IATUL
International Association
of University Libraries

IATUL Workshop on Information Literacy
23 - 24 November 2015
Prague, Czech Republic

Gwendolyn Ebbett
University of Windsor
Ontario, Canada

CARL members include Canada's twenty-nine largest university libraries as well as two national libraries. Enhancing research and higher education are at the heart of its mission. CARL develops the capacity to support this mission, promotes effective and sustainable scholarly communication, and public policy that enables broad access to scholarly information.

The Portage by [Winslow Homer](#), 1897

Two intertwined mandates:

▸ A library-based network of expertise on research data management, and

▸ National platforms for planning, preserving, and discovering research data

What do we mean by RDM?

RDM – Research Data Management – refers to the storage, access and preservation of data produced from a given investigation. Data management practices cover the entire lifecycle of the data, from planning the investigation to conducting it, and from backing up data as it is created and used to long term preservation of data deliverables after the research investigation has concluded.

Research Data Canada, Terms and Definitions
www.rdc-drc.ca/glossary/

Open science context

Open science means:

- Sharing and access to all types of research outputs
- Transparency of research findings
- Open peer review & open usage metrics
- Equitable flow of knowledge

O Began with open access to publications, but moving to data and other types of research outputs

O Parallels to Open Government/Open Data movement

What is driving this trend?

1. Verification, reproducibility and transparency of scientific results
2. New scientific discoveries through re-use datasets and greater diffusion of knowledge
3. Greater social and economic impacts through application of research outputs
4. And because we can…

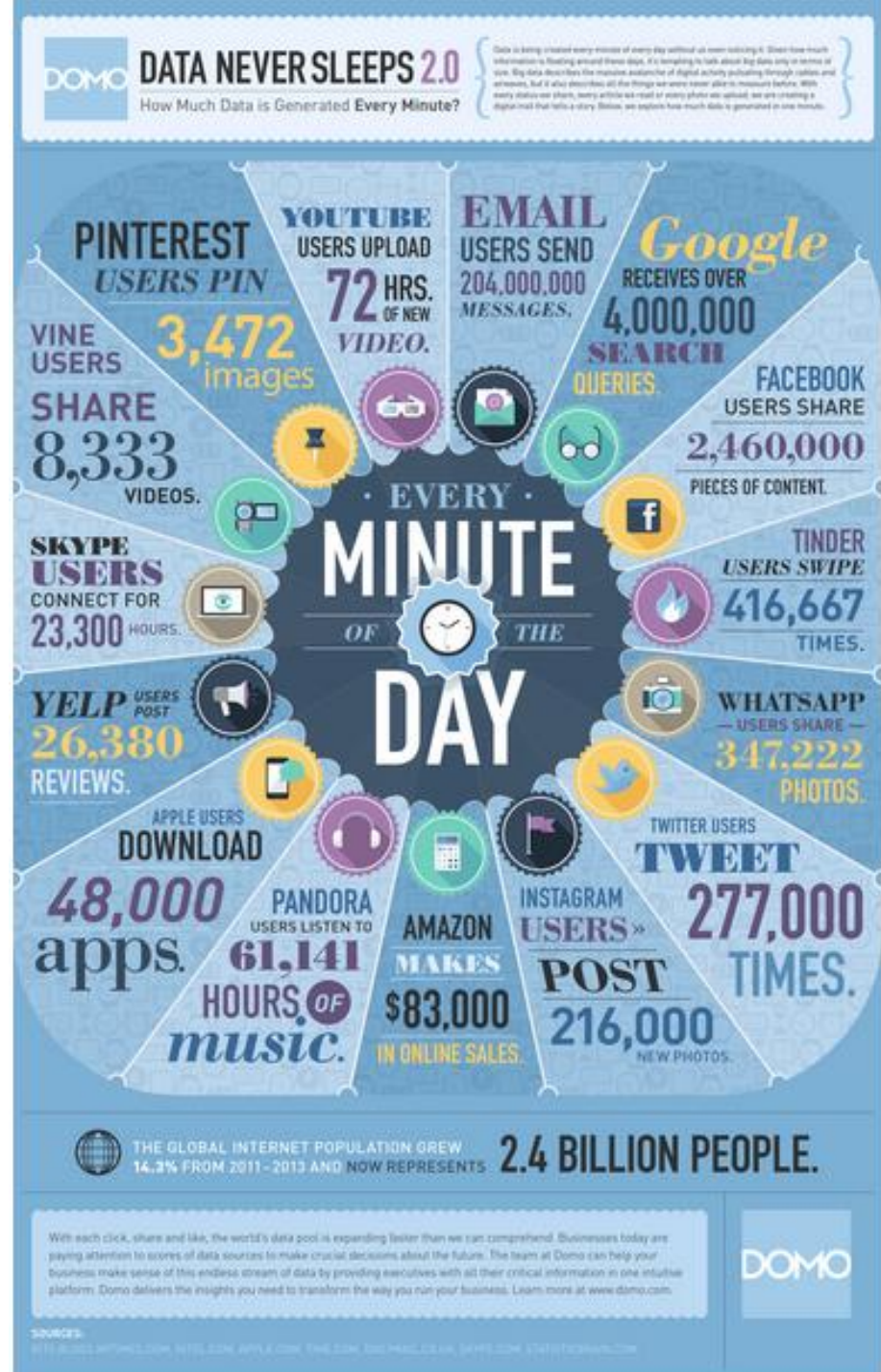**Figure ES.1 Average data storage cost for consumers**

