compute canada

Compute Canada and Globus Data Publication

April 15, 2015

Compute Canada
Federated RDM Pilot

- 2014-15 Pilot brought together CC, Research Data Canada (RDC), Canadian Association of Research Libraries (CARL)
  - CC staff
  - Librarians

- Aim: understand requirements for a future Canadian RDM repository system
  - CC would be heavily involved in hardware/software

- Evaluated a number of data repository technologies
  - Islandora + Archivematica
  - Dataverse + Archivematica (Summer 2015)
  - Globus Publishing

- Islandora, Dataverse, Globus Publishing: repository tools
- Archivematica: Preservation tool

compute canada | calcul canada

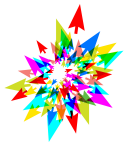
Datasets

- Two different datasets:
  - Social sciences: Pacific Herring
    - Videos: 50 x 50MB files
  - Physics: High Energy Physics Simulations
    - Custom data format: 200GB datasets

- Different use cases, different scales
  - Users of rich media often prefer to view in place
  - Repositories and technologies that work well at the MB-GB scale can have issues growing to TB-PB scale

- Datasets were ingested into (Islandora + Archivematica) and Globus Publishing

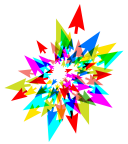
compute canada | calcul canada



Repository tools

- Globus Publishing deployed on Compute Canada resources at Simon Fraser University, University of Toronto
  - Pacific Herring data + metadata ingested at SFU
  - HEP Simulations data + metadata ingested at UofT
  - Discovery from both collections possible through single Globus querying point
- Islandora + Archivematica deployed on CC resources at SFU
- Islandora:
  - Nicely handled and presented small video dataset; in-place viewing at repository (Herring)
  - had technical problems with ingestion, transfer of large, non-media dataset (HEP simulations)
- Globus Publishing:
  - Easily handled both datasets
  - Not many in-place interaction features with data

compute canada | calcul canada



Globus Publishing Impressions

Strengths

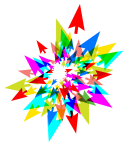
- Discovery process can span multiple collections/repositories
- Almost arbitrarily large datasets (total and per-file size) can be accommodated
- Storage and transfer highly scalable
  - (have not yet tested scalability of metadata discovery)
- Wide variety of data formats can be accommodated
- File transfer protocols for ingest and download are high-performance
- replication to other sites via Globus is high-performance, easy on CC infrastructure

Challenges

- Currently requires Globus intervention to set up metadata form for project
- Files stored in Globus Publishing do not undergo preservation steps
- No ability to view videos, pictures from within Globus Publishing: download or transfer files only
- Configuration/customization of front end done by Globus (only)

Globus is SaaS

compute canada | calcul canada



Replication

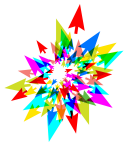
- Many motivators for automated replication of data for repositories
  - fault tolerance
  - Availability
  - Funding agency regulations
  - Librarians like to know things are safe (LOCKSS)

- As part of pilot, demonstrated automatable replication (one-way mirroring) of data
  - Globus API-based python code, shared endpoint
  - Used to mirror data from SFU to Toronto after ingestion into Globus Publishing at SFU

- Potential for extension to Globus-supported tools
  - Globus developing client for data replication, discussions with CC

- Fast/easy/reliable
  - leverages the CC Globus infrastructure we have already built

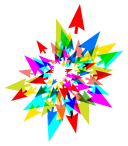
compute canada | calcul canada



Feature Wishlist

- HTTP file access
  - anonymous (no Globus account) download
- API for Globus Publishing
  - customizability of user interface is a driver
- Self-service form configurator
- Ingestion from existing collections with existing metadata
  - bypass manual data entry
- Ingestion in-place without transferring data
- Expansion of download features to get multiple datasets at a time

compute canada | calcul canada



The Future

- Compute Canada sees the scalability inherent in Globus Publishing as promising

- Compute Canada and Globus have just begun 3-month project to integrate Archivematica with Globus
  - Ingesting data and preservation products from existing Canadian Polar Data Network collection
  - Librarians interested in preservation capabilities of Archivematica

compute canada | calcul canada