Research Data Canada – Infrastructure Subcommittee

Activity 1: Identify infrastructure available for data in Canada

Research Data Canada (RDC) has established the Infrastructure subcommittees (RDC-I) to tackle research data management infrastructure issues. RDC-I established a workplan in March 2013 and obtained approval for on it April 8. The work plan consists of a number of activities.

This document summarizes the findings of the subcommittee related to activity 1: "identity infrastructure available for data in Canada". The expected outcome of the activity is a "list of infrastructure in existence or planned, with capabilities and management structure outlined".

The document presents the results of the RDC-I team investigation, to the best of their knowledge. It does not claim to provide an exhaustive list of facilities and infrastructures but makes every effort to cover the subject with ample details.

The document is divided up in three sections and three appendices. The first section establishes a few definitions that frame the concepts behind "Digital Infrastructure" when it relates to research data. The second section presents some of the existing actors on the Canadian scene who play an enabling role by providing some of the necessary technical elements. In a third section, the document lists the larger such infrastructures i.e., those with a national scope that cover one or more connected science disciplines, have a reasonably long life expectancy and reasonably sustained funding.
The appendices provide more details and compare research data support infrastructures briefly listed in the sections below. The last appendix covers more specifically research data that are typically attached to short- to medium duration research projects. Those will typically require either continued funding beyond the project end or a well planned transfer of their data holdings to long-lived data centres to ensure their maintenance and accessibility.

I.   Context and definitions

The following table defines terms and sets the stage for the remainder of the document.

Term	Definition in the context of RDC
Research Data	The collection of artifacts that have been assembled during research projects, regardless of their discipline. Research projects will be assumed to have followed a scientific methodology. Research data may consist in raw observations of phenomena, calibrated data in which observational biases have been removed, reduced data, analysis results, model output and conclusions. Research data is to be accompanied with metadata that explains the context of the observations (e.g., sampling methodology), and the precise description of the methods followed to go from raw data to final conclusions. Research data can be analog or digital.

Digital Infrastructure	The set of interconnected structural elements that provide a framework supporting collection, archival, search, processing, and analysis of research data. They consist of an ensemble of essential enablers that include networking, computing, storage, metadata, and curation personnel. Moreover, the Digital Infrastructure is capable of interoperating with other similar infrastructures.
Advanced Digital Infrastructure	Advanced Digital Infrastructure is the aggregate of hardware, software, protocols & human expertise, coordinated in an integrated & broadly accessible environment, to support data-intensive research, education and innovation in Canada and in connection with the world. (Source: CUCCIO).
Networking	The methods used to transport data from one place to the other, regardless of its digital or analog nature.
Computing	The set of resources that can be used to perform transformations and translations of data (e.g., digital to analog conversion) and implement data access through user–storage mediation.

Table 1: Definitions of terms used.

II.   Summary list of Infrastructure in support of Research Data in Canada

The following table is a non-exhaustive list of infrastructures that provide enabling, technical capabilities in support of Research Data in Canada. Some of its members are presented in more details in Appendix A.

Name	Discipline-specific	Description	Location(s)
Compute Calcul Canada	No	Initially focusing on high performance computing, but now also providing longer-term data storage capability	Across the country.
Most universities	No	Most universities have some form of data management facilities for their researchers and smaller research projects. Larger projects would presumably be invited to use Compute Canada facilities.	Across the country
National Research and Education Network (NREN)	No	The interdependent aggregate network comprised of the federal research and education network operated by CANARIE, and the independent provincial and territorial networks known as ORANs (Optical Regional Advanced Network)	Across the country.

III.   Non-exhaustive list of large research data repositories

The table below contains a non-exhaustive list of established, national in scope, discipline-focused research data organizations that provide not only storage, but also curation and support. Together with other smaller projects, they are described in more details in Appendix B. Appendix C will provide examples of various projects, recently funded by CFI, that have a data production and storage component.