1998-2012, per Gbit
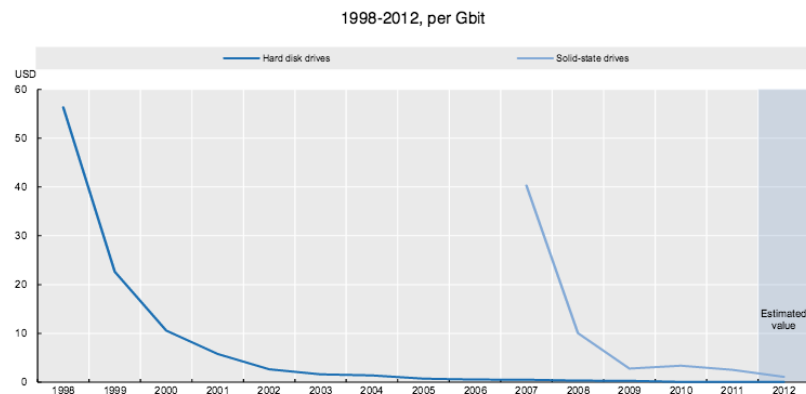
Source: OECD (2014), *Measuring the Digital Economy: A New Perspective*, OECD Publishing, Paris.
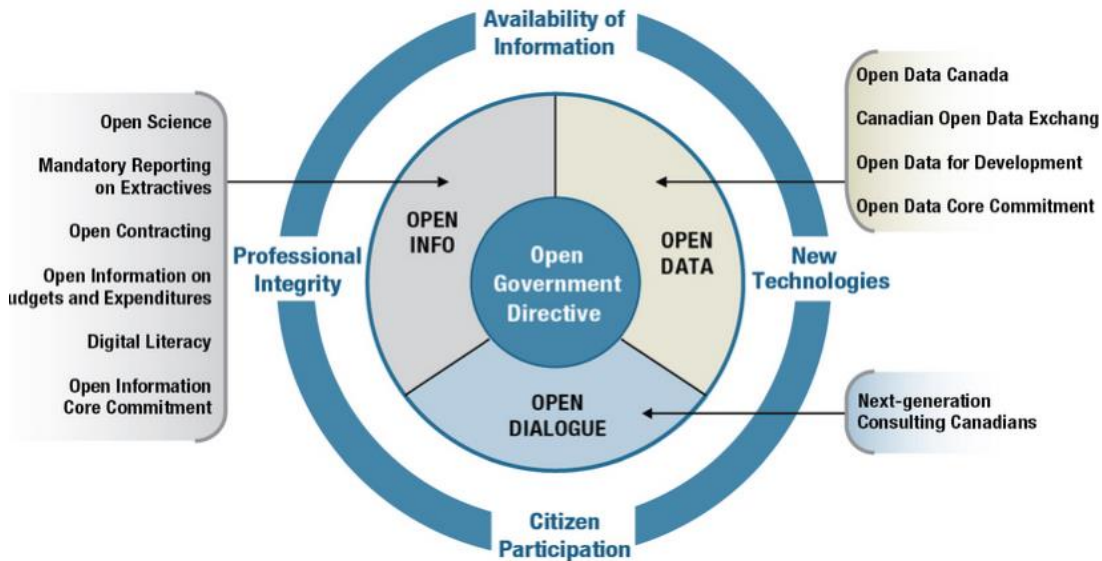
BIG DATA!

