

[332] https://dans.knaw.nl/en/about/services

[333] https://dans.knaw.nl/en/projects; https://dans.knaw.nl/en/about/organisation-and-policy/collaboration

[334] https://fair-software.nl

[335] https://www.dfg.de/en/research_funding/programmes/nfdi/index.html

Administration

- Consortia are generally organised by research domain or method. Their aim is to improve and safeguard access to and use of research data in their relevant area.
- Spokespersons of each consortia make up the Consortia Assembly. The Consortia Assembly makes research, operational and service-related decisions on behalf of the NFDI consortia and the NFDI and decides on the introduction of cross-disciplinary procedures, services and/or standards relating to the NFDI within the scope of the principles approved by the NFDI Scientific Senate.
- The Scientific Senate is the strategic body of the NFDI. It advises on matters relating to ongoing development of national research data infrastructure to ensure that the NFDI connects with national and international infrastructures.
- A local directorate consists of a full-time director and managing office. The Directorate ensures that cross-disciplinary topics, such as education, data protection and data ethics, are discussed, communicated and coordinated among the NFDI consortia.

Implementation

Over a period of three years, starting in 2019, the NFDI will be established as a cooperative network of consortia in three consecutive selection rounds. In each round, new consortia can be added to the NFDI in a research-driven process. The aim is to create a comprehensive framework of interconnected consortia providing a national research data service to the research community. A total of 22 proposals for NFDI consortia were received in response to the call for proposals on 15 October 2019. A total of 142 different institutions were involved in the proposals. Further calls are planned for 2020 and 2021.

The NFDI's programme aims for consortia include:

- Establishment of data handling standards, procedures and guidelines in close collaboration with the community of interest
- Development of cross-disciplinary metadata standards
- Development of interoperable data management measures and services tailored to community of interest
- Increased reusability of existing data, also beyond subject boundaries
- Improved networking and collaboration with partners outside the German academic research system with expertise in research data management
- Involvement in developing and establishing generic, cross-consortia services and standards in research data management together with other consortia

China

China Science and Technology Cloud (CSTCloud) [336]

Mandate

The China CSTCloud Federation provides the Chinese education, research, scientific and technical communities, relevant government departments and hi-tech enterprises a range of cyber infrastructure and internet services such as network access and identity management, along with computing power, cloud storage and research software.[337]

Administration

The CSTCloud is managed by The Computer Network Information Center (CNIC) which is a part of the Chinese Academy of Science (CAS).[338] The CNIC mission is to build ICT to support scientific innovations and management activities, promote the R&D of information technology, integrate e-science and e-management resources, and strengthen the spread of scientific ideas.[339]

Within the CNIC a Science and Technology Committee advises the CNIC on priorities for the CSTCloud,[340] and the Strategic Advisory Committee provides advice on development strategies.[341]

Implementation

The CSTCloud initiative was announced during the Fourth World Internet Conference in 2017, followed by the initial platform launch in 2018. The most recent version was launched in late 2019.[342] Conceptually the CSTCloud components covering network, compute and data services are delivered in 4 layers of service:

- Infrastructure services including internet access (CSTNET), cloud services, VPN, HPC (CNGrid) and AI cloud services.
- Operational services for data archiving, long term data preservation and disaster recovery
- Collaboration services for video, web and data conferencing and mail and mobile services

---

336 https://www.cstcloud.net/

337 http://english.cnic.cas.cn/patform/202001/t20200106_228938.html

338 http://cstcloud.net/index.html

339 https://www.egi.eu/about/newsletters/egi-and-the-chinese-academy-of-sciences-collaborate-to-boost-science-beyond-national-boundaries/

340 http://english.cnic.cas.cn/about/stc/

341 http://english.cnic.cas.cn/about/sac/

342 http://english.cnic.cas.cn/patform/202001/t20200106_228938.html

- A suite of discipline specific platforms for materials science, high energy physics, computational chemistry and microbial research along with dedicated platforms for data visualization, research software development, and AI modelling

CSTCloud also provides services to other CAS initiatives including CASEarth, CAS Space Science Missions, the Five-hundred-meter Aperture Spherical Telescope,[343] and the Large High-Altitude Air Shower Observatory.[344]

