

**Dataverse + Archivematica = **

** = research data preservation**

A lightning talk and lil demo for TAATU 2020  
by Grant Hurley (Digital Preservation Librarian, Scholars Portal)

# 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 5 years. I originally hail from unceded Passamaquoddy territory on the shores of the Bay of Fundy in the province now known as New Brunswick.

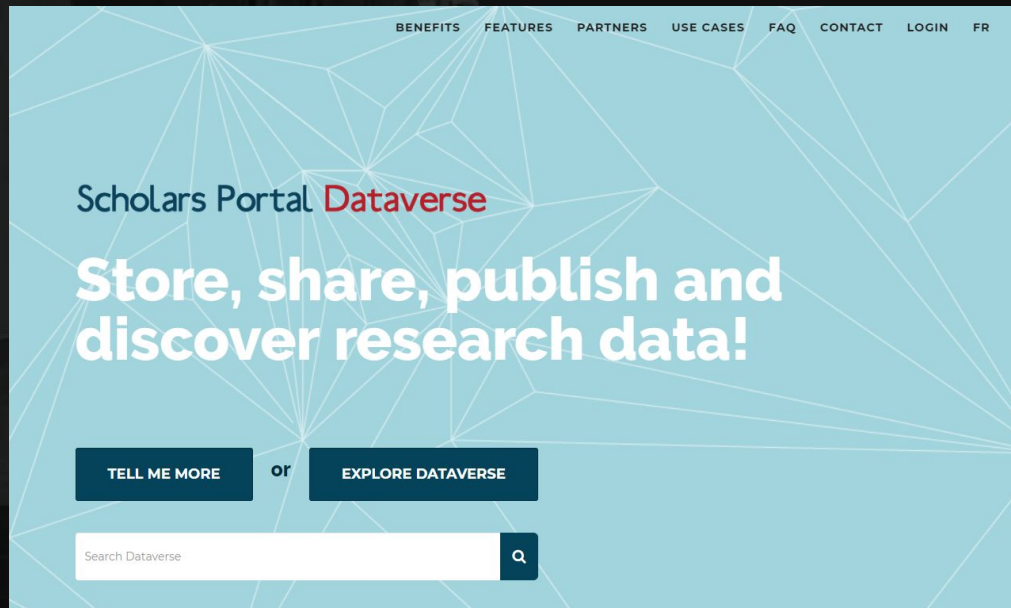
# What is research data?

A *genre* of digital objects better defined by the context of their creation than form:

- Outputs resulting from research activities, whether scholarly, technical or artistic.
- Constitute the source records/evidence of a research activity: observations, facts, measurements.
- Not the published products/interpretations of these primary source records.

# What is Dataverse?

- Open-source repository platform to store, share, publish and discover research data
- Developed and maintained since 2006 by the Institute of Quantitative Social Science (IQSS) at Harvard University
- 60+ installations around the world
- Developed with quantitative social science data in mind, but not limited to any domain
- Automatic data citation with DOI



# Topical examples:

## Black Experience Project

Version 1.1

Black Experience Project, 2018, "Black Experience Project", <https://doi.org/10.5683/SP2/A6MAMR>, Scholars Portal Dataverse, V1, UNF:6:+ilxaW4S0kyeQhiilCDN7WA== [fileUNF]

Cite Dataset [Learn about Data Citation Standards.](#)



**Description** Data (2017-07-01)  
**Subject** Social Sciences; Other  
**Keyword** African Canadian, Black, Greater Toronto Area

Files Metadata Terms Versions

Search this dataset...

Filter by  
File Type: All Access: All File Tag: All

1 to 5 of 5 Files

<input type="checkbox"/>	 <b>BEP Detailed Data Tables - v22 July 11-2017.pdf</b> Adobe PDF - 10.8 MB - Nov 7, 2018 - 1 Download MD5: ac28423c665e9683841b5eadd5f5cd57 Detailed Data Tables <input type="button" value="Data"/>	<input type="button" value="Download"/>
<input type="checkbox"/>	 <b>BEP2017.xls</b> MS Excel Spreadsheet - 24.5 MB - Nov 7, 2018 - 1 Download MD5: 7448bf712d8f532348e177e6f8a2fbd6 <input type="button" value="Data"/>	<input type="button" value="Request Access"/>



[Survey data](#) from the *Black Experience Project in the GTA* (York University Dataverse)  
- [see the report](#)



# Topical examples:

## Replication Data for: Measuring and Comparing Municipal Policy Responses to COVID-19

Version 1.0

Armstrong, David A. II; Lucas, Jack, 2020, "Replication Data for: Measuring and Comparing Municipal Policy Responses to COVID-19", <https://doi.org/10.5683/SP2/MXWJAZ>, Scholars Portal Dataverse, V1, UNF:6:sH1HjqRme9p/+1p4x3JsQ== [fileUNF]

[Cite Dataset](#) [Learn about Data Citation Standards.](#)

**Description** ? Replication code and data for CJPS COVID-19 research note, "Measuring and Comparing Municipal Policy Responses to COVID-19". (2020-05-03)

**Subject** ? Social Sciences


**Keyword** ? COVID-19, coronavirus, municipal policy, municipal politics, Canadian policy, Canadian politics

Files Metadata Terms Versions

Search this dataset... [Find](#)