90% of world's data has been generated over last two years

Current policy context in Canada
Canada's Action Plan on Open Government 2014-2016

"…the Government of Canada will establish a government-wide approach to Open Science to increase access to federally-funded scientific publications and data."

Tri-Agencies (CIHR, NSERC, SSHRC)

- Consultation in 2013 (Report: *Capitalizing on Big Data: Toward a Policy Framework for Advancing Digital Scholarship in Canada*)
- Tri-Agency Open Access Policy on Publications, February 2015
- Comprehensive Brief on Data Management Policies, April 2015
- *Draft Tri-Agency Statement of Principles* on *Digital Data Management*

SCIENCE AND SCIENTISTS

"We will appoint a **Chief Science Officer** who will ensure that government science is fully available to the public, that scientists are able to speak freely about their work, and that scientific analyses are considered when the government makes decisions."

But we need more than just policies
Research Data Management is like a three-legged stool…

Culture

Infrastructure and services

Policies

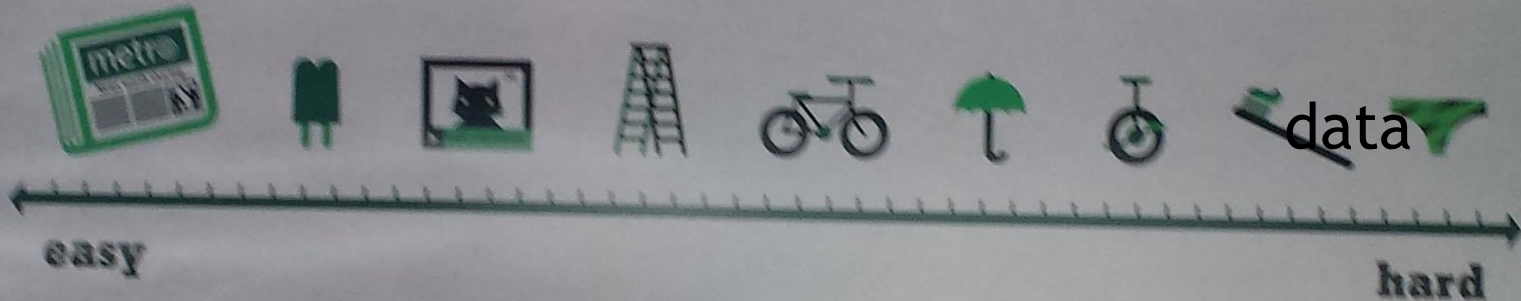# A lot of this is about ~~cultural change~~ socializing the community

Many researchers would rather share their toothbrush than their data...

Sharability

easy   hard

Infrastructure and services

Disciplinary repositories
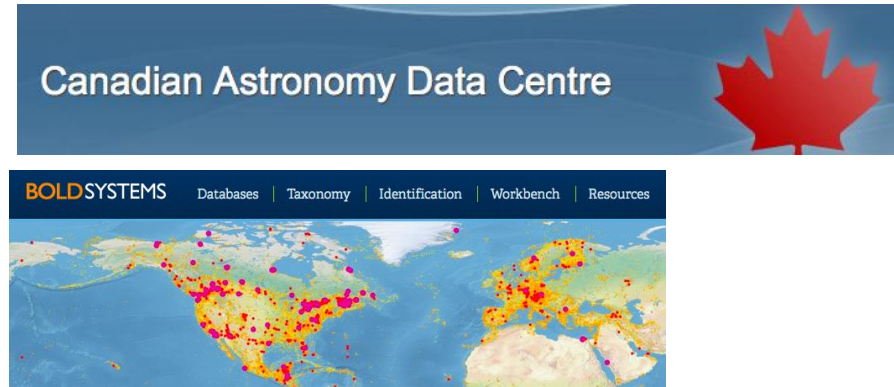
Integrated domain data centres
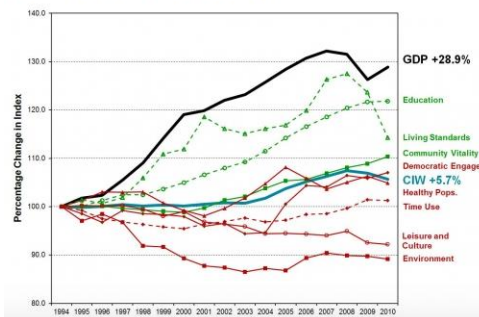
But these only collect a small portion of research data produced in Canada

