



# Data Curation Tool Scholars Portal Dataverse

Meghan Goodchild  
Tech Tools for RDM Showcase - Portage Webinar  
June 16, 2020

# National grant for RDM development (CANARIE)

---

- “Dataverse for the Canadian research community: Developing reusable and scalable tools for data deposit, curation, and sharing”
- Led by Scholars Portal and UTL with support from CARL and Portage
- 3 key focus areas:
  - Authentication
  - Data Curation
  - Scalability and large-file support
- November 2018 - March 2020
- [Blog post](#) on project completion



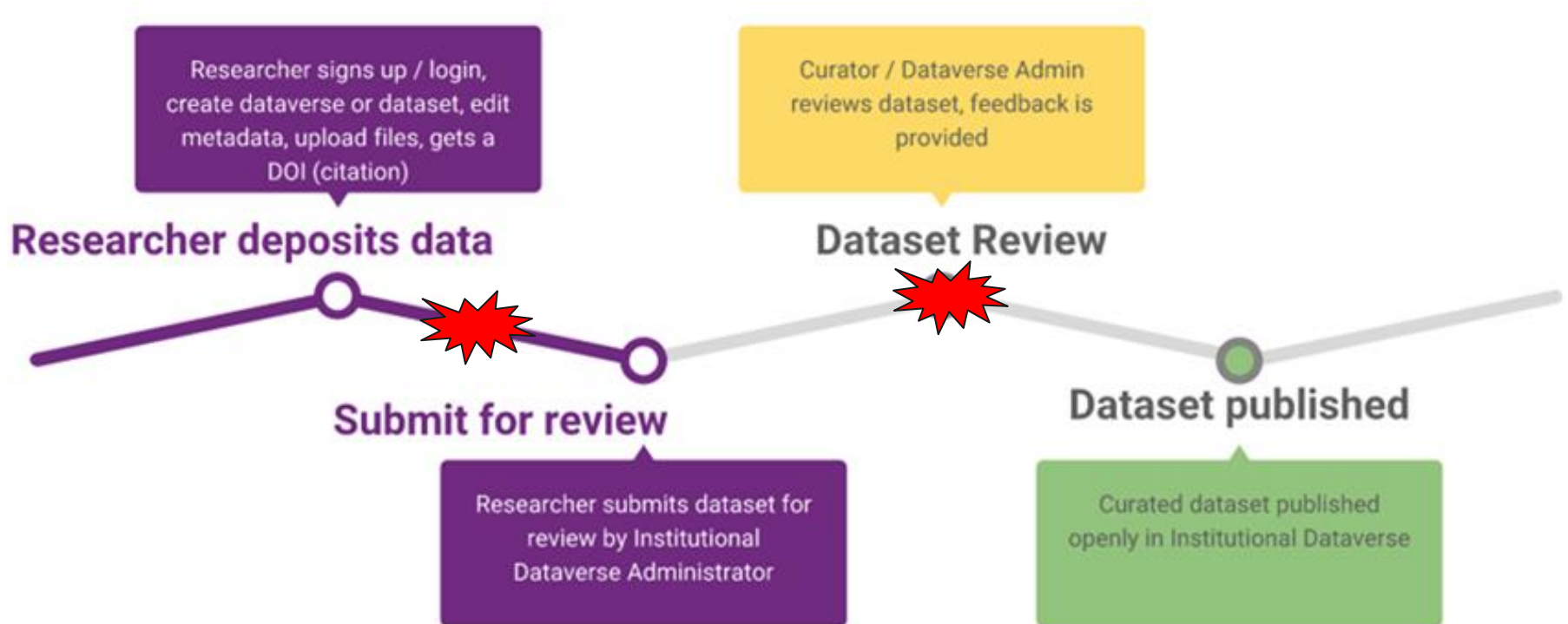
# Data Curation Tool



- “As a user, I want to add details about the variables in my dataset after I’ve uploaded it to Dataverse.”
- Developed the Data Curation Tool (DCT)
  - External web application that connects to Dataverse to create and edit metadata at the variable level using DDI standard
- Aim to improve data curation workflows and promote adoption of standards and best practices