Name	Discipline	Description	Location(s)
Canadian Research Data Centre Network (CRDCN)	Social Sciences	The Research Data Centres (RDCs) offer Canada's research community secure access to detailed microdata from population and household surveys, censuses and administrative files in universities across the country.	Across the country
Ocean Networks Canada	Ocean-related disciplines	ONC's Data Management and Archiving system has a 25-year mandate to acquire, manage and make available data pertaining to the Ocean.	UVic/UofS
Canadian Astronomy Data Centre	Astronomy	Data Management of space- or ground-based astronomy project in which Canada is participating	NRC/ Victoria, UofS
Network Enabled Platforms	Yes, various	Middleware, which is funded in part by CANARIE, to simplify access to network based resources such as HPC, and data.	Across the country.

A. Detailed overview of Research Data Support Infrastructure in Canada

This appendix lists the various major science data support infrastructures in Canada, organizations that will enable data management from the technical side.

Facility	CANARIE
Description	Organization to advance Canada's knowledge and innovation infrastructure
Mission statement	CANARIE designs and delivers digital infrastructure and drives its adoption for research, education and innovation
Funding sources	Primarily Government of Canada
Headquarters	Ottawa, Ontario
President	Jim Roche
Board Chair	Dr. Howard Brunt
Staff total	30
DI contact	Jim Ghadbane, CTO
DI key personnel	Mark Wolff, Senior Director Technology Innovation Thomas Tam, Chief Network Engineer Chris Phillips, CAF technical architect
DI staff	15
Research Data Holdings	None
Data types supported	All
Storage Capacity	N/A
Data protection measures	
Expected DI lifetime	
Plans beyond funded lifetime	

Facility	Compute Canada
Description	This national platform integrates High Performance Computing (HPC) resources at six partner consortia across the country to create a dynamic computational resource. Compute Canada integrates high-performance computers, data resources and tools, and academic research facilities around the country. These integrated resources represent close to a petaflop of computing capability and online and long term storage with rapid access and retrieval over Canada's national, provincial and territorial high-performance networks.
Mission statement	To create a world-class sustained national platform of shared high performance computing and data resources and personnel, accessible by researchers in all disciplines independent of resource or researcher location and to promote high performance computing nationally and internationally.
Funding sources	Canada Foundation for Innovation, Provinces, Institutions and other various sources
Headquarters	Virtual

Interim President	*Jill Kowalchuk*
Board Chair	*Don Hathaway*
Staff total	*~150*
DI contact	John Morton, Technical Manager, SHARCNET
DI key personnel	Rob Simmons, CTO, Westgrid Suzanne Talon, Coordinator, CQ Greg Lukeman, CTO, Acenet Chris Loken, CTO, SciNet Chris MacPhee, CTO, HPCVL
DI staff	~150
Research Data Holdings	Various
Data types supported	
Storage Capacity	> 15PB available across the platform with ~10 major storage pools
Data protection measures	- Data behind a protected network - Some data backed up and some backed up off site
Expected DI lifetime	
Plans beyond funded lifetime	strategic plan being developed now

B. Detailed overview of active major Research Data organization in Canada

Facility	Canadian Astronomy Data Centre
Description	The Canadian Astronomy Data Centre (CADC) is part of the National Research Council of Canada (NRC) and operates under its mandate to serve the university research community. It hosts data collections from most of Canada's major observatories and serves 4500 users world-wide.
Mission statement	To provide essential services to support astronomy research in Canada and internationally.
Funding sources	National Research Council Canada, CANFAR.
Headquarters	5071 West Saanich Road, Victoria, BC
Group Leader	Dr. David Schade
Staff total	~22
DI contact	David Schade, Group Leader
DI key personnel	Severin Gaudet, Manager, Software Development John Ouellette, Manager, Operations JJ Kavelaars, Science Coordinator
Research Data Holdings	Observational data from major Canadian observatories and other science projects.
Data types supported	Astronomy data in various frequency ranges. Over 100 instruments from multiple observatores on the ground and in space.
Archive holdings	> 400TB > 50 million individual files Four distinct copies
Data protection measures	- Data replicated in 3 geographical locations. - Protected by authorization mechanism
Expected DI lifetime	Designed to support ongoing operations.
Plans beyond funded lifetime	Funding is ongoing.

