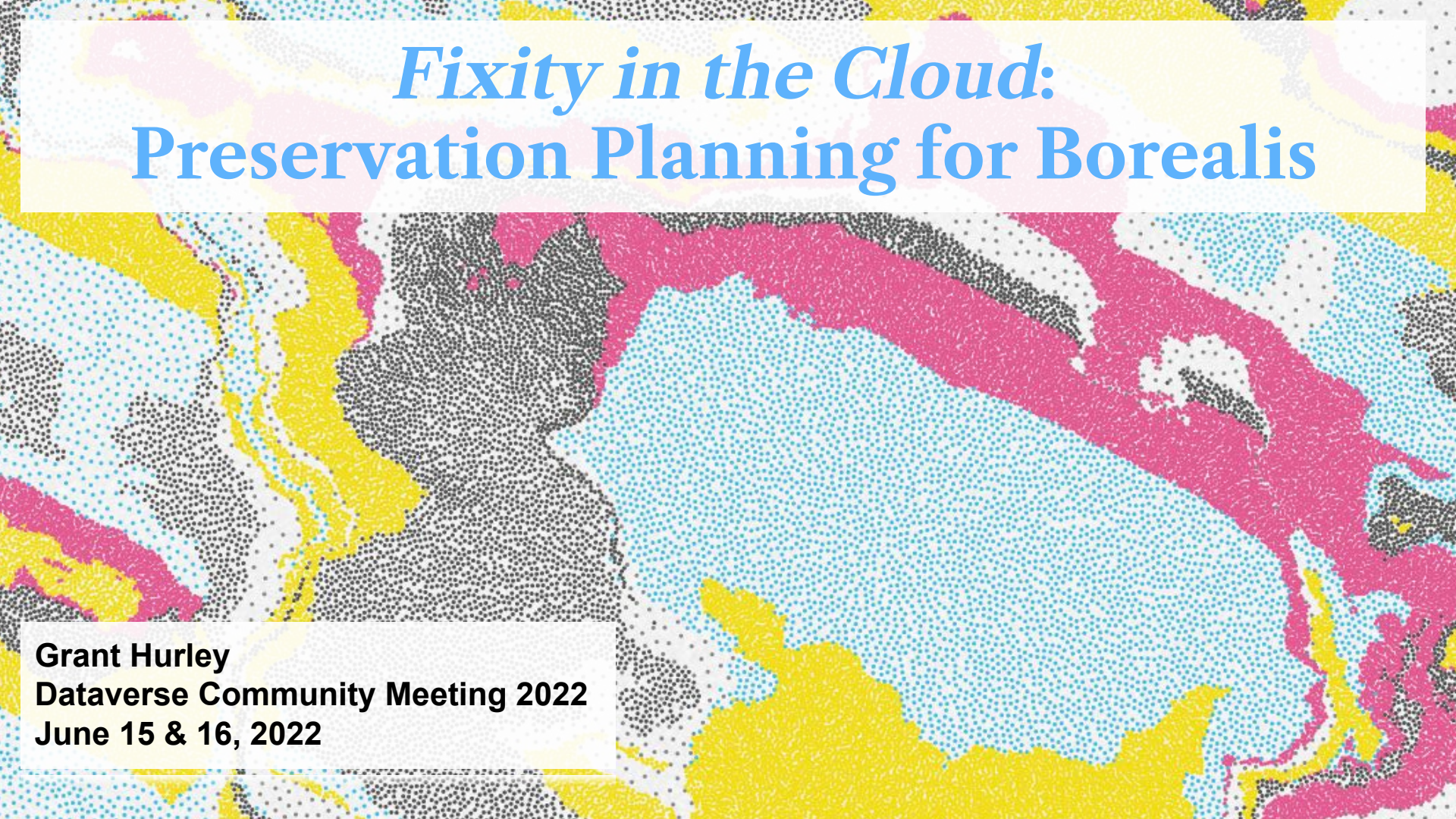
Fixity in the Cloud: Preservation Planning for Borealis



Grant Hurley
Dataverse Community Meeting 2022
June 15 & 16, 2022

Digital preservation happens on Indigenous lands

I am speaking from Tkaronto (Toronto), the traditional territory of many nations including the Mississaugas of the Credit, the Anishnabeg, the Chippewa, the Haudenosaunee and the Wendat peoples.