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16
1	1.040e+14	4	26	104012	99	1	1973	2	1642	1040	1040	1	6	1	5	5
2	1.040e+14	1	19	104008	99	2	1943	1	99	99	99	1	1	6	4	5
3	1.040e+14	4	98	104003	99	1	1990	1	99	99	99	1	6	3	99	11
4	1.040e+14	2	98	104010	99	1	1983	2	1756	1040	1040	1	6	3	99	11
5	1.040e+14	1	18	104007	99	2	1927	1	99	99	99	1	1	6	4	1
6	1.040e+14	2	19	104007	99	1	1983	1	99	99	99	1	6	2	5	3
7	1.040e+14	4	15	104005	99	2	1970	1	99	99	99	1	1	2	4	6
8	1.040e+14	4	19	104006	99	1	1942	1	99	99	99	2	1	6	5	88
9	1.040e+14	4	15	104005	99	1	1965	2	1040	1276	1040	7	4	1	5	5
10	1.040e+14	7	15	104004	99	2	1955	1	99	99	99	1	1	2	5	5
11	1.040e+14	7	15	104003	99	2	7777	1	99	99	99	1	1	6	77	77
12	1.040e+14	4	18	104005	99	2	1938	1	99	99	99	1	3	6	5	5
13	1.040e+14	7	17	104005	99	1	1945	1	99	99	99	1	1	6	5	4
14	1.040e+14	4	18	104005	99	2	1949	2	1040	1380	1040	1	4	6	5	5
15	1.040e+14	2	15	104003	99	2	7777	1	99	99	99	1	1	2	4	5
16	1.040e+14	4	32	104007	99	1	1974	1	99	99	99	1	6	2	5	5
17	1.040e+14	4	98	104006	99	1	7777	1	99	99	99	1	6	3	99	11
18	1.040e+14	4	25	104010	99	2	1968	1	99	99	99	1	1	2	5	1
19	1.040e+14	4	40	104007	99	1	1967	1	99	99	99	1	1	2	6	2
20	1.040e+14	1	23	104004	99	1	1932	1	99	99	99	1	1	6	4	3
21	1.040e+14	2	18	104003	99	2	1965	1	99	99	99	1	6	2	5	5
22	1.040e+14	1	27	104011	99	1	1956	1	99	99	99	2	1	1	5	2
23	1.040e+14	4	18	104004	99	2	1923	1	99	99	99	1	3	6	3	3
24	1.040e+14	1	19	104006	99	2	1952	1	99	99	99	1	4	2	5	3
25	1.040e+14	3	16	104004	99	2	1947	2	1276	1040	1040	1	1	6	3	3

# Curation Workflows



# Data Curation Tool - Features



- View summary statistics for variables
- Add labels, groups, weights
- Create and edit variable-level metadata
- DDI XML file saved back to Dataverse and can be exported by users
- DDI-formatted HTML codebook for entire dataset that also includes variable-level metadata

# Development details



- Works with tabular data files (CSV, SAV, XLSX, etc.)
- Uses Angular 7, Angular Material Theme, & Dataverse API
- Completed UX testing with 5 participants
- French translations provided by the University of Ottawa
- GitHub: <https://github.com/scholarsportal/Dataverse-Data-Curation-Tool>

# DCT Demo





# Scholars Portal Dataverse

Search User Guide Support English Meghan Goodchild 29

Scholars Portal Dataverse > Survey test

Contact Share Link Edit



## Survey test

Version 1.1

Goodchild, Meghan, 2020, "Survey test", <https://doi.org/10.80240/FK2/288QUY>, Scholars Portal Dataverse, V1, UNF:6:W2m43fybeKUaPgWVg8qJmA== [fileUNF]

Cite Dataset

[Learn about Data Citation Standards.](#)

### Dataset Metrics

0 Downloads

Description asfasdf

Subject Earth and Environmental Sciences

Files Metadata Terms Versions

Upload Files

1 File

Edit Files



SampleSurveyData.tab

Tabular Data - 101 B - Apr 22, 2020 - 0 Downloads  
5 Variables, 10 Observations - UNF:6:W2m43fybeKUaPgWVg8qJmA==

Configure

Explore

Download

# Launch from “Configure” button for tabular file

The screenshot displays a file management interface with the following elements:

- Navigation tabs: Files, Metadata, Terms, Versions.
- Buttons: Upload Files, Edit Files.
- File list header: 1 File.
- File entry: LFS2016-01\_PUMF\_EN.tab, Tabular Data - 11.9 MB - Sep 19, 2019 - 6 Downloads, 75 Variables, 101887 Observations - UNF:6:EerkQFr2ySzwCu4oV5jExA==.
- Action buttons: Configure, Explore, Download.
- Callout box: Data Curation Tool.