Facility	Canadian Research Data Centre Network
Description	Since 2000, the Canadian Research Data Centre Network (CRDCN), in partnership with Statistics Canada's Research Data Centre Program, has transformed quantitative social science research in Canada. In secure computer laboratories on university campuses across Canada, university, government and other approved researchers are able to analyse a vast array of social, economic and health data. The Research Data Centres (RDCs) offer Canada's research community secure access to detailed microdata from population and household surveys, censuses and administrative files in universities across the country. Staffed by Statistics Canada analysts, RDCs operate in accordance with the confidentiality provisions of the Statistics Act and are accessible only to researchers with approved projects and security clearance

Mission statement	1) To improve data access by giving researchers across the country access, free-of-charge, to detailed micro-data from an increasing range of survey, census and administrative data. 2) To expand the pool of skilled quantitative researchers in Canada and train the next generation of researchers. 3) To make research count by improving communication between social scientists and the potential users of the knowledge they create.
Funding sources	Canada Foundation for Innovation, SSHRC/CIHR
Headquarters	McMaster University, Hamilton, ON
President	
Board Chair	
Staff total	
DI contact	
DI key personnel	
DI staff	
Research Data Holdings	
Data types supported	
Archive holdings	
Data protection measures	
Expected DI lifetime	
Plans beyond funded lifetime	

Facility	Ocean Networks Canada
Description	Ocean Networks Canada (ONC) is a world-leading organization supporting ocean discovery and technological innovation. ONC is a not-for-profit society, established in 2007 by the University of Victoria under the BC Society Act. Under a Management Agreement with the University, the purpose of ONC is to govern, manage and develop: the Ocean Networks Canada Observatory (comprised of the VENUS and NEPTUNE Canada networks) as a national research platform; and the ONC Centre for Enterprise and Engagement as a federal centre of excellence for commercialization and research.
Mission statement	To enable transformative ocean research for the advancement of science and technology and for the benefit of Canada.
Funding sources	Canada Foundation for Innovation (40%), Province of BC (40%), various sources (20%)
Headquarters	Victoria, BC (UVic Campus)
President	Dr. Kathryn Moran
Board Chair	Andrew Bjerring
Staff total	~80
DI contact	Benoît Pirenne, Associate Director, Digital Infrastructure
DI key personnel	Eric Guillemot, Manager, Software Development Martin Hofmann, Manager, Systems and Operations Reyna Jenkyns, Manager, Data Stewardship
DI staff	~30

Research Data Holdings	Observation data from the ocean. Raw and calibrated data.
Data types supported	- Time series of scalar data (e.g., physical and chemical measurements) - Multi-dimensional snapshots of ocean conditions (e.g., current velocities, backscatter maps, still images, ...) - Data streams (e.g., video, audio data) Data are coming from > 120 different instruments, representing > 50 different types of instruments
Archive holdings	> 120TB as of mid-2013 > 10^{10} individual scalar measurements > 6.6 million individual files
Data protection measures	- Data behind a protected network - Data replicated in quasi real-time at a secondary site at UofS (on Compute Canada resources)
Expected DI lifetime	Designed to support 20+ years of operations
Plans beyond funded lifetime	Data to be transferred in a Canadian, US or European "World Data Centre".