There is a current effort to develop 20 national data centres, covering all types of research data. These 20 national data centres are then planned to feed into CSTCloud.[345]

Japan

National Institute for Informatics (NII) Research Data Cloud [346]

Mandate

The 5th Science and Technology Basic Plan (2017-2021) sets out the basic policy on science and technology initiatives and promotes a foundation of Open Science in Japan.[347] In support of the Plan's commitment to open science, the NII created the Research Data Cloud as an e-infrastructure where research data and other related files can be managed, stored, and discovered. "*Our infrastructure will operate at all times as if giraffes act without rest, and will manage various kinds of data which differ like coat pattern of giraffes. Users will be able to access necessary one from a large amount data as if giraffes find the necessary food from high perspective in savanna.*"[348]

Administration

The Research Data Cloud is built and maintained by the Research Center for Open Science and Data Platform (RCOS). One component, the JAIRO Cloud, is a joint initiative of NII and the Japan Consortium for Open Access Repository (JPCOAR).[349]

[343] https://www.nature.com/articles/d41586-019-02790-3

[344] http://english.ihep.cas.cn/lhaaso/

[345] https://researchdata.springernature.com/posts/54209-is-china-ready-for-open-data

[346] https://rcos.nii.ac.jp/en/service/

[347] https://rcos.nii.ac.jp/en/about/

[348] https://rcos.nii.ac.jp/en/about/mascot/

[349] https://rcos.nii.ac.jp/en/service/weko3/

Implementation

The NII Research Data Cloud became operational in 2020 and utilizes the Science Information Network (SINET5) to provide authentication, cloud infrastructure and academic content to promote open science. It consists of three platforms:

- a research data management platform (GakuNin RDM[350]), for active research data management while projects are being executed; built on the Open Science Framework.[351]
- a repository platform (WEKO3), to store data, derived publications and their relationships to build the scholarly communication network graph. Built on JAIRO Cloud, which is a nationally provided institutional repository cloud service. It includes a DOI service.
- a discovery platform (CiNii Research) that harvests metadata from institutional repositories and other open databases.[352]

These platforms underpin NII's Research Center for Medical Big Data, a platform for R&D initiatives including cloud-based AI technology for analyzing medical images.

Access to services from institutions across Japan is supported via Shibboleth and managed by the Academic Access Management Federation, also referred to as GakuNin. GakuNin is also responsible for pursuing inter-federation access globally.[353]

Korea

KISTI/Korean Research Data Platform [354]

Mandate

The Korean Institute of Science and Technology Information (KISTI) is a government funded research institute that supports competitiveness of Korean R&D through research and intelligence gathering, developing national standards and services, and providing advanced DRI infrastructure, including high speed research network, supercomputing, and national repository network.

Korea has been building its Korea Research Data Platform (KRDP) since 2018 with the intention that it should systematically support researchers to share, manage, search, analyse and use research data.

---

350 https://rdm.nii.ac.jp/

351 https://www.cos.io/our-products/osf

352 https://rcos.nii.ac.jp/en/service/research/

353 https://www.gakunin.jp/en

354 https://www.kisti.re.kr/eng/rnd/pageView/250

Administration

Currently, development of KRDP is under the direction of KISTI's Research Data Sharing Center, Division of National S&T Data.[355] Availability of more detailed information online is limited.

Implementation

Once completed, the KRDP will fulfill the following functions:[356]

- Integrated RDM environment to preserve share and reuse research data
- Federated search tool for distributed research data
- Real-time collaborative analysis environment
- Data management planning tool integration

KRDP will also integrate into KISTI's other open science initiatives, for instance open access policies and Korea Open Access Repository & Archive (KoaRXiv).[357]

Australia

Australian Research Data Commons [358]

Mandate

The ARDC was formed in 2018, building on legacy initiatives (the Australian National Data Service (ANDS), National eResearch Collaboration Tools (Nectar) and the Research Data Services (RDS)), to bring together the people, data, skills and resources to enable world class data intensive research. The overarching goal of the ARDC is to accelerate Australian research by developing, testing, and supporting platforms where investigators can store, discover, share, access, and interact with digital objects (data, software, etc.). It has a mandate to provide national coherence to data and e-research platform capability, including:

- High performance computing - National Computational Infrastructure (NCI) and Pawsey Supercomputing Centre
- Research networks - Australian Academic Research Network (AARNet)
- Access and authentication - Australian Access Federation (AAF)

ARDC has identified five strategic themes to frame the implementation of its vision:

---

355 https://www.kisti.re.kr/eng/rnd/pageView/250

356 https://github.com/pragmagrid/pragma-meetings/blob/master/pragma36/24/talk10-krdp.pdf

357 https://zenodo.org/record/3232912

358 https://ardc.edu.au/

- Theme 1 - Coordination and Coherence: Facilitating an Australian research data commons
- Theme 2 - People and Policy: Transforming culture and community
- Theme 3 - Data and Services: Maximising the value of Australia's data assets
- Theme 4 - Software and Platforms: Enabling research insights
- Theme 5 - Storage and Compute: Providing foundation infrastructure

Administration

The ARDC has an operating budget for the period of covered by its 2019-2023 strategic plan of $110m and a capital investment budget for the same period of $72m.

An Executive Team lead core activities, consulting with the community to set the direction of work and supporting co-creation activities.

Implementation

ARDC recognizes that partnerships with stakeholders at various levels in each of its service areas are central to supporting the notion of a data commons. A report by Sarah Jones maps out stakeholders in the commons (Figure 4).

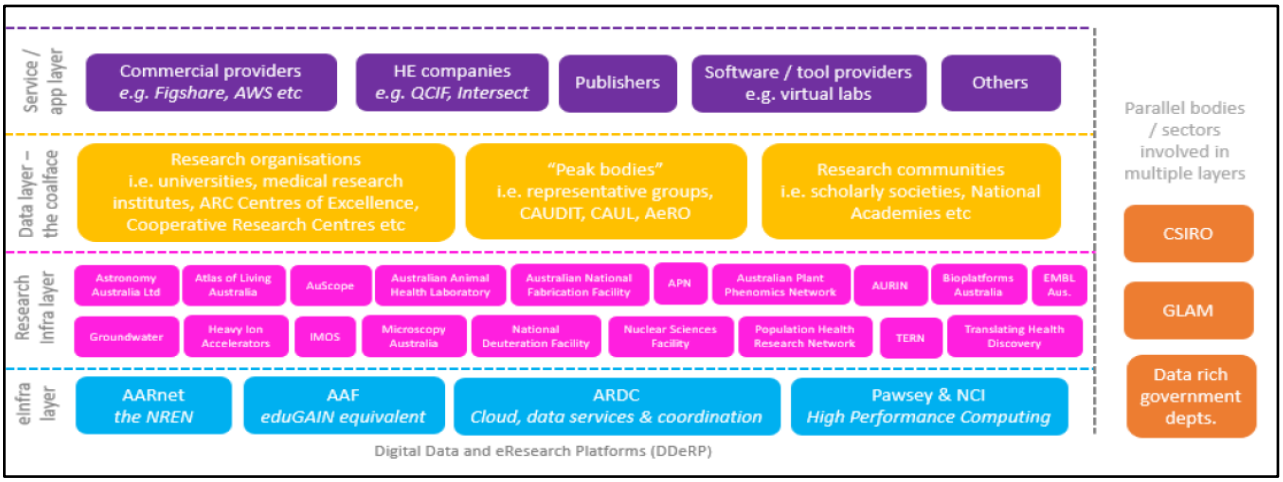


Figure B3. Mapping stakeholders in the ARDC (Sarah Jones, DCC)[359]

[359] Figure reproduced from https://op.europa.eu/en/publication-detail/-/publication/be6c8944-216c-11ea-95ab-01aa75ed71a1/language-en

ARDC has identified 3 key areas for its 4-year strategic plan:[360]

Software and Platforms

- Enable research through use of advance research software and platforms
- Support the creation and maintenance of an ecosystem of FAIR research software and services
- Transform practice by encouraging new funder/publisher/institutional policies and creation of software culture that addresses issues of governance, citation, stewardship, and attribution
- The ARDC is running a program of investment in platforms. An explicit goal of the program will be to increase the number of researchers with access to platforms, both in terms of absolute number and in terms of diversity of disciplines. The program will also seek to support a community of platform operators.