The majority of datasets fall into the "long tail"

What is the size of the largest
data set that you have used or
generated in your research?
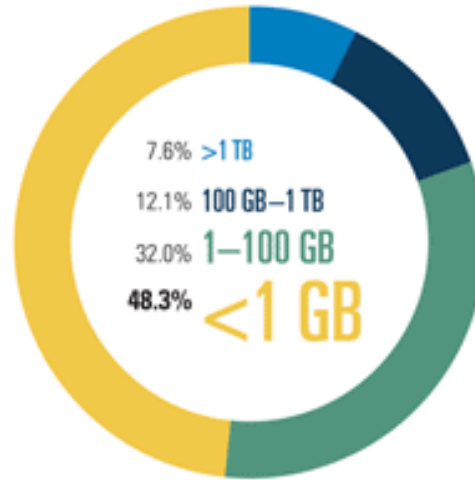
7.6% >1 TB
12.1% 100 GB–1 TB
32.0% 1–100 GB
48.3% <1 GB

1 TB
100 GB
1 GB

Size of data sets

Number of data sets

Where do you archive most of
the data generated in your
lab or for your research?

Even within a single
institution there are no
standards for storing
data, so each lab, or
often each fellow, uses
ad hoc approaches.

0.5% It is not stored
50.2% Our Lab
38.5% University servers
7.6% Community repository
3.2% Other
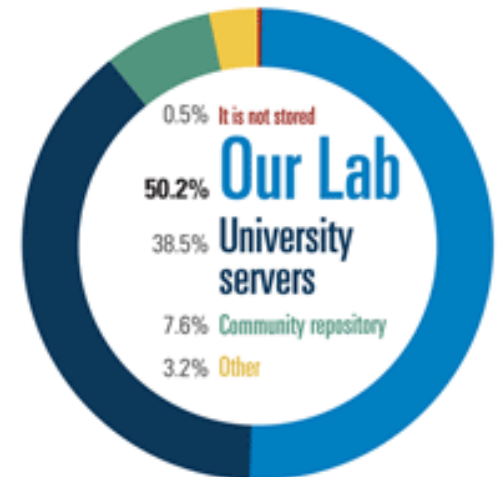
Libraries and RDM

Portage was initiated and supported by CARL, to build on past national and regional initiatives

▸ Canadian Research Data Summit 2011

▸ Research Data Canada launched in 2012: multiple stakeholders, together advancing research data stewardship in Canada

▸ CARL, a RDC stakeholder, began library RDM services planning in December 2013 and Project ARC (now Portage) in early 2014 with one-year project mandate

▸ RDC Federated Pilot began in early 2014, including CARL, Compute Canada, domain data centres

▸ CARL agreed to operationalize Portage in May 2015

**Chuck Humphrey
appointed Director in
September 2015**

Building on Regional Library Strengths

OCUL Ontario Council of University Libraries

OLRC Ontario Library Research Cloud

The OLRC Storage Service is a cloud storage network, initially created for the use of OCUL members. This storage network seeks to provide an alternative to commercial cloud storage providers to allow OCUL members to more closely control their storage costs and maintain control of their data. This storage service utilizes industry-standard APIs and interfaces to maximize compatibility with existing library use cases, including institutional repositories and other preservation repositories.