Facility	Ocean Tracking Network
Description	The Ocean Tracking network is an international infrastructure that uses electronic telemetry to provide knowledge of the movements, survival, and habitats of valued aquatic species, and of how all are linked to environmental conditions. The OTN fosters technological and operational innovation ins support of its scientific objectives. OTN operates oceanographic equipment in support of its science, including profiling (2 each Slouch Electric Gliders) and surface (1 ea. Liquid Robotics Wave Glider) autonomous vehicles, and moored oceanographic instrumentation (benthic pods, ADCPs). Canadian and International researchers work with the infrastructure to address compelling scientific questions. A social science component assists with translating research results into sustainable management of valued marine resources, and policy development.
Mission statement	To create a global partnership to construct and sustain a scientific platform and the associated trained personnel to collect, store, share, analyze, and use aquatic tracking and environmental data to support sustainable management of valued aquatic species.
Funding sources	Canada Foundation for Innovation International Joint Venture Fund (Infrastructure); NSERC Network Grant (For Canadian researchers to work with the Infrastructure); SSHRC; Operation and Maintenance support from a broad array of international partners drawn from industry, academia, government, and NGO sectors.
Headquarters	Dalhousie University, Halifax, Nova Scotia
President/PI	Dalhousie University President Richard Florizone. Dalhousie University is ultimately responsible for the OTN. The PI is Dr. Sara Iverson of Dalhousie University.
Board Chair	The OTN Council is chaired by Peter Harrison, and is appointed by the Dalhousie University President.
Staff total	CFI funded staff- 9. NSERC funded staff -2. Dalhousie funded staff – 1.5
DI contact	Bob Branton
DI key personnel	Bob Branton, Lenore Bajona
DI staff	5

Research Data Holdings	Fisheries Management, Oceanography; Marine Conservation; MPA/EBSA Planning; Marine Biodiversity; Polar, Temperate and Tropic Aquatic Sciences; Fish Movement and Migration
Data types supported	Scalar
Archive holdings	53+ million records consisting of 32+ million detections from 3,930 receiver stations, 31+ thousand acoustic/satellite/sensor tagged marine animals, 52 species, 15 ocean regions, 164 projects, 73 institutions, 14 countries and growing, 92 GB (PostGIS enabled postgresql database), plus ADCP and Benthic Pod data approx. 3 GB (submitted to DFO for processing/archive/distribution).
Data protection measures	Protected by Dalhousie University firewalls. Triple backup of existing records. Regular offloads to archival facilities at DFO
Expected DI lifetime	Permanent
Plans beyond funded lifetime	DFO will maintain a permanent archive.

C. Research Projects with Data Storage Components

In this appendix, a non-exhaustive list of recently approved CFI research project with a data storage component explicitly requested by the PI are listed. Beyond the summary description obtained from CFI, the PIs were asked about their understanding of the data management and about their plans related to the data management resources to be devoted to their project.

Project Host	Carleton University
Description	A Distributed Data Management Infrastructure for Indigenous Knowledge Research with Inuit and First Nations Communities
Keywords	local knowledge, data management, arctic, Inuit, cartography, digital, citizen science, policy, sea ice, routes, rescue, development
Funding sources	Canada Foundation for Innovation, LEF2012
Headquarters	Carleton University
Principal Investigator	Fraser D. R. Taylor
Project Staff total	
Digital Infrastructure contact	
Digital Infrastructure staff number	
Research Data Holdings	Geography/native studies
Data types supported	
Archive holdings	
Data safeguarding measures	
Expected Digital Infrastructure lifetime	
Plans for the data beyond funded lifetime	

Project Host	Queen's University
Description	Leading The Search For Dark Matter and Double Beta Decay:Improved Discovery Potential For DEAP-3600 and SNO+ Experiments
Keywords	Dark Matter particles, neutrino-less double beta decay a rare form of radioactivity, low radioactivity argon
Funding sources	Canada Foundation for Innovation, LEF2012
Headquarters	Queen's University
Principal Investigator	Arthur Mc Donald
Project Staff total	300
Digital Infrastructure contact	Jeff Harnell (SNO+); Chris Jillings (DEAP) Jeff Hartnell j.j.hartnell@sussex.ac.uk, jillings@snolab.ca
Digital Infrastructure staff number	
Research Data Holdings	astrophysics
Data types supported	Binary data based on common open-source standards, including root
Archive holdings	

Data safeguarding measures	Data at two separate centers. Strict secure login to servers.
Expected Digital Infrastructure lifetime	~ 10- 20 years
Plans for the data beyond funded lifetime	Not developed as yet.

