compute canada

Research Data Management

Novembre 18, 2015

RDM in Compute Canada

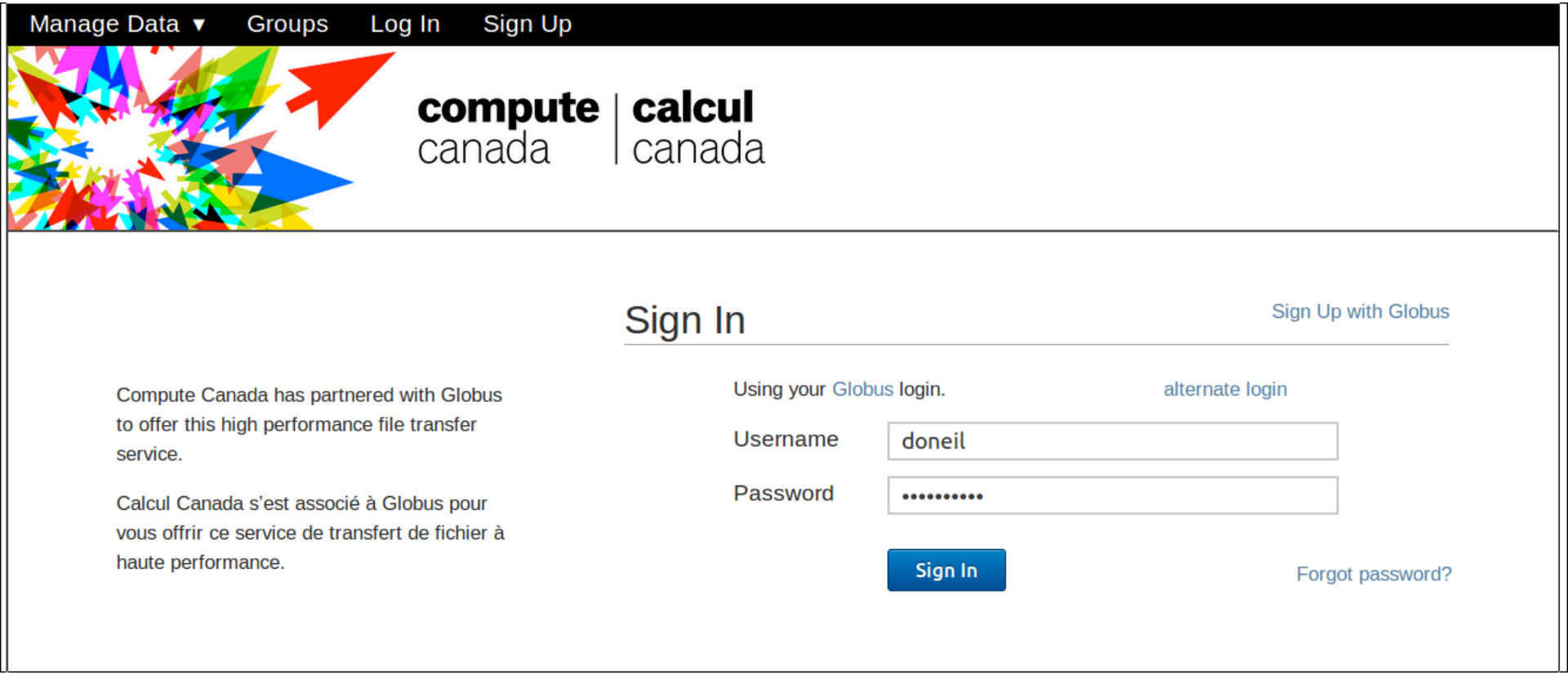
- Traditionally, Compute Canada gave access to significant storage and left Research Data Management to the researcher.

- This worked when only a small number of very experienced research teams had significant RDM needs (e.g., HEP, Astronomy)

- Today, everybody is accumulating data they want to store, analyze, preserve and make available to others.

- We are set up to provide common infrastructure. We should provide services that make RDM easier:
    - National deployment of Globus (software) - 2014
    - Owncloud (software) - 2014
    - National storage infrastructure (hardware) - RFP coming soon!