5 duplicate nodes

11 Ontario Libraries to date

INSTITUTIONAL PORTALS

Other text mining portals at OCUL libraries or university research units

Canadian Text Archive Centre

Service of UTL that provides for enhanced discovery and analysis of UTL collections and shared OCUL collections

WEB SERVICES

CORE

Discovery Services (e.g specialized indexing, enhanced metadata, corpus creation)

Data processing services (e.g. Cloud provisioned VMs Hadoop)

EXTENDED

e.g. Data visualization Speech recognition Translation

STORAGE SERVICES

UTL/SP Storage Cluster

Subscription services for non-partner libraries

OCUL Storage Cluster 200TB

Contains institutional collections

SP storage for shared collections 200TB (e.g. Journals, Books, ODESI, GeoPortal, Gale, Adam Matthews)

UTL storage for UTL collections 800TB (e.g. IA, Hathi Trust, born-digital and digitized UTL resources)

OCUL

UTL

With RDM Stakeholders

1. Network of Expertise

Build on expertise and interests across the country in specific areas of research data management:

- data management planning
- privacy, security, and confidentiality
- data curation
- data preservation
- data dissemination and access
- data discovery
- skills and training

2. Develop and support national platform services for data planning, preservation and discovery

▸ Operations
- ➚ Data management planning tool
- ➚ Ingest and repository services
- ➚ Preservation services to maintain data long-term
- ➚ Aggregated discovery tool
- ➚ Metadata guidelines and procedures

▸ Service model
- ➚ Collaborative development with constellation of partners
- ➚ Local, regional and national configurations
- ➚ Institutions lacking local mechanisms will have access to services via other participants acting as host sites

Infrastructure-as-a-Service

Platform-as-a-Service

Software-as-a-Service

Infrastructure:
- CPUs
- Disk Drives
- Networks
- Data Centers

Platform:
- Development and Runtime Tools and Environment

Applications:
- Enterprise Apps
- Desktop Apps
- Mobile Apps
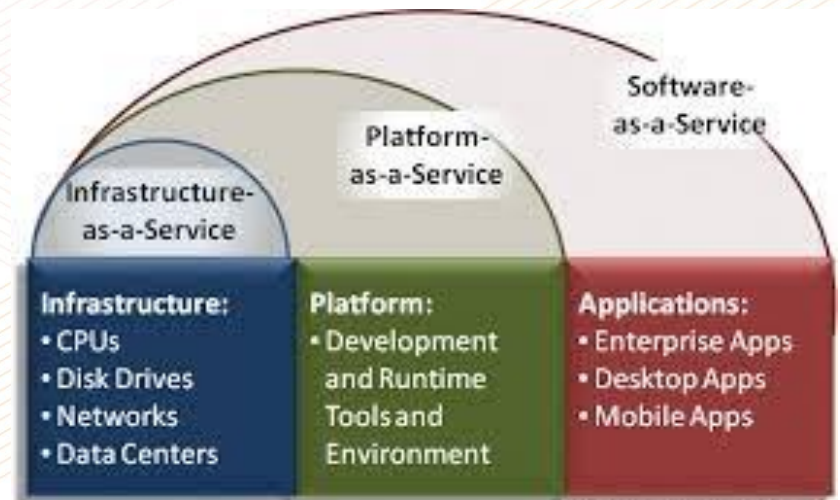
Source: ChainLink Research

First service launch

http://portagenetwork.ca
http://reseauportage.ca

Portage DMP Expert Group

Eugene Barsky, UBC
Jay Brodeur, McMaster
Talia Chung, UOttawa
Carla Graebner, SFU
Alex Guindon, Concordia
Chuck Humphrey, UAlberta
Amber Leahey, Scholars Portal
Jeff Moon (Chair), Queens
Carol Perry, Guelph
Kathleen Shearer, CARL
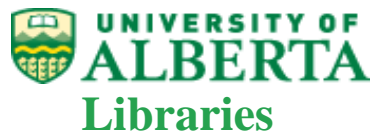
- *DMP Assistan*t was adapted from the (UK) Digital Curation Centre's *DMPOnline* tool and is hosted at the University of Alberta. The French user interface was produced with support from the Université de Montréal Library

- This national, open, bilingual data management planning tool will assist researchers throughout their projects.

portage

Shared stewardship of research data

Home   About   Help

DMP Assistant is a bilingual tool for preparing data management plans (DMPs). The tool follows best practices in data stewardship and walks researchers step-by-step through key questions about data management.

Sign in +

If you have an existing account with DMP Assistant or previous

Step 1 Sign up with DMP Assistant

Step 2 Sign in and select a template under Organizations. The Portage template is the default.

Step 3 Answer the questions that are relevant to your work. Guidance and examples are provided.

