

A Metadata-Driven Data Curation Tool

Amber Leahey & Victoria Lubitch, Scholars Portal
NADDI 2019, Ottawa (Statistics Canada)



Background

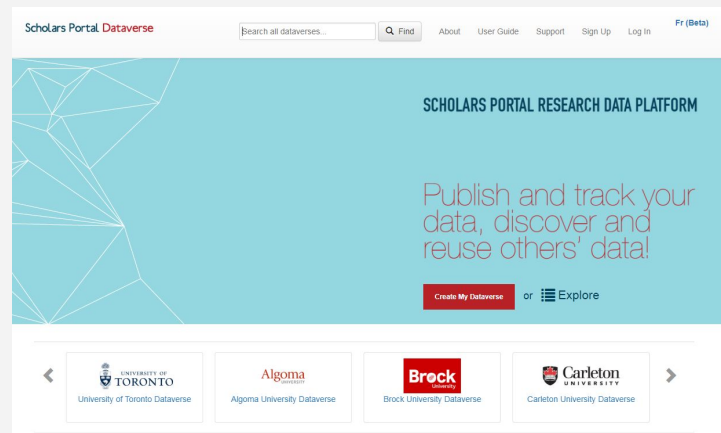
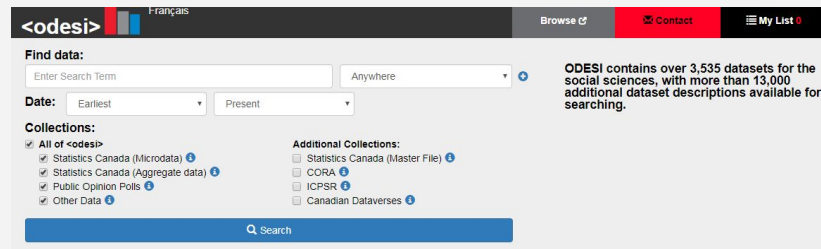
- Scholars Portal (University of Toronto) is a digital library project, providing services to Ontario University Libraries and others across Canada
- SP hosts and manages digital data + scholarly collections (journals, ebooks, geo and statistical data, research data)



SP Data Services

- In 2008, SP launched <odesi> a social science survey data repository
- In 2011, SP launched Scholars GeoPortal
- In 2012, SP launched Dataverse (open-source from Harvard University) for Ontario libraries and researchers

Would like to bring the same level of curation support to Dataverse for research data...



Data Curation

- In archives and libraries data curation involves managing and enhancing data with access and reuse in mind.
- Curation services can focus on enhancing access to data and documentation (publishing in open repositories), metadata creation for discovery and reuse, and data quality review.

A curators toolkit can include a variety of tools:

- ❖ Data Management Plans
- ❖ Storage tools (Google, AWS, local file drive)
- ❖ Open Science Framework
- ❖ Survey and other research administration tools (Blaise, SurveyMonkey, Qualtrics)
- ❖ Data analysis software + tools (SPSS, SAS, STATA, R, Python)
- ❖ Code review + analysis (Github, R, Python, Jupyter Notebooks)
- ❖ DDI metadata tools (Nesstar, StatTransfer, Colectica, in-house solutions)
- ❖ Data repository/publishing tools (Nesstar, Dataverse, Colectica, in-house solutions)



Who are data curators?

- Archivists and librarians
- Research support staff
- Researchers
- Research managers
- Research assistants
- Data analysts

Really isn't one job title for data curators!