Globus File Transfer



- Web interface to CC storage
- Users define their own access groups, sharing enabled
- Optimized, scalable data transfers (labs, laptops, etc.)

compute canada | calcul canada

OwnCloud

- More like “Dropbox for Scientists”
- Smaller data sharing needs
- Open source, many plugins (e.g., shared calendars)

compute | calcul
canada | canada

National Data Infrastructure

- Traditionally, we bought large computing systems and bought the storage “to go with them”.
- With our recently announced funding, we will be doing something quite different - designing and purchasing a national data infrastructure.
- Deploy storage (disk and tape) at four locations across the country. Object and block storage, geo-replication service.
- Users will be able to login to any site and see datasets across the network.
- All four sites to be co-located with significant compute facilities, to come a few months later.
- RDM services to sit “on top” of hardware designed to support it.

Essential Components of RDM

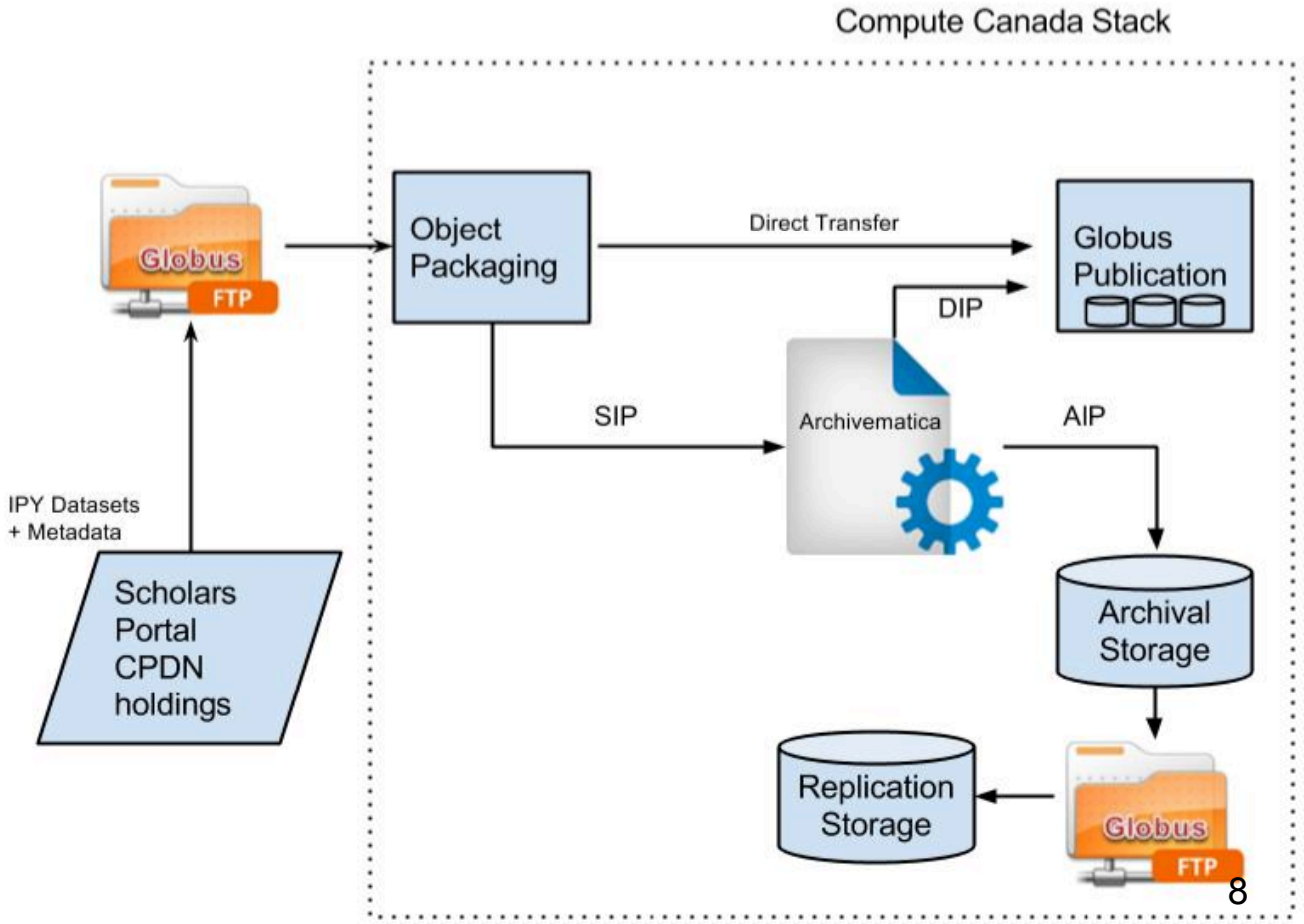
- We have learned a lot from our friends at CARL (library community) and we are still working with them.
- RDM services should include:
  - Provision of reliable repositories to store datasets
  - Collection of metadata during ingestion of those datasets
  - Curation of datasets into collections, QA on metadata and data
  - Services to find (discover) datasets in repositories
  - Preservation and archival processes
    - Transformation (normalization) of file formats into durable / long-term standards
    - Migration to new file formats when standards change
    - Creation and storage of archival packages