The "Configure" button is highlighted with a white callout box labeled "Data Curation Tool".

# Labour Force Survey (Curated)

LFS2016-01\_PUMF\_EN.tab

Tester, Curation, 2019, "Labour Force Survey (Curated)", <https://doi.org/10.5072/FK2/YLJJAY>, Scholars Portal Dataverse, V1, UNF:6:EerkQFr2ySzwCu4oV5JExA== [fileUNF]

< Hide Groups

 Download

 Save

Add Group +


Search 


















Items per page: 25

1 - 25 of 75

< >

All Variables

Number of hours 

<input type="checkbox"/>	ID	Name	Label	Weight	View	
<input type="checkbox"/>	v14110	REC_NUM	Order of record in file			
<input type="checkbox"/>	v14130	SURVYEAR	Survey year			
<input type="checkbox"/>	v14106	SURVMNTH	Survey month			
<input type="checkbox"/>	v14127	LFSSTAT	Labour force status			
<input type="checkbox"/>	v14165	PROV	Province			
<input type="checkbox"/>	v14168	CMA	3 largest CMAs			
<input type="checkbox"/>	v14155	AGE_12	Five-year age group of respondent			
<input type="checkbox"/>	v14136	AGE_6	Age in 2 and 3 year groups			

# Labour Force Survey (Curated)

LFS2016-01\_PUMF\_EN.tab

Tester, Curation, 2019, "Labour Force Survey (Curated)", <https://doi.org/10.5072/FK2/YLJJAY>, Scholars Portal Dataverse, V1, UNF:6:EerkQFr2ySzwCu4oV5JExA== [fileUNF]

< Hide Groups


Download Save




















Add Group +

Search

Items per page: 25 1 - 25 of 75 < >

All Variables

Number of hours 

<input type="checkbox"/>	ID	Name	Weight	View	
<input type="checkbox"/>	v14110	REC_N			
<input type="checkbox"/>	v14130	SURVY			
<input type="checkbox"/>	v14106	SURVM			
<input type="checkbox"/>	v14127	LFSSTA			
<input type="checkbox"/>	v14165	PROV			
<input type="checkbox"/>	v14168	CMA			
<input type="checkbox"/>	v14155	AGE_12			
<input type="checkbox"/>	v14136	AGE_6			
<input type="checkbox"/>	v14158	SEX			

## Variable Information

ID: v14158 Name: SEX

Label: Sex of respondent

Literal Question

Interviewer Instructions

Post Question

Universe

Notes

# Labour Force Survey (Curated)

LFS2016-01\_PUMF\_EN.tab

Tester, Curation, 2019, "Labour Force Survey (Curated)", <https://doi.org/10.5072/FK2/YLJJAY>, Scholars Portal Dataverse, V1, UNF:6:EerkQFr2ySzwCu4oV5jExA== [fileUNF]

< Hide Groups

Download

Save

Add Group +

Search

Items per page: 25

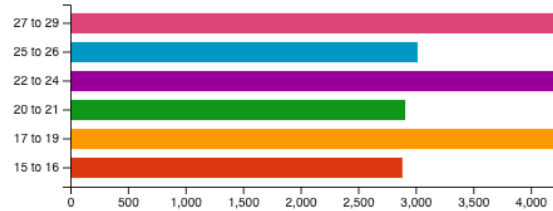
1 - 25 of 75

All Variables

Number of hours

<input type="checkbox"/>	ID	Name
<input type="checkbox"/>	v14110	REC_N
<input type="checkbox"/>	v14130	SURVY
<input type="checkbox"/>	v14106	SURVM
<input type="checkbox"/>	v14127	LFSSTA
<input type="checkbox"/>	v14165	PROV
<input type="checkbox"/>	v14168	CMA
<input type="checkbox"/>	v14155	AGE_12
<input type="checkbox"/>	v14136	AGE_6
<input type="checkbox"/>	v14158	SEX

