computecanada

Compute Canada-CARL-Globus
Canadian National Research Data Repository

April 5, 2016

Compute Canada – CARL

- **Compute Canada**: Canadian Advanced Research Computing organization
- **CARL**: Canadian Association of Research Libraries

- In November 2015, CC and CARL signed a 2-year MOA to collaborate on the development of an at-scale repository and preservation technology framework

- CC+CARL struck a 2-year development project (Jan 1/2016 to Dec 31/2017)
  - build a tool that could be used to run a national research data repository (NRDR) service for Canada

Compute Canada – Globus

- **Globus**: Non-profit RDM organization out of UChicago/ Argonne

- Spring 2016: CC and Globus collaborate on development for NRDR project

- CC development team and Globus are working closely together to jointly extend functionality of Globus Publishing

- Globus licensing current version of Globus Publishing to CC

Canadian NRDR: Goals

- Provide a platform for the "long tail" of Canadian researchers to preserve their data, allow discovery and reuse

- Integrate existing institutional and domain-specific data repositories

- Build partnerships that span the country's library and IT communities to work at national scale

Canadian NRDR:
Features of Planned Solution

- Federated Storage Model
  - Not monolithic; storage and repositories can be geographically distributed, and owned operated by organizations / institutions (as well as CC)

- Scalable
  - Designed to be operated at national scale for Canada
  - Many files; large files; large total amounts of storage (Globus)

- National Data Discovery
  - Single search to discover data, whether in CC-hosted NRDR platform or in one of many existing repositories

Canadian NRDR:
Features of Planned Solution

- Data Preservation Pipeline
  - Integration of archivist tool suite, Archivematica, to normalize / preserve data (AIPs)

- Suitable for broad range of data types

- Automatic geographic data replication
  - Leverages Compute Canada national data infrastructure being deployed in 2016
    - Deployed at all four CC hosting sites based on SBB design
    - Many types of storage services: disk, tape, Object storage, HSM, backup

compute | calcul
canada | canada

Canadian NRDR: Project Timeline

- Development project runs Jan 2016-Dec 2017

- By end of 2017, goal is fully functioning and scalable RDM repository software tools, deployed on CC hardware at modest scale

- First milestone June 2016:

  - Canadian instance of Globus Publishing running on CC hardware

  - Support direct ingestion / upload of researcher data and metadata

  - Support discovery of data at selected institutional repositories

  - Characterize the preservation pipeline improvements required