RDM in Compute Canada

- In addition to these essential components, CC sees the need for:
    - Repository federations - CC would never “own” all of the repositories (institutions, libraries)
    - Scalable data movement, storage.
    - Discoverability across many data collections and repositories.
    - Automatic geo-replication, data security
- Tested several sets of tools, combined a set into a viable candidate solution.
- Worked closely with CARL. Key collaboration in this space:
    - Researchers
    - Librarians
    - Infrastructure providers

CC-RDC-PORTAGE Federated RDM Pilot: Canadian Polar Data Network

CPDN is the domain repository for the Canadian International Polar Year (IPY) and Northern research data



globus

Publish | Manage Data | Groups | Support | hlady

Browse & Discover | Data Publication Dashboard | Communities & Collections

Search

Canadian International Polar Year Data Collection

Collection home page

Browse

Issue Date | Author | Title | Subject

Submit to This Collection

Collection's Items (Sorted by Submit Date in Descending order) 1 to 20 of 118

next >

Issue Date	Title	Author(s)
4-May-2015	Arthropod monitoring on Herschel Island	-
5-Mar-2013	Small mammal monitoring on Herschel Island and Komakuk Beach	-
3-Feb-2012	Breeding activity of shorebirds on Herschel Island	-
5-Mar-2013	Plant primary production on Herschel Island	-
4-May-2015	Breeding activity of passerine birds on Herschel Island	-
6-Feb-2015	Weasel population monitoring on Herschel Island	-
5-Mar-2013	Hershel Island Snow Depth and Snow Cover	-
12-Mar-2015	Shorebird surveys conducted to determine habitat types in Queen Elizabeth Islands, Canadian Arctic Archipelago	-
31-Mar-2015	Monitoring of the abundance and activity of Arctic Foxes in the Hudson Bay Lowlands	-
12-May-2014	ArcticNet - Arctic SOLAS 0706b - Northwest Passage CTD data	-
4-May-2015	Monitoring of breeding activity of Long-tailed Jaegers at Alert, Nunavut	-
31-Mar-2015	Long-term monitoring of plant net primary production, flowering times and goose grazing impact on vegetation of the Cape Churchill Peninsula, Manitoba	-

Future of RDM and CC

- We have realized that RDM is a rich topic: technology, policy, education. We have changed our approach.
- A service provider like CC only provides part of the picture.
- Working with CARL, Globus to design RDM production system. Federated repositories, global discovery, scalable, etc.).
- Build RDM system on top of new national storage architecture.
- Build flexible system based on robust tools at the level of PBs at 2-3 locations. 2 year project (beta testers sooner).
- Training - national partnership with Software Carpentry is coming down the road. The emerging RDM community need to be trained.
- We need your input! There will be a SPARC 2 consultation coming soon. Please, tell us what you think and how you would like CC to evolve with respect to RDM.

Conclusions

- Researchers and Compute Canada do not work in isolation. Strong research collaborations exist between many community (EGI, Globus, CARL, etc.).
- It is crucial that we adopt open, interoperable solutions.
- We must develop solutions which lower the barrier to proper data management and sharing.
- While enabling sharing, we must preserve data privacy.
- It is crucial that the data not be merely stored in a virtual warehouse - it must be closely connected to significant computational resources.
- Compute Canada is very open to collaboration and to receive input from your community.