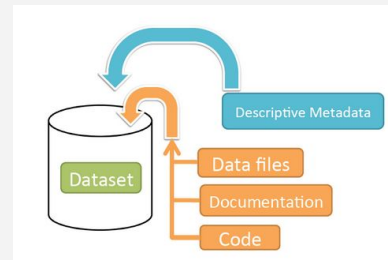
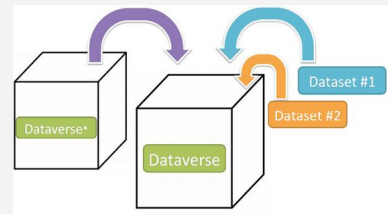
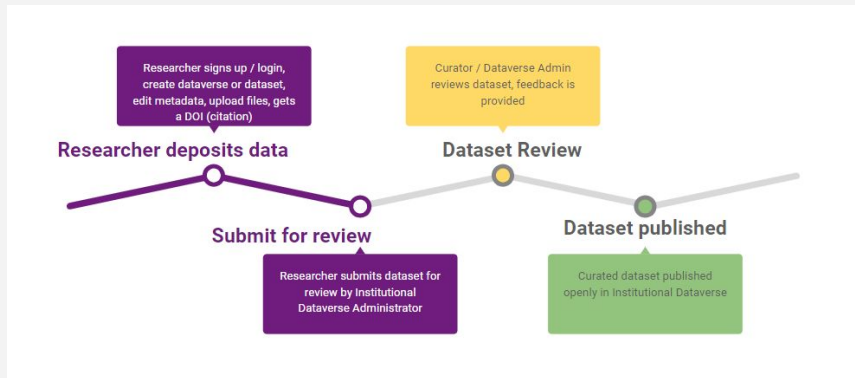


DDI & Data Curation

- DDI supports description of data at the study (project) level and variable / element levels (and pretty much everything in between!)
 - Study methodology
 - Sampling
 - Study concepts
 - Questionnaire procedures & logic
 - Questions
 - Variables (measures)
 - Values
 - Missing values
 - Universes
 - Weights
 - Summary statistics
 - Coding / analysis
 - Software used
 - Etc.



What is Dataverse?



- Open-source repository developed by Harvard University's Institute for Quantitative Social Science (IQSS)
- Offers flexible data deposit and access models (self-deposit, mediated, curated, restricted)
- Supports general and disciplinary metadata standards (Dublin Core, Social Science (DDI), Health, Life Science, Astronomy)
- API-driven
- Modular application support



SP Dataverse

- Multi-institutional installation (over 40 universities in Canada)
- Varied local support models (open and mediated deposit)
- Formal data curation is limited, often self-deposit driven (on part of researchers)
- Limited metadata provided



Example 1

Black Experience Project Dataverse (York University)

JAC

Scholars Portal Dataverse > York University Dataverse > Black Experience Project Dataverse > **Black Experience Project**



Metrics

2 Downloads

Contact Share



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:+ilxaW4S0kyeQhilCDN7WA==

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](#)



Example 2

[Scholars Portal Dataverse](#) > [University of Toronto Dataverse](#) > [Debra Wunch Dataverse](#) > **GTA Bike Surveys August 10-18, 2017**

 Metrics

85 Downloads

 Contact  Share



GTA Bike Surveys August 10-18, 2017 Version 1.0

Wunch, Debra; Arrowsmith, Colin; Heerah, Sajjan, 2017, "GTA Bike Surveys August 10-18, 2017",
<https://doi.org/10.5683/SP/6WJXRX>, Scholars Portal Dataverse, V1, UNF:6:fjyazNAvCI4i4y5XaYV1Vg==

 Cite Dataset ▼

 Learn about [Data Citation Standards](#).

Description

These are preliminary, uncalibrated, synchronized data from the University of Toronto LGR multigas analyser and Airmar weather station while transported in a bike cargo trailer. These measurements were taken throughout the GTA on 3 surveys: August 10, 15, and 18, 2017. (2017-08-31)

Subject

Earth and Environmental Sciences; Physics

Keyword

greenhouse gases, methane, carbon dioxide, carbon monoxide, water vapour

Files

Metadata

Terms

Versions



Example 3

I-Rep Canada Poll 2007

Ipsos Canada, 2017, "I-Rep Canada Poll 2007", <https://doi.org/10.5683/SP/P2VFJR>, Scholars Portal Dataverse, V1, UNF:6:mxzDqEjfg4jSF6aWUcpXQQ==

Q		849 Results	Download
162158	q2_9	2.9) Please indicate how well you feel you know: MasterCard ?	5
162561	q2_10	2.10) Please indicate how well you feel you know: Petro-Canada ?	5
162335	q2_11	2.11) Please indicate how well you feel you know: Purolator Courier ?	5
162389	q2_12	2.12) Please indicate how well you feel you know: Shell ?	5
162625	q2_13	2.13) Please indicate how well you feel you know: Tim Hortons ?	5
162208	q2_14	2.14) Please indicate how well you feel you know: UPS (United Parcel Service) ?	5
161966	q2_15	2.15) Please indicate how well you feel you know: Visa ?	5

First

«

3

4

5

6

7

»

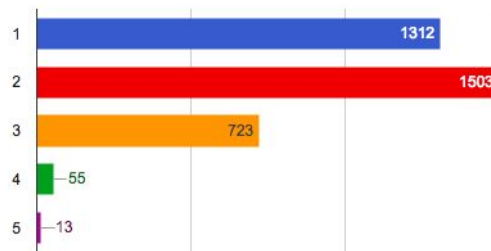
Last

Records Per Page 10

Chart View

Table View

Variable q2_13: 2.13) Please indicate how well you feel you know: Tim Hortons ?