Step 4 Revisit the tool throughout your research to review or revise your answers.

gle sign-in authentication. For
nt. You will have the option to
ID when that feature

Summary of DMP Sections and related questions

Sections	Questions
Data Collection	- What types of data will you collect, create, link to, acquire and/or record? - What file formats will your data be collected in? Will these formats allow for data re-use, sharing and long-term access to the data? - What conventions and procedures will you use to structure, name and version-control your files to help you and others better understand how your data are organized?
Documentation and Metadata	- What documentation will be needed for the data to be read and interpreted correctly in the future? - How will you make sure that documentation is created or captured consistently throughout your project? - If you are using a metadata standard and/or tools to document and describe your data, please list here.
Storage and Backup	- What are the anticipated storage requirements for your project, in terms of storage space (in megabytes, gigabytes, terabytes, etc.) and the length of time you will be storing it? - How and where will your data be stored and backed up during your research project? - How will the research team and other collaborators access, modify, and contribute data throughout the project?
Preservation	- Where will you deposit your data for long-term preservation and access at the end of your research project? - Indicate how you will ensure your data is preservation ready. Consider preservation-friendly file formats, ensuring file integrity, anonymization and de-identification, inclusion of supporting documentation.
Sharing and Reuse	- What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final). - Have you considered what type of end-user license to include with your data? - What steps will be taken to help the research community know that your data exists?
Responsibilities and Resources	- Identify who will be responsible for managing this project's data during and after the project and the major data management tasks for which they will be responsible. - How will responsibilities for managing data activities be handled if substantive changes happen in the personnel overseeing the project's data, including a change of Principal Investigator? - What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?
Ethics and Legal Compliance	- If your research project includes sensitive data, how will you ensure that it is securely managed and accessible only to approved members of the project? - If applicable, what strategies will you undertake to address secondary uses of sensitive data? - How will you manage legal, ethical, and intellectual property issues?

portage

Data Collection

– What **types of data** will you collect, create, link to, acquire and/or record?

– What **file formats** will your data be collected in? Will these formats allow for data re-use, sharing and long-term access to the data?

– What **conventions and procedures** will you use to **structure, name and version-control your files** to help you and others better understand how your data are organized?

Collecting information and 'educating' researchers

Second: National Platform Service with Compute Canada (in progress)
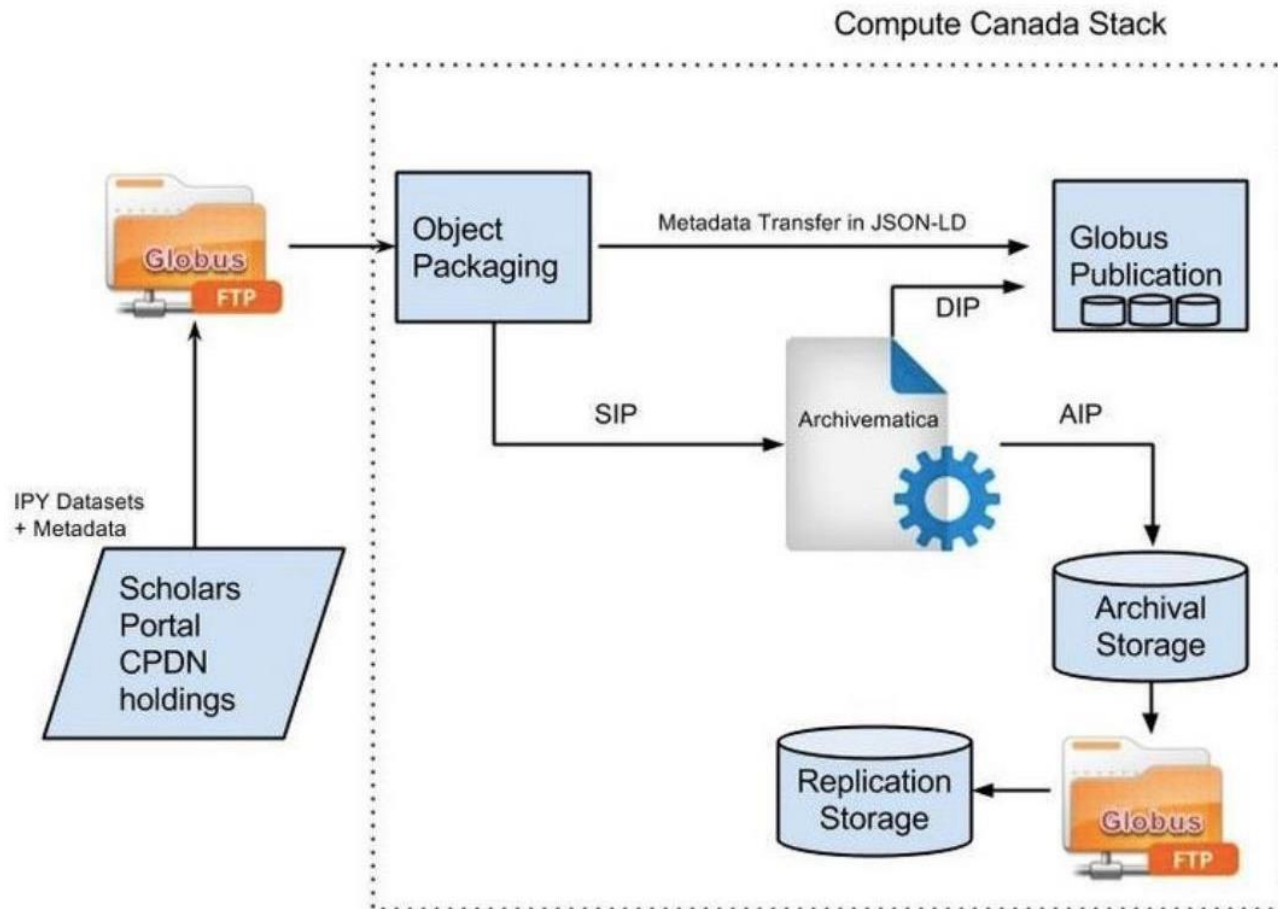
National Platform Service refers to a production-ready service for the ingest of research data and metadata, processing micro-service functions related to research data preservation such as characterization and normalization, generation of Archival Information Packages (AIPs) and Dissemination Information Packages (DIPs), transfer of AIPs to archival replicated storage, transfer of DIPs to a data repository for access control and discovery purposes. This is an application of a Platform as a Service for research data.