Storage and Compute

- Derive maximum benefit from existing resources
- Ensure evolution of compute resources meets stakeholder needs
- Develop reliable measures of impact
- Develop sustainable funding models

A key initiative is refreshing ARDC's Nectar Research Cloud compute and storage infrastructure, which has reached end of life, to achieve capacity required to meet future demand to be able to:

- host research applications in a scalable and flexible cloud environment
- access a computational resource which complements existing and new supercomputing facilities
- rapidly deploy and share innovative research applications

Data and Services

Two strategic objectives will be pursued in the Data and Services theme:

1. Enabling new research from existing data
2. Increasing the integrity and reproducibility of research across the whole research system by increasing the FAIR-ness of the data arising from research.

For the first objective, ARDC will partner with research communities, facilities, government agencies, and research organisations to target data collections with high re-use value, strong

[360] Australian Research Data Commons. (2019). ARDC Strategic Plan 2019-2023. https://ardc.edu.au/wp-content/uploads/2019/05/ARDC-Strategic-Plan-2019-2023.pdf

community ownership, and a national scope. ARDC's role will be to provide seeding resources to develop these national and community collections and their long-term sustainability.

For the second objective, ARDC will promote institutional data curation and management capacity as well as greater coherence of data collections and services across sectors of the national research data ecosystem. ARDC's role will be in supporting, facilitating, and communicating best data practice, interoperability, capacity, and capability across the sector. This will include a commitment to ensuring that Australia is aligned with international standards and initiatives and has more data that is FAIR.

Areas of initial focus will include:

- transformational data collections
- sensitive data and approaches, platforms and services to manage, collaborate over, and share these data
- institutional roles and approaches in the data commons

Services

- Nectar Research Cloud
    - Federated Research Cloud
- Research Data Australia
    - Find research data from Australian research organizations, government agencies, and cultural institutions
- Identifier Services
    - Create and manage PIDs for data
- Research Vocabularies Australia
    - Find and use controlled vocabularies

Collaborations

National Data Assets Program

- ARDC has launched a National Data Assets initiative to establish a portfolio of national scale data assets. Six initial programs have been proposed, each of which will have their own open call for participation with unique EOI and RFP requirements.
    - Cross-NCRIS National Data Assets program
        - Partner with clusters of National Collaborative Research Infrastructure Strategy facilities to establish interconnected collections
    - National Data Partnerships

        - Partner with data stewards, institutions, research communities, industry and public sector to establish national data assets

    - Public Sector to Research Sector Bridges
        - Partner with public sector data stewards and users to improve use and access to public sector data for research purposes beyond core business
    - Emerging Collections
        - Partner with research communities and organizations to incubate development of emerging national scale collections
    - Institutional Underpinnings
        - Partner with institutions to catalyse adoption of FAIR practices
    - Health Studies Australian National Data Asset
        - Partner with health research institution consortia to build distributed national data asset from output of national health funded research

Skilled Workforce

The ARDC will establish partnerships to facilitate collaboration and coordination on skills development, with a focus on:

- advancement of cultural change through policy and funding frameworks
- skilled workforce planning for the sector
- development of key communities of practice, including connections with international communities and initiatives.

Piecemeal National Initiatives

United Kingdom

National research data services in the UK are supported by a patchwork of national research funders and higher education sector organizations, where much of the focus has been placed on capacity development and on repository data storage.

UK Research and Innovation [361]

Administration: UKRI is an independent agency supported through the UK Department for Business, Energy, and Industrial Strategy. UKRI brings together the UK's seven Research Councils and provides an overarching framework for individual Research Council policies on data policy.

Implementation: Several of the Research Councils currently operate data repositories and are engaged in supporting platform and service development to support data resulting from funded research or third-parties. The Natural Environment Research Council operates a number of data centres and is developing a data commons approach, which will provide both storage and computing power to enable researchers to bring data and computation to its archive of datasets. [362] The Economic and Social Research Council supports the UK Data Service,[363] which works with a number of organisations including Government agencies and departments to provide users with access to a wide range of data resources. It provides detailed guidance for researchers, and engages on projects advancing issues in metadata, interoperability, and web technologies.

Jisc [364]

Mandate: Jisc provides network and shared digital infrastructure to UK higher education institutions.

