Dataverse at the University of Alberta

Larry Laliberte and Chuck Humphrey
University of Alberta Libraries
2015
The University of Alberta Context

- Research data management services is part of Digital Initiatives, which also includes the institutional repository, digitization production & collections, digital preservation, and scholarly communications.
- Digital Initiatives has ten librarians, one post-doc, three programmers, three student assistants.
- Support from two metadata librarians, one discovery librarian, and two system administrators.
HYDRA NORTH DAMS PROJECT
2014/15 1.0

CONTENT TYPES | MODELS
- theses
- text
- digitization
- research and spatial data
- multimedia

DIGITAL COLLECTIONS
INDEX | BLACKLIGHT
- University Archives
- Faculty Blogs
- Research Data
- Spatial Data
- Multimedia

COLLECTIONS | METADATA | VIEWS
- Examples Only
- creators and contributors

DATAVERSE

DAMS
- ETDs, books, papers, newspapers
- reports, gov pubs, research data
- images, maps, multimedia, etc...

API

LOCKSS PLNs
Canadian TDR Network
HathiTrust
Internet Archive
Canadian Polar Data Network

NOTES
- DAMS = Digital Asset Management System; connection to preservation storage, i.e. Swift + OpenStack
- Main discovery in Blacklight, not full text index; representation of some materials at collection level, e.g. research data, digitized newspaper title (not full text)
- Content models = e-thesis (multipart), book, journal, article, newspaper (article segmented), image (large, small), audio, video, data set, spatial data, etc.
- DAMS development driven by user stories
- Access to original and derivative files to support accessibility and digital scholarship, e.g. OCR, TIFF/P2
- Archivematica may be used as a processing/packaging tool for certain types of content going into DAMS, e.g. research data, archival
- Utilize digital object identifiers (DOIs) or equivalent persistent IDs, where logical Respect metadata in its native format, whenever possible
- Use OAIharvest digital records to be included at later stage.
Dataverse & Researcher Needs

- Back-up storage for research data.
- An easy way to share data with a research team that does not involve the use of email or Dropbox.
- A way of tracking which dataset was used to produce a table in a publication or of tracking other version issues.
- A way to satisfy a publisher’s requirements, e.g. a DOI for a dataset associated with an article or providing access to a dataset for reviewers.
Dataverse & Curator Needs

- A simplified approach to capture project-level metadata that can be used repeatedly with other applications.
- A one-step method of minting PIDs for data files.
- A service that helps organize data for preservation processing.
Use Cases

- Humanities uptake
  - Linguistic collections
    - Audio recordings of aboriginal elders
    - Audio recordings of aboriginal accents
  - Translation of Spanish texts in the public domain
    - Image collection of texts and current translations
  - Gamegate
    - Social media web-scrapping on gamegate discussions
Use Cases

• Physical Education & Recreation
  o PLOS publication requiring the researcher to share her data for anonymous review

• Canadian Century Research Infrastructure
  o Dissemination of spatial data for the 1911, 1921, 1931, 1941, and 1951 Censuses

• Biology (in need of a quick DOI)
  o Grizzly bear habitat
Next Steps

- Install the new version (4.0) of Dataverse.
- Tailor Dataverse workshops for graduate students.
  - Promote the sharing of thesis data through Dataverse.
  - Provide experience using dataset DOIs with dissertations.
  - Promote the rewards of data sharing.
The Canadian Context

- Research Data Canada (RDC) and the Canadian Association of Research Libraries (CARL) federated pilot program
  - Emerged from the January 28-29, 2014 Digital Infrastructure Summit in Ottawa
    - Prove that existing infrastructure can be coordinated to support (i) research council data management plan policies and (ii) a national research data preservation service
PROJECT OR RESEARCH PROGRAM LEVEL

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 

In-Situ approved Repositories

User knows DOI

User doesn't know DOI

Search Options

Search subset of DataCite DOI Registry

General Search

Search by

DataCite Canada

Custom search app or domain-specific search apps

Repartition

NO Refer to Archive

YES Refer to repo

Repository score?

Deposited in standard format

Archival metadata managed by CASR/ROC

Perform descriptive crosswalk as required

Index Data Package

Data processing

Data is used to provide

Compute Canada-hosted infrastructure

Compute Canada Storage Systems

Funded

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

Create a Standard Data Interoperability Package

Deposit 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

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 (PPR), possible other sources.

Local/Regional Storage Repositories (e.g., ODRL, EIDC-Out. Institutional Storage)

Facilitate re-use on local platforms

Facilitate re-use on local platforms

Local/Regional Compute Facilities

Access/increases data for replication/review; if missing, extracts data back into approved repository.
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 #

Data repository

Project or Research Program Level

Exchange

Data repository level

Data repository level

Compute Canada-hosted infrastructure

Local/Regional Storage Resources (e.g., Compute Canada, institutional storage)

Researchers

Facilitate re-use on local platforms

Local/Regional Compute Facilities

Facilitates re-use on local platforms

Researcher

Access retrieves data for replication/test; if testing, returns data back into approved repository.

Create 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

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 Archinormica Format Policy Registry (FPR), possible other sources.

Funded

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

- Scholars Portal of the Ontario Council of University Libraries is working with IQSS and Artefactual to move studies from Dataverse to Archivematica for the production of Archival Information Packages (AIPs)
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 IDs

DATA REPOSITORY LEVEL

- Hydra Repos
- Valentina VREs
- OAI Archives
- CASIRNet
- CAOIC

Network Level

EXCHANGE

- Website is used to
- Index Data Package in
- Archivematica
- Interoperability
- Compute Canada Storage Systems
- Compute Canada-hosted infrastructure
- Compute Canada HDF5 Facilities
- Local/Regional Storage
- Resources
- e.g., ODIN, Eedi/Out. Institutional Storage
- Local/Regional Compute Facilities

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 crosswalks

A 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 (PPR), possible other sources.
Next Steps

- Complete the RDC-CARL Federated Pilot with the Dataverse-Archivematica software stack.
- Find ways for Canadian Dataverse Networks to collaborate.