AIPing Research Data: Integrating Dataverse and Archivematica

Grant Hurley (Digital Preservation Librarian, Scholars Portal)
Dataverse Community Meeting 2021
June 17, 2021

Land Acknowledgement

I would like to begin by acknowledging that the land I am speaking from is the traditional territory of many nations including the Mississaugas of the Credit, the Anishnabeg, the Chippewa, the Haudenosaunee and the Wendat peoples and is now home to many diverse First Nations, Inuit and Métis peoples. Toronto is covered by Treaty 13 signed with the Mississaugas of the Credit, and the Williams Treaties signed with multiple Mississaugas and Chippewa bands.

I am grateful to pursue my life and livelihood on these lands, where I have lived for 6 years. I originally hail from unceded Passamaquoddy territory on the shores of the Bay of Fundy in the province now known as New Brunswick.

Credit: Ogimaa Mikana

Outline

- Project Context
- Integration Development
- Assumptions, Workflow & Functionality
- Demo
- Looking Ahead

Context

- Scholars Portal is the information technology service provider for the Ontario Council of University Libraries (OCUL), a consortium of 21 academic libraries in Ontario, Canada
- Scholars Portal was established in 2002 as a way to implement OCUL's planning and direction
- We are a unit of the University of Toronto Libraries
- Our role is to help our members meet the needs of their students, faculty, researchers, and other stakeholders in the provision of shared technology services

OCUL Ontario Council of University Libraries

Scholars Portal
A Service of the Ontario Council of University Libraries
Un service du Conseil des bibliothèques universitaires de l'Ontario

Context

- Dataverse instance installed at Scholars Portal in 2012
- More focused development for OCUL members began in 2015
- Expanded to institutions in Québec in 2019
- Today: 59 institutions across Canada use Scholars Portal Dataverse
- Connection with NDRIO/Portage Network via Dataverse North advisory group + many Dataverse-supporting staff who serve members on Portage Expert Groups
- New consortium recently created to organize national governance of the service

Integration Development

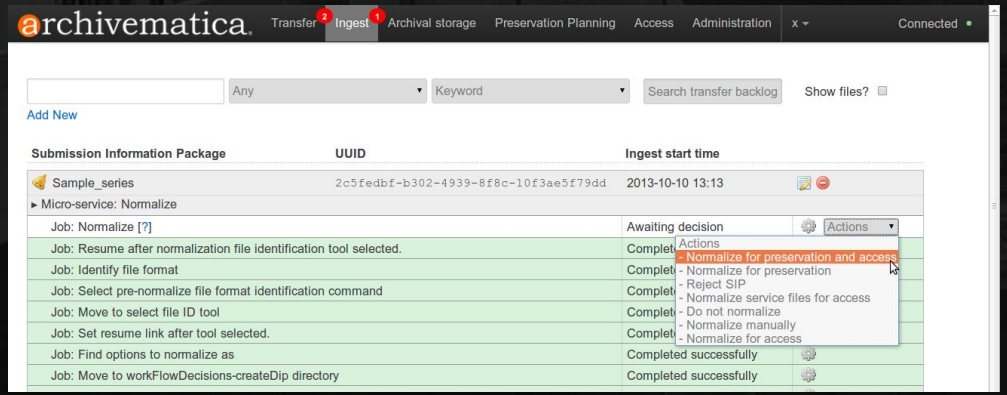
- Motivated by desire to enable functional workflows for dataset preservation between Dataverse and Archivematica

- OCUL sponsored integration project with Artefactual Systems Inc.
  - Phase 1 - Proof-of-Concept (2015)
  - Phase 2 - Public release in Archivematica v. 1.8 (2018)

- Archivematica can be configured to use a connected Dataverse instance as a transfer source location as of v. 1.8 + some fixes in 1.9!

What is Archivematica?

- Open-source, standards-based workflow tool for processing digital objects for preservation and access
- Configurable workflow based on series of microservices, including:
  - Checksum generation and verification
  - File format identification, characterization, and validation
  - Normalization (generate preservation and/or access copies) … and more!
- Generates Archival Information Package (AIP) for safe preservation storage and Dissemination Information Package (DIP) for access