Implementation: Jisc is supporting open science through several initiatives, including an Open research hub. Built in partnership with the UK research sector, the open research hub will allow researchers to manage digital outputs in one place, offer secure storage and management, and archive and preserve research data.

Digital Curation Centre [365]

Mandate: Launched in 2004, the DCC resulted from recommendations made in the JISC Continuing Access and Digital Preservation strategy, as a national organization to solve challenges in digital curation of research data.

---

[361] https://www.ukri.org/

[362] https://nerc.ukri.org/

[363] https://ukdataservice.ac.uk/

[364] https://www.jisc.ac.uk/

[365] https://www.dcc.ac.uk/

Administration: Principal Partners include the University of Edinburgh, HATII, and UKOLN.

Implementation: The DCC provides expert advice and practical help to research organisations wanting to store, manage, protect and share digital research data. They also provide support for issues such as policy development and data management planning, and run a number of training programmes to develop skills of researchers and data managers to support FAIR principles.

United States

In the U.S., there is no national analogue to the Alliance for supporting digital research infrastructure or providing data management services. However, a number of organizations aim to support the coordination of research data management at a national-level and for various communities.

Infrastructure/Institutional Coordination

NIST RDaF [366]

Mandate: The U.S. National Institute of Standards and Technology's mission is to promote U.S innovation and competitiveness by advancing measurement science, standards and technology. The development and use of standards is a core competency, and they have developed a number of domain-specific and broader-field standards in areas of networks and scientific data systems, cybersecurity and privacy, and measurement science.

Implementation: NIST recently convened a multi-stakeholder working group to explore the creation of a NIST-led Research Data Management Framework (RDaF), recognizing that more national coordination is needed to support planning, deploying, and operating research infrastructure. The goal of the RDaF is to provide organizations a structured approach to develop a coherent data management strategy, by providing a common language and basis for coordination. The RDaF will touch on all aspects of data management practices in all phases of the data life cycle. The main target for the RDaF will be at an institutional or organizational level (e.g. CIO, CDO) - someone with broad responsibilities for the management of research data across an organization. The long-term plan for developing the RDaF includes a scoping study, pilot studies in specific disciplines such as astronomy, materials science, agriculture, or economics or on stakeholder communities such as the university library or research laboratory community.

Domain-Centred Support and Coordination

NIH New Models of Data Stewardship[367] / Office of Data Science Strategy [368]

Mandate: The NIH is the primary U.S. funding agency for supporting biomedical and public health research. The New Models of Data Stewardship (NMDS) program was supported by the NIH's Common Fund that supports strategic investments in emerging areas that no single NIH institution

[366] https://www.nist.gov/news-events/events/2019/12/research-data-framework-rdaf-workshop

[367] https://commonfund.nih.gov/data

[368] https://datascience.nih.gov/

can address on its own. The NMDS program ended in 2018 and initiatives transitioned to the newly formed Office of Data Science Strategy.

Implementation: From FY 2017-2018 the NMDS supported two initiatives to develop and test new strategies for data management:

1. The NIH Data Commons Pilot Phase explored new ways to store, access, and share biomedical data and associated tools in the cloud so they were FAIR. This included guidelines and metrics, unique identifiers, shared workspaces to find and interact with data, indexing and search functionality, support for research ethics, privacy and security, and teaching and outreach.

2. The Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative established partnerships with commercial cloud service providers to reduce economic and technological barriers to accessing and computing on large biomedical datasets to accelerate biomedical advances.

The NIH Office of Data Science Strategy currently supports elements of the NMDS programs, including the STRIDES initiative, which has collaborated with Google Cloud and Amazon Web Services to support projects who want to prepare, migrate, upload and compute on data in the cloud. The NIH also supports a Researcher Authentication Service, which expects to deploy in late 2020 a set of APIs that will allow seamless login and access across integrated domain data repositories. While the NIH supports a number of domain-specific data repositories, they are also running a pilot generalist repository on figshare and recently released two funding opportunities to support biomedical data repositories and knowledge bases. Finally, a number of training and outreach initiatives to encourage adoption of FAIR tools (e.g. Fast Healthcare Interoperability Resources) and practices are run through the Office.

Services Support and Coordination

Several organizations have spun up in the last few years to support development and coordination of data services provided by academic research institutions, typically via their university libraries. Two of note are the Data Curation Network and the CURE Consortium.