Values	Categories	N
1	1	1312
2	2	1503
3	3	723



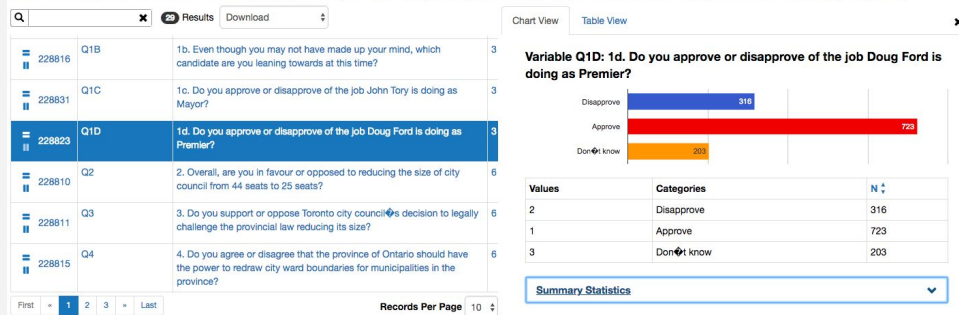
Modular Applications / Integrations

- Modular configuration of the Dataverse code + APIs
- Allows for community sourced development
 - Data Explorer (right)
 - Metrics
 - Archivematica integration
 - Open Science Framework
 - TwoRavens
 - Open Journals System
 - Code Ocean
 - And more

Forum Research Political Poll - Municipal Issues (Toronto) 2018

Toronto - August 2018 - Dataset.tab

Forum Research Inc, 2018, "Forum Research Political Poll - Municipal Issues (Toronto) 2018", <https://doi.org/10.5683/SP2/QCPM89>, Scholars Portal Dataverse, V3, UNF:6:9LBIMLYvRPhQhQyBA== [fileUNF]



Data Curation Tool - DV

Files

Metadata

Terms

Versions

1 File

dct.tab

Tabular Data - 47.2 KB - Apr 17, 2019 - 0 Downloads

3 Variables, 3045 Observations - UNF:6:fbFW653KFIDctLBrYqUA8A==

+ Upload Files

Edit Files ▾

Configure ▾

Download ▾

Data Curation Tool

Data Curation Tool

✕

Data Curation Tool for curation of variables

Continue

Cancel



 Save

Add Group		Search 🔍					
All		ID	Name	Label	Weight	View	
Demography		⋮ - v134615	Q2	2. Do you approve or disapprove of the job Stephen Harper is doing as Prime Minister?		👁️	✎️
		⋮ + v134374	Q1A	1a. If a federal election were held today, which party are you most likely to vote for?		👁️	✎️
		⋮ + v134345	Q1B	1b. Even though you may not have made up your mind, which party are you leaning towards at this time?		👁️	✎️
		⋮ + v134478	Q3	3. Do you approve or disapprove of the job Tom Mulcair is doing as Leader of the Opposition?		👁️	✎️
		⋮ + v134600	Q4	4. Do you approve or disapprove of the job Justin Trudeau is doing as leader of the Liberal Party?		👁️	✎️
		⋮ + v134634	Q5	5. Which party do you expect to win the next federal election?		👁️	✎️
		⋮ + v134338	Q6	6. Do you agree or disagree prostitution should be legal for sex trade workers and their clients?		👁️	✎️
		⋮ + v134522	Q7	7. Parliament recently passed Bill C36 which makes the advertisement and purchase of sexual services illegal. While performing sexual services would still be legal for sex workers, they are prohibited from charging for their services, advertising or emplo		👁️	✎️
		⋮ + v134310	Q8	8. The government has pledged \$20 million dollars to encourage sex workers to leave the sex trade and earn their living in another manner. Do you approve or disapprove of this?		👁️	✎️
		⋮ + v134373	Q9	9. Do you agree or disagree \$20 million dollars is an adequate sum to encourage sex workers to leave the sex trade?		👁️	✎️
<div>Items per page: 10 ▾ 1 - 10 of 39 ⏪ ⏩</div>							