Project ARC Preservation Workflow Concept

Approved repositories will:
- be able to transmit data to the Interoperability Layer
- provide a crosswalk from any internally used DMD formats to a standard "lowest common denominator" DMD standard. e.g., DataCite, DC
- be able to link deposited data to grant #

User knows DOI

DataCite Canada

DOI resolution request

DOI Transactions

YES Refer to repo

Repository exists?

NO Refer to Archive

Search subset of DataCite MD repository

User doesn't know DOI

TC3+ Approved Repositories

Custom search app or domain-specific search apps

General Search

Search Options

Hydra Repos

Islandora VREs

preformed Standard Data Interoperability Packages

Compute Canada-hosted infrastructure

Metadata is used to populate

Dataverse Networks

CADC

Deposited in standard format

Archivematica Instance(s) managed by CARL/RDC

Writes Data Package to

Compute Canada Storage Systems

Facilitates re-use on CC platforms

Compute Canada HPC Facilities

Funded researcher

ATLAS

CBRAIN

Researcher deposits data in any approved repository in which they are comfortable.

Replicated to

Standard Data Interoperability package includes:
- Data as provided
- Data as transformed for reuse/preservation purposes, if necessary
- Domain specific DMD in native format
- DMD in transformed, "lowest common denominator" format (DataCite, DC?)

Resources for transformation include Archivematica Format Policy Registry (FPR), possible other sources.

Local/Regional Storage Resources (e.g., ODLRC, EduCloud, Institutional Storage)

Researcher

Accesses/processes data for replication/reuse. If reusing, deposits data back into approved repository.

Creates a Standard Data Interoperability Package

Deposit is made through existing CC endpoints such as Globus and FTP access.
Deposit should include:
- DOI (if assigned already)
- CASRAI ID / Grant ID
- Domain-specific descriptive metadata that is crosswalkable per the repo's provided crosswalk

Facilitates re-use on local platforms

Local/Regional Compute Facilities

Next steps

- Communications plan
- Governance model
- Data preservation expert group

portage

SHARED STEWARDSHIP OF RESEARCH DATA

Acknowledgements and thanks to

Chuck Humphrey, Portage Director, CARL
Jeff Moon, Chair, Portage DMP Expert Group
Kathleen Shearer, Research Associate, CARL
Martha Whitehead, President, CARL

Questions?