Data Curation Network [369]

Mission: The Data Curation Network (DCN) serves as the “human layer” in the data repository stack and connects local data sets to expert data curators via a cross-institutional shared staffing model. Their vision is to

- provide expert data curation services for Network partners and end users,
- create and openly share data curation procedures and best practices,
- support training and development opportunities for an emerging data curator professional community

[369] https://datacurationnetwork.org/

Administration: The Data Curation Network is supported by grants from the Alfred P. Sloan Foundation and the Institute of Museum and Library Services. The Data Curation Network project team includes representatives from 10 partner institutions. DCN representatives are typically managers and directors of their local curation services and often have supervisory responsibilities for the DCN Curators who contribute staff time to the project.

Implementation: Curation experts located within partner institutions share their time and expertise in curating datasets with other member institutions. This enables all member institution's data repositories to collectively, and more effectively, curate a wider variety of data types (e.g., discipline, file format, etc.) that expands beyond what any single institution might offer alone.

CURE Consortium[370]

Mission: To support curation of research data and review of code and associated digital scholarly objects for the purpose of facilitating digital preservation, reuse, and reproducibility of published science.

Implementation: The CURE Consortium is developing a network of expert curators to establish standards, sharing practices, and promoting adoption of their Data Quality Review procedure. Membership in CURE is open to academic institutions, organizations, and individuals who support and promote the principles of Curating for Reproducibility and are committed to, currently have, or are in the process of implementing workflows for transparent and reproducible research.

Appendix C - International/National Research Data Management Associations

This document provides an overview of organizations within and outside Canada that are currently engaged in supporting and advancing research data management through the development of communities of practice, and which could be strategically engaged in the Alliance's data management stream to support the Canadian data management community. This document records how these organizations are advancing the state of research data management, and could inform awareness and coordination of existing Portage and RDC working groups, as well as future Alliance working groups.

Research Data Alliance (RDA)

https://www.rd-Alliance.org/

[370] http://cure.web.unc.edu/

RDA was launched in 2013 by the European Commission, US National Science Foundation, National Institute for Standards and Technology, and Australian Department of Innovation. As of March 2020 it has almost 10,000 members based in 144 countries. It is a platform where international research data experts meet to exchange views and advance topics related to best practices, standards, and protocols. Outputs include RDA Recommendations, which are documents that may include specifications, taxonomies or ontologies, workflows, schemas, data models, etc., that are endorsed by RDA.

- Working groups typically have a short-term mandate (18 month lifespan). There are 36 as of May 2019.
- Interest groups have a long-term mandate. There are 66 as of May 2019.
- Groups tend to take two forms:
    - Domain-based groups (e.g. Chemistry, Health, Indigenous, Agriculture, Biodiversity, etc.)
    - Functional groups (e.g. Federated ID Management, Domain Repositories, Active DMP, Data Policy, Sharing Rewards and Credit, Vocabulary services, Virtual Research Environments)

RDA North America

https://www.rd-Alliance.org/groups/rda-north-america

The objective of RDA North America is to build relationships between RDA members and other potential collaborators on the North American continent.

World Data Systems International Programme Office (WDS-IPO)

https://www.icsu-wds.org/organization/international-programme-office

WDS is an interdisciplinary body of the International Science Council, with the mission of supporting access and stewardship of trusted scientific data and data services, products, and information. For instance, it supports the CoreTrustSeal program.

WDS-IPO coordinates the operations of the WDS and is under the direction of the WDS Executive Director with guidance from the Scientific Committee. The IPO is hosted and supported by the Japanese National Institute of Information and Communications Technology.

World Data Systems International Technology Office (WDS-ITO)

https://wds-ito.org/what-we-do

WDS-ITO is based at the University of Victoria and is supported by three Canadian host organizations: Ocean Networks Canada, Polar Data Catalogue, and Canadian Astronomy Data

Centre. The ITO is under the oversight of the WDS Scientific Committee. The ITO supports member organizations and partners via technical infrastructure and services to support access to scientific data. The ITO supports infrastructure necessary for repositories with data analytics and visualization.

- ITO Technical Advisory Committee advises on infrastructure strategies and technology roadmaps, trends and activities in the global DM community, provides counsel and advocates for ITO programs.