Project Host	**Simon Fraser University**
Description	Upgrade to the ATLAS Tier-1 Data Analysis Centre
Keywords	Higgs particle, supersymmetry, Standard Model, high performance computing, experimental particle physics, high energy proton-proton collisions
Funding sources	Canada Foundation for Innovation, LEF2012
Headquarters	Simon Fraser University
Principal Investigator	Michel Vetterli
Project Staff total	
Digital Infrastructure contact	
Digital Infrastructure staff number	
Research Data Holdings	Physics
Data types supported	
Archive holdings	
Data safeguarding measures	Handled by the ATLAS collaboration
Expected Digital Infrastructure lifetime	
Plans for the data beyond funded lifetime	

Project Host	**University of British Columbia**
Description	Genomic approaches to personalizing cancer diagnosis and treatment
Keywords	DNA sequencing, bioinformatics, personal genomics, mutation detection, cancer genomics
Funding sources	Canada Foundation for Innovation, LEF2012
Headquarters	University of British Columbia
Principal Investigator	Marco Marra
Project Staff total	8
Digital Infrastructure contact	Robyn Roscoe, rroscoe@bcgsc.ca
Digital Infrastructure staff number	6
Research Data Holdings	Genomics

Data types supported	Standard data files for genomics/bioinformatics
Archive holdings	7PB
Data safeguarding measures	Off-site tape backup
Expected Digital Infrastructure lifetime	5 years
Plans for the data beyond funded lifetime	The GSC pursues sustainability options on an ongoing basis, and would work to provide data access for as long s researchers or funders require it

Project Host	University of British Columbia
Description	Molecules to Human: Enhanced phenotyping for discovery, prevention, & treatment of heart, lung, & blood vessel disease
Keywords	cardiovascular, pulmonary, imaging, phenotyping, gene x environment, translational research, patients, disease, risk, prevention
Funding sources	Canada Foundation for Innovation, LEF2012
Headquarters	University of British Columbia
Principal Investigator	Darryl Knight
Project Staff total	
Digital Infrastructure contact	
Digital Infrastructure staff number	
Research Data Holdings	Health
Data types supported	
Archive holdings	
Data safeguarding measures	
Expected Digital Infrastructure lifetime	
Plans for the data beyond funded lifetime	

Project Host	University of British Columbia
Description	Systems analysis of single stem cells
Keywords	Stem Cells and Regeneration, Cancer Stem Cells, High Throughput Single Cell Analysis, Single Cell Genomics, Single cell Heterogeneity, Microfluidics Atomic Mass Spec-Based Cytometry, Microfluidics, Automated library construction
Funding sources	Canada Foundation for Innovation, LEF2012
Headquarters	University of British Columbia
Principal Investigator	Fabio Rossi
Project Staff total	
Digital Infrastructure contact	
Digital Infrastructure staff number	
Research Data Holdings	Genomics
Data types supported	

Archive holdings	
Data safeguarding measures	
Expected Digital Infrastructure lifetime	
Plans for the data beyond funded lifetime	

Project Host	Université du Québec à Montréal
Description	Hexagram, le Centre de recherche-création en arts médiatiques et technologiques, phase 2 UQAM
Keywords	recherche-création technologique, réflexivité, méthodologie, expérimentation, production, monstration, diffusion, transdisciplinarité
Funding sources	Canada Foundation for Innovation, LEF2012
Headquarters	Université du Québec à Montréal
Principal Investigator	Gisèle Trudel
Project Staff total	
Digital Infrastructure contact	
Digital Infrastructure staff number	
Research Data Holdings	Creative Arts
Data types supported	
Archive holdings	
Data safeguarding measures	
Expected Digital Infrastructure lifetime	
Plans for the data beyond funded lifetime	

Project Host	Université Laval
Description	Human and Microbial Integrative Genomics
Keywords	Bioinformatics, Computational Biology, Bioanalysis, Proteomics, Metabolomics, Cancers, Infectious Diseases, Microbiome, Genetic Predisposition, Biomarker
Funding sources	Canada Foundation for Innovation, LEF2012
Headquarters	Université Laval
Principal Investigator	Jacques Simard
Project Staff total	
Digital Infrastructure contact	
Digital Infrastructure staff number	
Research Data Holdings	Biology
Data types supported	