Group Variables

Forum Research Political Poll - Federal Issues (Canada) 2015

 Save


Add Group
































Search



All

 Demographics



	ID	Name	Label	Weight	View	
	 v134374	Q1A	1a. If a federal election were held today, which party are you most likely to vote for?			
	 v134345	Q1B	1b. Even though you may not have made up your mind, which party are you leaning towards at this time?			
	 v134615	Q2	2. Do you approve or disapprove of the job Stephen Harper is doing as Prime Minister?			
	 v134478	Q3	3. Do you approve or disapprove of the job Tom Mulcair is doing as Leader of the Opposition?			
	 v134634	Q5	5. Which party do you expect to win the next federal election?			
	 v134600	Q4	4. Do you approve or disapprove of the job Justin Trudeau is doing as leader of the Liberal Party?			
	 v134338	Q6	6. Do you agree or disagree prostitution should be legal for sex trade workers and their clients?			



Variable Metadata

- Labels
- Question Text
- Interviewer Instructions
- Universe
- Notes
- Variable Group
- Weighting

Variable Information

ID	Name
v46	Var1

Label

gender

Literal Question

Literal Notes

Interviewer Instructions

Interview Instructions

Post Question

Post Question

Universe

Notes

Notes

Group

Group1

Weight Variable

v48

Is Weight

☐



Select Weight Variable(s)

All

Demographics

ID

v134542

v134351

v134575

v134576

v134565

v134436

v134441

v134566

v134333

Variable Information

ID

v134351

Name

D2

Label

D2. How old are you?

Literal Question

Interviewer Instructions

Universe

Notes

Group

Weight Variable

Is Weight

☐

Weight

View

Items per page: 10

31 - 39 of 39

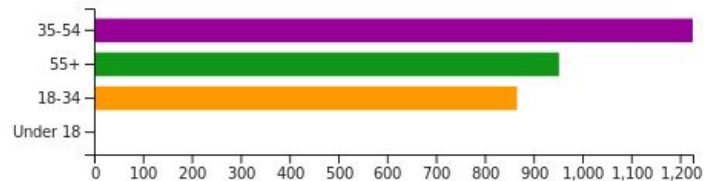
<

>



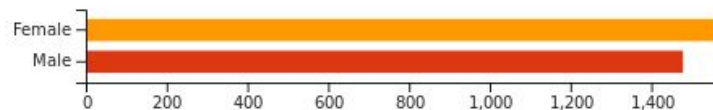
View Frequencies

Var2: age_rollup



Values	Categories	N	NW
1	Under 18	0	0
2	18-34	691	866.4391632769999
3	35-54	1262	1226.3455293050008
4	55+	1092	952.2153135640008

Var1: gender



Values	Categories	N	NW
1	Male	1537	1478.6407127819975
2	Female	1508	1566.3592933639995



Select/Edit Multiple Variables

Forum Research Political Poll - Federal Issues (Canada) 2015

Save

Add Group		Search		Add Selected to Group		
All		ID	Name	Label	Weight	View
Demographics	<input checked="" type="checkbox"/>	v134374	Q1A	1a. If a federal election were held today, which party are you most likely to vote for?		
	<input checked="" type="checkbox"/>	v134345	Q1B	1b. Even though you may not have made up your mind, which party are you leaning towards at this time?		
	<input checked="" type="checkbox"/>	v134615	Q2	2. Do you approve or disapprove of the job Stephen Harper is doing as Prime Minister?		
	<input checked="" type="checkbox"/>	v134478	Q3	3. Do you approve or disapprove of the job Tom Mulcair is doing as Leader of the Opposition?		
	<input type="checkbox"/>	v134600	Q4	4. Do you approve or disapprove of the job Justin Trudeau is doing as leader of the Liberal Party?		
	<input type="checkbox"/>	v134634	Q5	5. Which party do you expect to win the next federal election?		
	<input type="checkbox"/>	v134338	Q6	6. Do you agree or disagree prostitution should be legal for sex trade workers and their clients?		