International Science Council - Committee on Data (CODATA)

https://codata.org/

Committee on Data of the International Science Council, whose goal is to promote global collaboration to improve availability and usability of data for research and policy, technological and cultural changes. Convenes a range of standing committees, task groups and working groups. It is more of a top-down organization compared to RDA, with connections to other influential international organizations.

- Working Groups address immediate short term needs. Current groups are related to vocabularies (IRIDIUM Glossary), training, repository business models, FAIR data, and some domain-specific topics (e.g. nanomaterials, materials).

- Task Groups are selected biennially and cover a range of topics addressing data needs or policy issues. E.g. Digital Representation of Units of Measure, Improving Data Access and Reusability, Citizen Science, Linked Open Data for Global Disaster Research, Preservation of data in Developing Countries.

CNC/CODATA

https://codata.org/canada/ (and old site https://www.codata.info/canada/about.shtml)

The Canadian member organization of the CODATA parent organization. Previously concerned more with experimental measurement standards in science and technology fields. Emphasis was given to data management problems common to data used outside of the field in which it was generated. CNC/CODATA was previously sponsored by CISTI. Currently supported by 2-year funding commitment by NRC to support revitalization. RDC leading steering committee to revise terms of service. Current website lists focus on developing RDM culture, supporting RDC and Portage's work to support RDM from national perspective, supporting institution-based approach to RDM services and infrastructure, supporting early career scientists, and science

communication. The current Terms of Reference are undergoing an update (not yet public/approved by the host, NRC).

GOFAIR

https://www.go-fair.org/

New initiative formed by joint commitment by CODATA, RDA, and WDS. Model is open and inclusive or individuals and organizations working together to support FAIR principles via Implementation Networks, which are self-funded and governed and focus on building technology, culture and training as part of an envisioned Internet of FAIR Data & Services.

FAIRsFAIR

https://www.fairsfair.eu/

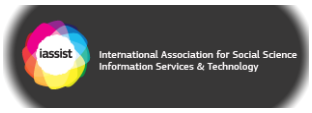
Role in the development of global standards for FAIR certification of repositories and data within them. Aims to provide platform for using and implementing FAIR principles in work of European research data providers and repositories. Programs support:

- Repository certification
- Training within universities and Professionalization of Data Stewardship
- Policies (e.g. surveying communities on FAIR policies at EU universities, semantics and interoperability)

Global Indigenous Data Alliance

https://www.gida-global.org/

An international network of Indigenous researchers, data experts, and policy makers devoted to advancing Indigenous control over Indigenous data through advocating for data sovereignty and Indigenous data governance at the international level and within nation-states.



International Association for Social Science Information Services & Technology (IASSIST)

https://iassistdata.org/

International organization of over 300+ members who are information professionals supporting data services in the social sciences. Convenes Action Groups to undertake specific tasks, find solutions to problems, as well as Interest Groups who share information on topics with larger membership.

- Current Interest Groups are domain focused: Geospatial, Qualitative Data, Health Data
- Previous Interest Groups had broader data management focuses on data visualization, open source tools, Data Management and Curation, Data Citation

OPENAIRE

https://www.openaire.eu/

Mission is to provide barrier free, open access to research outputs financed by public funding in Europe. Service programs include funding National Open Access Desks (NOAD) – network of 34 experts embedded in institutions across Europe to support Open Science, and delivering training opportunities.

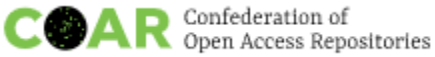
- NOADs and partners form task forces to advance open science, and includes and RDM task force
- Convenes a Community of Practice for Training Coordinators, which is an informal network to share experiences and map out related activities to strengthen training capacity.
- Develops open RDM software (e.g. Amnesia, ARGOS) and Dashboards for repository and funder metrics
- Convenes domain-focused communities to gather open data, software, and publications as “Community Gateways” (e.g. DARIAH-EU, EPOS).

FORCE11

https://www.force11.org/

Missions to improve research practices by supporting innovations in the ways knowledge is created and shared across research disciplines and communities by connecting the global communities interested in communications in research, providing space for discussion and collaborative work, facilitating development of new approaches and tools for effective digital communications in research.