Credit: Ogimaa Mikana

A new identity for Scholars Portal Dataverse

borealis

The Canadian Dataverse Repository
Le dépôt Dataverse canadien

About Borealis

- Dataverse instance installed by Scholars Portal in 2012
- More focused development for academic libraries in Ontario began in 2015
- Expanded to institutions in Québec in 2019 and nationally in 2019-20
- Launching as *Borealis, the Canadian Dataverse Repository* on June 23, 2022
- Today the repository hosts over 8,000 datasets held by over 60 member institutions

Designing a preservation approach

Needs:

- Basic, ongoing integrity assurance for *all* files uploaded to the repository
- Scalable, cost-effective, automated
- Make use of existing preservation functionality in Dataverse software
- Support for more advanced preservation processing and independent package creation via Archivematica integration

Community input:

- Dataverse North Policy Working Group draft preservation plan outline (2020)
- Digital Research Alliance of Canada Preservation Expert Group report *Preservation for Dataverse in Canada: Recommendations Report* (2021)

A useful preservation precursor: Cloud storage

- SP migrated Borealis' file storage to the Ontario Library Research Cloud (OLRC) in August 2021

- Cloud storage provides a baseline for data availability, reliability and scalability
- Replication across multiple geographic locations reduces the risk of a single point of failure
- Cloud service and software manages functions such as security, data ingest and replication, and disk monitoring and maintenance separately

What is the OLRC?

Objective: Provide cost-effective, subscription-based, scalable and reliable storage for libraries and archives to house and support access to digital collections

Background: Developed with project partners in the Ontario Council of University Libraries between 2012-2015, in production since 2016; available nationally in 2021

OLRC Components

- **OpenStack Swift** object storage architecture
- **5 data centres (“nodes”)** located at academic libraries in Ontario:
  Toronto, York, Guelph, Queen’s & Ottawa
- ORION and GTAnet **high speed private network** to connect them

OLRC Features

Availability: 3 copies are replicated across 5 nodes to ensure files are always accessible

Reliability: Files are internally checked for integrity, and if a copy is found to be corrupted, is automatically replaced from a good copy in another location

Flexibility: Variety of methods to integrate with, and access, the OLRC via Horizon and DuraCloud interfaces, command line interface (CLI), and S3 API

Why does cloud storage not equal preservation?

- “Replication” does not mean *independent* copies - for preservation purposes, this is still considered 1 complete copy
  - More than one complete copy is a minimum for preservation assurance
  - See: NDSA Levels of Preservation
- Internal fixity checking is a feature, but the storage network will accept corrupted data
  - For example, a file is corrupted during upload or transfer
  - A second, independent source of integrity verification is required

Preservation Plan: Objectives

1. Ensure a minimum level of fixity assurance for all files uploaded to the repository by registered users
2. Store files using a secure, reliable and scalable preservation storage strategy
3. Install and maintain all preservation features that are core to the Dataverse application, resulting in selected preservation metadata and format conversion for tabular data uploads

   [Together = Level 1 preservation strategy]

4. Support Participating Institutions who wish to export independent packages of selected dataset files and metadata from institutional collections in Dataverse

   [= Level 2 preservation strategy]

Preservation Plan: Level 1

Fixity checks for all user-uploaded files are conducted every 30 days:

- Includes draft and restricted datasets; does not include derivatives, thumbnails, etc. created by Dataverse
- When users upload files, MD5 checksums are automatically generated and stored in the Dataverse database
- A script uses the Dataverse Native API's Physical Files Validation in a Dataset API call, which downloads each file from storage, computes its checksum and compares with the value stored in the database
- The record of each fixity check (both positive and negative) is stored in an internal MySQL database

Preservation Plan: Level 1

- Storage strategy
  - All files are stored in the OLRC
  - Files are backed up to IBM TSM tape nightly
  - Up to 7 versions of files are retained in backup for 30 days

- Key preservation-supporting functions in Dataverse:
  - File format identification using JHOVE
  - Transformation of tabular data formats into non-proprietary tabular text data files (.tab) upon ingest
  - Generation of UNFs (Universal Numeric Fingerprints) for tabular data files

Fixity error triage

1. Any fixity errors identified through monthly report
    - Examples: Empty (0-byte) file, missing file, checksum does not match stored value
2. Compare with copy in backup
3. If good copy from backup resolves issue, replace bad file(s) with good versions on file system
4. If issue not resolved via backup, contact depositor(s) and institutional contact for file replacement
    - This would usually indicate a failed upload

Preservation Plan: Level 2

- Scholars Portal will assist membership in the setup of connections to Archivematica instances or running BagIt exports
- Participating Institutions are responsible for determining which datasets are eligible for additional processing and export based on *appraisal* criteria
- See the guide *Appraisal Guidance for the Preservation of Research Data*

Dataverse + Archivematica

- Requires access to an Archivematica instance (hosted by service provider or locally installed)
- Enables creation of independent preservation packages to be stored in a completely separate preservation storage location
- Performs signature-based file format identification, file format validation, characterization, and normalization
- Archivematica validates Dataverse checksums, records tabular normalization events, transfers some descriptive metadata
- Automates creation of PREMIS-formatted digital preservation metadata
- Most important: requires some level of appraisal for preservation to decide which datasets to process
- Read more: project wiki, iPRES paper, DV 2021 Community Meeting slides & recording

Thank you!

Up next: preservation plan to be published at borealisdata.ca on June 23

Questions? dataverse@scholarsportal.info