Search and reorder variables

Forum Research Political Poll - Federal Issues (Canada) 2015 Save

Add Group ≡+

Search
liberal

All	<input type="checkbox"/>	ID	Name	Label	Weight	View	
Demographics ✎	<input type="checkbox"/>	<div>⋮</div>	<div>v134466</div>	Q17	17. Is Canada becoming more conservative or is it becoming more liberal?		
	<input type="checkbox"/>	<div>Drag Me</div>	<div>v134600</div>	Q4	4. Do you approve or disapprove of the job Justin Trudeau is doing as leader of the Liberal Party?		

Items per page: 10 1 - 2 of 2 < >



Development Progress

- As of Dataverse 10.12 (core github):
 - New tabular data ingest code
 - New variable metadata edit API
 - New variable metadata data table
 - New ‘configuration’ code for modular apps (DCT)
 - Updated DCT UI for ‘Edit’ variable window
 - Updated DCT UI for Weighted Frequencies



Next Steps

- Full usability / user testing of DCT UI
- Upgrade DCT UI to latest version of Angular
- Address user testing w/ new UI developments
- Adopt / setup in SP Dataverse production release (fall 2019)



Discovery & Reuse Use Cases

- Enhanced variable metadata (code/variable descriptions)
- Support weighting
- Automated Codebook generation to accompany data
- Easy-to-understand online visualization of metadata/data
- Structured metadata for reuse in future



Dataverse DCT Integration

Root > Test dataset for tabular upload and curation

Success! — This dataset has been created

Metrics 0 Downloads

Contact Share Publish Edit

Test dataset for tabular upload and curation Draft Unpublished

Leahey, Amber, 2016, "Test dataset for tabular upload and curation", <https://doi.org/10.5072/FK2/PO5RFL>, Root, DRAFT

VERSION

Cite Dataset

Learn about Data Citation Standards

Description

Subject

Files Metadata Terms Versions

1 File

assets2008.SAV
SPSS SAV - 6.2 MB - 13-Nov-2016 - 0 Downloads
NCS: 01a3b5d2c9039f1e9530e16c551272
data

Ingest in progress...

Upload Files

Edit Files

Download

Copyright © 2018

Powered by **Dataverse** v 4.9.4

Researcher uploads tabular data to DV

Root > Test dataset for tabular upload and curation

Metrics 0 Downloads

Contact Share Publish Edit

Test dataset for tabular upload and curation

Leahey, Amber, 2018, "Test dataset for tabular upload and curation", <https://doi.org/10.5072/FK2/PO5RFL>, Root, DRAFT

VERSION

Cite Dataset

Learn about Data Citation Standards

Description

Subject

Files Metadata Terms Versions

1 File

assets2008.tab
Tabular Data - 11.9 MB - 13-Nov-2016 - 0 Downloads
169 Variables, 31509 Observations - UNF:6 sOsLF57UJNq3SIsYagFmfw== [fileUNF]

Configure Explore Download

Data Curation Tool

Copyright © 2018

Powered by **Dataverse** v 4.9.4

Dataset is converted to .tab file and 'Explore' and 'Configure' buttons appear

Saves metadata back to DV

British Political Poll - Federal Issues (Canada) 2015

ID	Name	Label	Original	View
1	Q1	1. Do you agree or disagree with the statement that the government is doing a good job?		
2	Q2	2. Do you agree or disagree with the statement that the government is doing a good job?		
3	Q3	3. Do you agree or disagree with the statement that the government is doing a good job?		
4	Q4	4. Do you agree or disagree with the statement that the government is doing a good job?		
5	Q5	5. Do you agree or disagree with the statement that the government is doing a good job?		
6	Q6	6. Do you agree or disagree with the statement that the government is doing a good job?		
7	Q7	7. Do you agree or disagree with the statement that the government is doing a good job?		
8	Q8	8. Do you agree or disagree with the statement that the government is doing a good job?		
9	Q9	9. Do you agree or disagree with the statement that the government is doing a good job?		
10	Q10	10. Do you agree or disagree with the statement that the government is doing a good job?		

Data Curation Tool application opens > User edits metadata



Questions?

amber.leahey@utoronto.ca