## AGE\_6: Age in 2 and 3 year groups



Values	Categories	Count	Count Percentage(%)	Weighted Count
1	15 to 16	2,885	13.417	
2	17 to 19	4,224	19.644	
3	20 to 21	2,909	13.528	
4	22 to 24	4,219	19.621	
5	25 to 26	3,017	14.031	
6	27 to 29	4,249	19.76	

Age in 2 and 3 year groups

Sex of respondent

# Data Curation Tool Codebook

## Data Curation Tool Testing Dataset (ICPSR doi:10.5072/FK2/0TYIHL) (DCT Testing Dataset)

View: [Part 1: Document Description](#)  
[Part 2: Study Description](#)  
[Part 3: Data Files Description](#)  
[Part 4: Variable Description](#)  
[Part 5: Other Study-Related Materials](#)  
[Entire Codebook](#)

Document Description	
<b>Citation</b>	
<i>Title:</i>	Data Curation Tool Testing Dataset
<i>Identification Number:</i>	doi:10.5072/FK2/0TYIHL
<i>Distributor:</i>	Scholars Portal Dataverse
<i>Date of Distribution:</i>	2019-06-14
<i>Version:</i>	2
<i>Bibliographic Citation:</i>	Lubitch, Victoria; Leahey, Amber, 2019, "Data Curation Tool Testing Dataset", <a href="https://doi.org/10.5072/FK2/0TYIHL">https://doi.org/10.5072/FK2/0TYIHL</a>
Study Description	
<b>Citation</b>	
<i>Title:</i>	Data Curation Tool Testing Dataset
<i>Subtitle:</i>	DDI Test
<i>Alternative Title:</i>	DCT Testing Dataset
<i>Identification Number:</i>	doi:10.5072/FK2/0TYIHL
<i>Authoring Entity:</i>	Lubitch, Victoria (University of Toronto) Leahey, Amber (Scholars Portal)
<i>Producer:</i>	Leahey, Amber
<i>Date of Production:</i>	2019-05-22
<i>Grant Number:</i>	4445555
<i>Distributor:</i>	Scholars Portal Dataverse
<i>Date of Distribution:</i>	2019-06-14
<b>Study Scope</b>	
<i>Keywords:</i>	Astronomy and Astrophysics, test, smoking
<i>Topic Classification:</i>	Metadata

# Try it out!



- Launched in Scholars Portal Dataverse (October 2019)
  - Production Dataverse: [dataverse.scholarsportal.info](https://dataverse.scholarsportal.info)
  - Demo Dataverse: [demodv.scholarsportal.info](https://demodv.scholarsportal.info)
- GitHub: <https://github.com/scholarsportal/Dataverse-Data-Curation-Tool>



## Future developments

- Improvements to groups (sub-groups, re-ordering, viewing variables in group)
- Incorporating variable metadata within main Dataverse search (SOLR index)
- Investigating metadata elements/fields that could be added
- Improve accuracy of metadata in the DDI Codebook “Document Description” section
- Adding content from the DCT to the Data Explorer
- UI tweaks, including improved multi-select for variables



# Want to learn more?



- Blog post “Introducing the Data Curation Tool”:  
<https://spotdocs.scholarsportal.info/x/BIR0D>
- Basic documentation available in GitHub:  
<https://github.com/scholarsportal/Dataverse-Data-Curation-Tool#using-the-data-curation-tool>
- OCUl webinar on DCT (recording):  
[https://ocul.zoom.us/rec/share/yNxRIun2\\_zhOGqfL6VPtUbJiENzKT6a81SRIqKBfyB2m8tUX1e9ioLrZo6bQRInD](https://ocul.zoom.us/rec/share/yNxRIun2_zhOGqfL6VPtUbJiENzKT6a81SRIqKBfyB2m8tUX1e9ioLrZo6bQRInD)

---

**Comments, questions,  
suggestions?**

**[dataverse@scholarsportal.info](mailto:dataverse@scholarsportal.info)**