- Has several joint RDA Working Groups, including FAIRsharing organization, other groups on software citation, attribution.



Confederation of Open Access Repositories (COAR)

https://www.coar-repositories.org/

International association with 157 members that brings together individual repositories and repository networks to build capacity, align policies and practices, and act as global voice for repository community. Programs cover research data management, metadata and vocabularies, training and capacity building.

- RDM Interest Group aims to help community to expand operations, provides forum for community discussions about managing research data, best practices and strategies, and capacity building for RDM in the repository community.

Allied Initiatives

There are also a number of allied international initiatives focused on supporting communities of practice in fields related to research data management, which could also be strategically engaged in the Alliance's data management stream to support the Canadian data management community. These include:

Software/Tools

Research Software Alliance

https://www.researchsoft.org/

ReSA aims to bring research software communities and organizations together to develop a community of practice to address challenges in software productivity, quality, reproducibility, and sustainability, to achieve their shared long-term goal of research software valued as a fundamental and vital component of research.

ReSA currently leads a number of taskforces on FAIR principles for research software, evidence for the importance of research software, funding opportunities, and landscape analysis.

Research Information Management

Metadata2020

http://www.metadata2020.org/

Metadata 2020 is a collaboration that advocates for richer, connected, and reusable, open metadata for research outputs. Convenes a series of Community Groups who are responsible for defining metadata challenges, barriers and opportunities for their area of scholarly communication. Communities include researchers, publishers, librarians, data publishers and repositories, services, platforms and tools, and funders.

ORCID-CA

https://orcid-ca.org/home

ORCID-CA is the ORCID Consortium in Canada. They provide institutions and organizations with membership to ORCID at a reduced cost as well as access to community support services.

DataCite

https://datacite.org/steering.html

DataCite is a world leading provider of persistent identifiers for research outputs. To help guide their development, DataCite has formed three Steering Groups: Sustainability and Business, Services and Technology, and Community and Engagement. Steering Groups provide a venue for open participation by interested community members who support strategies related to sustainability planning, services, and outreach.

Global Research Collaborations

Within various research domains, a number of global collaborations are advancing RDM support and best practices. Examples with strong ties to the Canadian research landscape include:

Global Research Collaboration for Infectious Disease Preparedness (GloPID-R)

https://www.glopid-r.org/

GloPID-R is an Alliance of research funding organizations to facilitate effective and rapid research on significant outbreaks of new or re-emerging infectious diseases.

Organizes a number of working groups related to two main work streams: Preparedness and Response. Related to RDM, a Data Sharing Working Group coordinates initiatives to support timely and transparent sharing of data in public health emergencies, and has released a related roadmap to guide this work.

Global Open Data on Agriculture and Nutrition (GODAN)

https://www.godan.info/

A network of over 1,000 members from national governments, NGOs, international and private sector organizations. Combining open data advocacy and consultancy with innovative products and solutions, GODAN and partners are looking to improve food security for generations to come, ensure zero hunger and improve the lives and livelihoods of farming communities across the globe.

Partners aim to build high level policy and private sector support for open data. They also encourage collaboration and co-operation across existing agriculture, nutrition and open data activities and stakeholders to solve long-standing global problems.

Global Alliance for Genomics and Health (GA4GH)

https://www.ga4gh.org/

An international Alliance that brings together 500+ organizations in healthcare, research, patient advocacy, life science, and information technology. The GA4GH community advances frameworks and standards to enable the responsible, voluntary, and secure sharing of genomic and health-related data.

The Alliance convenes a number of Foundational and Technical work streams in the areas of regulatory, ethics, and data security in genomics. Current Foundational work streams include data security and Regulatory & Ethics. Current Technical work streams include clinical and phenotypical data capture, cloud, data use and researcher identities, discovery, genomic knowledge standards, and large scale genomics.

International Virtual Observatory Alliance (IVOA)

https://ivoa.net/

An international Alliance of more than 20 international astronomical data centres, with the objective facilitating coordination among tools, systems, and organizational structures via the development of a shared set of standards. The IVOA constitutes Working Groups that propose recommendations for interoperability standards and technologies. The IVOA also has Interest Groups that discuss experiences using virtual observatory technologies for their improvement.