Filter by  
File Type: All Access: All [Sort](#)

1 to 3 of 3 Files [Download](#)






<input type="checkbox"/>	 <a href="#">armstrong_lucas_replication.rda</a> Gzip Archive - 20.2 KB - May 6, 2020 - 4 Downloads MD5: e585be3afc01ba2666389f029fcb6384	<a href="#">Download</a>
--------------------------	--	--------------------------

[Survey data](#) (University of Calgary Dataverse) + associated paper in the [Canadian Journal of Political Science](#)

# What is Dataverse?

## Preservation-friendly features in Dataverse:

- User-friendly upload of data files, codebooks, documentation, and metadata
- File format identification
- File verification - MD5 checksums
- Tabular data transformation processes
  - Converts variety of formats (SPSS, Strata, RData, CSV, Excel) to TAB text format
  - Tabular files receive UNF checksums to verify semantic content of derivatives

<input type="checkbox"/>	 <a href="#">LaurelCreek_Invertebrate_Water_Spring2017.csv</a> Comma Separated Values - 4.9 KB - Jan 9, 2018 - 0 Downloads MD5: 08a302dd8603f5faf71c90485f96fde3
<input type="checkbox"/>	 <a href="#">LaurelCreek_Invertebrate_Water_Winter2017.csv</a> Comma Separated Values - 3.2 KB - Jan 9, 2018 - 0 Downloads MD5: 899ff2e8f35326b824b06ab7e43e2595
<input type="checkbox"/>	 <a href="#">LaurelCreek_Map.tif</a> TIFF Image - 2.4 MB - Jan 9, 2018 - 1 Download MD5: 2b8a158126afb8d346ebb0fc9853768d
<input type="checkbox"/>	 <a href="#">LaurelCreek_SiteFeatures_2017.csv</a> Comma Separated Values - 11.8 KB - Jan 9, 2018 - 0 Downloads MD5: c8c0440a9e2864b15b017a850f654998
<input type="checkbox"/>	 <a href="#">README_LaurelCreek_Water_Invertebrate_2017</a> Plain Text - 2.4 KB - Jan 9, 2018 - 0 Downloads MD5: 2b18152570ef8dca95398cac19438e34

# 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) and Dissemination Information Package (DIP)

The screenshot displays the Archivematica web interface. At the top, the navigation bar includes 'archivematica', 'Transfer', 'Ingest', 'Archival storage', 'Preservation Planning', 'Access', 'Administration', and 'Connected'. Below the navigation bar is a search area with a search box, a dropdown menu set to 'Any', a 'Keyword' dropdown, a 'Search transfer backlog' button, and a 'Show files?' checkbox. The main content area shows a table of submission information. The table has columns for 'Submission Information Package', 'UUID', and 'Ingest start time'. The first row shows a submission named 'Sample\_series' with UUID '2c5fedbf-b302-4939-8f8c-10f3ae5f79dd' and an ingest start time of '2013-10-10 13:13'. Below the table, a 'Micro-service: Normalize' section is expanded, showing a list of jobs. The first job is 'Job: Normalize [?]' with a status of 'Awaiting decision'. A dropdown menu is open over the 'Awaiting decision' status, showing options: 'Actions', 'Normalize for preservation and access', 'Normalize for preservation', 'Reject SIP', 'Normalize service files for access', 'Do not normalize', 'Normalize manually', and 'Normalize for access'. The subsequent jobs in the list are all marked as 'Completed' or 'Completed successfully'.

Submission Information Package	UUID	Ingest start time
Sample_series	2c5fedbf-b302-4939-8f8c-10f3ae5f79dd	2013-10-10 13:13
Micro-service: Normalize		
Job: Normalize [?]		Awaiting decision
Job: Resume after normalization file identification tool selected.		Completed
Job: Identify file format		Completed
Job: Select pre-normalize file format identification command		Completed
Job: Move to select file ID tool		Completed
Job: Set resume link after tool selected.		Completed
Job: Find options to normalize as		Completed successfully
Job: Move to workFlowDecisions-createDip directory		Completed successfully



# Integration Context

- Dataverse service began at Scholars Portal in 2012
- More focused development began in 2015
- Now: 40 institutions across Canada use SP's Dataverse instance, work with Portage to establish SP Dataverse as national repository
- Growing interest in research data preservation both within OCUL but also nationally, internationally
- Initial aim of project was to investigate how Dataverse datasets could be processed into AIPs
- 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!

# Demo demo!

archivematica. Transfer Backlog Appraisal Ingest Archival storage Preservation planning Access Administratio

Dataverse  Test transfer    Browse

Transfer type Transfer name Accession no. Access system ID  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)

Transfer	UUID	Transfer start time
<input checked="" type="checkbox"/> ViolenceRisk-GH	458c06ad-7d26-472c-aa2e-9d1c39139b45	2018-10-16 11:55
▸ Microservice: Create SIP from Transfer		
▸ Microservice: Complete transfer		

# Technical Questions

Improvements to functionality:

- Messaging back to Dataverse
- Automation
- Interface search and browse
- Use of DIP?

[Issues](#) re: Dataverse METS

# Policy/Governance Questions

- Appraisal/selection for preservation + curation workflows
- File format policies
- Shared preservation infrastructures

# Resources!

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