Integration Assumptions

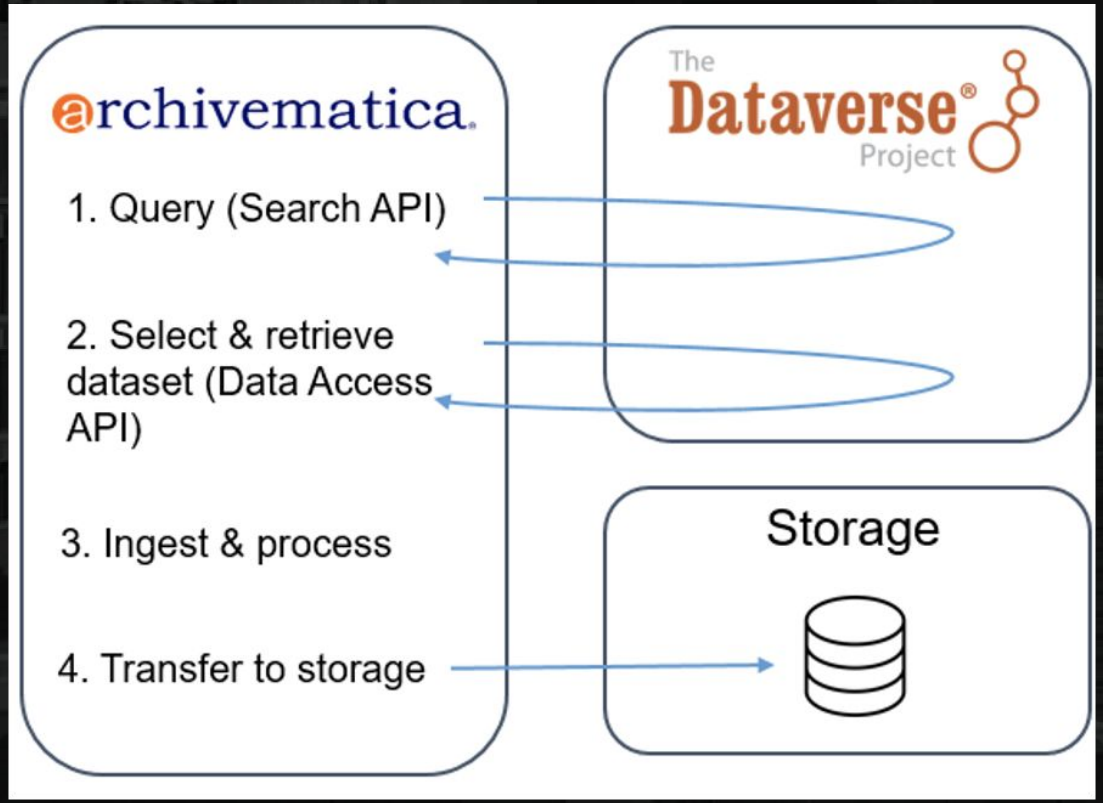
The Preserver has:

- A Dataverse account and an API token
- Access to an Archivematica instance
- Obtained necessary rights and privileges to process and store dataset files independently of Dataverse
- Is interested in selecting and processing specific datasets in the target Dataverse instance for preservation

Other assumptions:

- 1 Dataverse dataset = 1 Archival Package (AIP)
- The access copy (Dissemination Information Package or DIP) is presumed to remain the copy available in Dataverse, though users can also choose to create new DIPs in Archivematica