Archive holdings	
Data safeguarding measures	
Expected Digital Infrastructure lifetime	
Plans for the data beyond funded lifetime	

Project Host	**University of Guelph**
Description	Digital Biodiversity - From DNA Barcode Libraries to Applications
Keywords	DNA-based identification, DNA barcoding, sequencing, informatics platform, taxonomy, species
Funding sources	Canada Foundation for Innovation, LEF2012
Headquarters	University of Guelph
Principal Investigator	Paul Hebert
Project Staff total	120
Digital Infrastructure contact	Sujeevan Ratnasingham, Director of Informatics Sratnasi@uoguelph.ca
Digital Infrastructure staff number	17
Research Data Holdings	Biology, Genomics, DNA Sequence
Data types supported	Textual specimen metadata, DNA sequence data , instrument data, image data. Visit www.boldsystems.org for further details
Archive holdings	Not applicable
Data safeguarding measures	Failover site, offline backups
Expected Digital Infrastructure lifetime	5 years
Plans for the data beyond funded lifetime	Goal is to establish BOLD as a component of national research infrastructure or relocation to an international organization such as the Global Biodiversity Information Facility

Project Host	**University of Regina**
Description	Western Canada Integrated Environmental Simulation and Risk Assessment associated with Climate Modeling
Keywords	environment, climate modeling, downscaling, risk, simulation, optimization, adaptation planning, impact assessment, uncertainty, decision support, informatics
Funding sources	Canada Foundation for Innovation, LEF2012
Headquarters	University of Regina
Principal Investigator	Guohe Huang
Project Staff total	15
Digital Infrastructure contact	Chunjiang An
Digital Infrastructure staff number	4
Research Data Holdings	Climatology

Data types supported	Results of the research will be made available in digital form in spreadsheet tables, tab-delimited files, or image files. Main research products will be available online in digital form. Manuscripts will appear in PDF, and contain text, calculations, drawings, plots, and images. The targeted journals for the results of this research project all provide a downloadable PDF copy of the manuscript on the web. All of the computer software, subjective test data and model parameters will be available to interested parties upon request.
Archive holdings	Institute for Energy, Environment and Sustainable Communities
Data safeguarding measures	All electronic data generated by proposal research will be redundantly archived. We have a secure server on which all information is stored. The server hard drives are set up in a RAID that is capable of full recovery even in the case of multiple simultaneous disk failure. Additionally, the server drives are backed up on an independent server. This will allow full recovery of data in the event of catastrophic failure of the local laboratory server.
Expected *Digital Infrastructure lifetime*	10 years
Plans for the data beyond funded lifetime	We will continue to maintain archived data beyond funded lifetime. During that period, long-term data storage will be examined.

Project Host	University of Regina
Description	Western Canada Integrated Environmental Simulation and Risk Assessment associated with Climate Modeling
Keywords	environment, climate modeling, downscaling, risk, simulation, optimization, adaptation planning, impact assessment, uncertainty, decision support, informatics
Funding sources	Canada Foundation for Innovation, LEF2012
Headquarters	University of Regina
Principal Investigator	Guo Huang
Project Staff total	
Digital Infrastructure contact	
Digital Infrastructure staff number	
Research Data Holdings	Climatology
Data types supported	
Archive holdings	
Data safeguarding measures	
Expected Digital Infrastructure lifetime	
Plans for the data beyond funded lifetime	

Project Host	University of Waterloo
Description	Privacy Enhancing Technologies at a Global Scale

Keywords	Privacy enhancing technologies, global-scale, experimentation, implementation, private information retrieval, anonymous communications, censorship resistance
Funding sources	Canada Foundation for Innovation, LEF2012
Headquarters	University of Waterloo
Principal Investigator	Ian Goldberg
Project Staff total	
Digital Infrastructure contact	
Digital Infrastructure staff number	
Research Data Holdings	Computer Sciences
Data types supported	
Archive holdings	
Data safeguarding measures	
Expected Digital Infrastructure lifetime	
Plans for the data beyond funded lifetime	