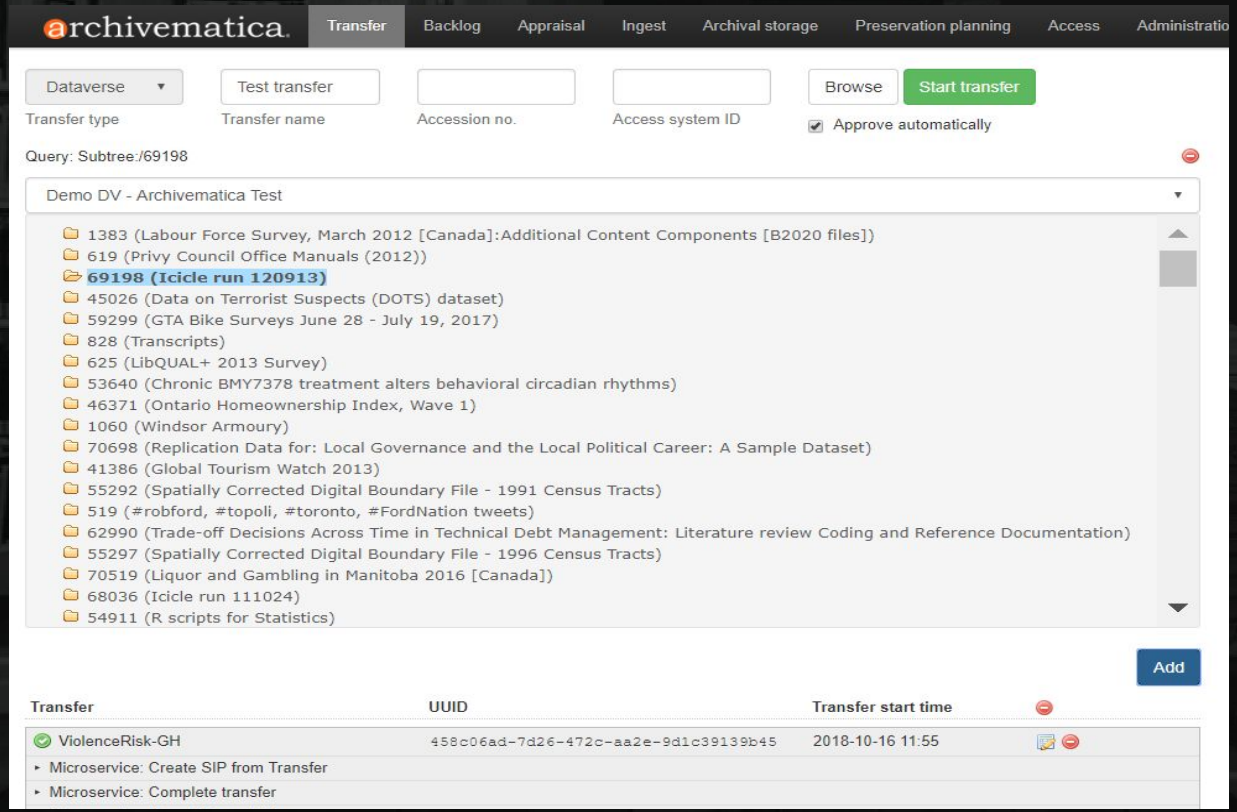
Workflow


archivematica
The Dataverse Project
1. Query (Search API)
2. Select & retrieve dataset (Data Access API)
3. Ingest & process
4. Transfer to storage
Storage

Overview of Functionality

1	Dataverse Access API retrieves package	• Original user-submitted data files • Dataset-related metadata (dataset.json file) • Dataverse-created derivatives of tabular files + citation metadata (if applicable)
2	Dataverse transfer type preconfigured steps	• Creates Dataverse METS file for original package • Checksums verified using Dataverse MD5s • Tabular generation recorded as PREMIS event; Dataverse instance linked as the agent • Other standard microservices as configured
3	Archivematica creates METS	• Includes dataset descriptive metadata (DDI format) • Outlines relationships of originals, metadata files & derivatives and actions undertaken through processing steps

Demo!


archivematica
Transfer
Backlog
Appraisal
Ingest
Archival storage
Preservation planning
Access
Administratio
Dataverse
Transfer type
Test transfer
Transfer name
Accession no.
Access system ID
Browse
Start transfer
Approve automatically
Query: Subtree:/69198
Demo DV - Archivematica Test
1383 (Labour Force Survey, March 2012 [Canada]:Additional Content Components [B2020 files])
619 (Privy Council Office Manuals (2012))
69198 (Icicle run 120913)
45026 (Data on Terrorist Suspects (DOTS) dataset)
59299 (GTA Bike Surveys June 28 - July 19, 2017)
828 (Transcripts)
625 (LibQUAL+ 2013 Survey)
53640 (Chronic BMY7378 treatment alters behavioral circadian rhythms)
46371 (Ontario Homeownership Index, Wave 1)
1060 (Windsor Armoury)
70698 (Replication Data for: Local Governance and the Local Political Career: A Sample Dataset)
41386 (Global Tourism Watch 2013)
55292 (Spatially Corrected Digital Boundary File - 1991 Census Tracts)
519 (#robford, #topoli, #toronto, #FordNation tweets)
62990 (Trade-off Decisions Across Time in Technical Debt Management: Literature review Coding and Reference Documentation)
55297 (Spatially Corrected Digital Boundary File - 1996 Census Tracts)
70519 (Liquor and Gambling in Manitoba 2016 [Canada])
68036 (Icicle run 111024)
54911 (R scripts for Statistics)
Add
Transfer
UUID
Transfer start time
ViolenceRisk-GH
458c06ad-7d26-472c-aa2e-9d1c39139b45
2018-10-16 11:55
Microservice: Create SIP from Transfer
Microservice: Complete transfer

Post-Integration Developments

- Some unresolved bugs with Dataverse METS file creation files documented in the Archivematica Issues repository
- Take up has been slow - many institutions:
  - Have not yet prioritized data preservation as a target for resources and staffing
  - Have not yet delineated responsibilities for research data preservation within their organizations
  - Need policies for collections development and digital preservation and procedures for curation and appraisal/selection

Post-Integration Developments

- CoreTrustSeal certification cohort via NDRIO/Portage pushing policy and procedure creation and documentation forward at SP and at member institutions
- Portage Preservation Expert Group just released *Preservation for Dataverse in Canada: Recommendations Report*
  - Outlines two-tier approach:
    - First level: Bit-level checks for all SP-hosted Dataverse data
    - Second level: Access to Archivematica services for dataset preservation
  - Advocates for funding for policy development, storage, Archivematica integration improvements, training

Additional Improvements?

- Messaging back to Dataverse
- Automation
  - From Dataverse, e.g. “push to preservation” button
  - For groups of datasets (from Archivematica end)
- Use of DIP? - see Archidora
- Cross-AIP reporting - see AIPScan

Additional Reading

Dataverse-Archivematica wiki page

Archivematica documentation

iPRES paper / DPC blog post

Acknowledgements!

- OCUL - for funding the integration!
- Artefactual Systems - for developing it!
- Advice and support: Allan Bell, Eugene Barsky, Peter Binkley, Eleni Castro, Alan Darnell, Kate Davis, Philip Durbin, Alex Garnett, Geoff Harder, Chuck Humphrey, Larry Laliberte, Amber Leahey, Victoria Lubitch, Steve Marks, Evelyn McLellan, Umar Qasim, Joel Simpson, Ross Spencer, Amaz Taufique, Leanne Trimble, and Dawas Zaidi