Introduction to Data Archiving with Borealis

At Scholars Portal using the Dataverse platform

Western Libraries

About Librarians / Archivists

- FYI – we can use your support
- We are seeking a fair deal from the employer
- Western needs more Librarians / Archivists

- Help us avoid a strike – potentially as early as **this Wednesday**

- **TAKE ACTION: https://www.uwofa.ca/take-action/**

UWOFA

Does your work as a faculty member or student include access to in-depth, subject-specific research and learning?

If you answered **YES**, please **SUPPORT** your Librarians and Archivists whose work is vital for:

- ✓ ensuring access to cutting edge information technologies
- ✓ protecting the integrity of our libraries
- ✓ ensuring access to quality resources for teaching and research
- ✓ maintaining Western's top-ranked contribution to new scholarship

Librarians & Archivists

…because librarians do it better than Google

Western Libraries

UWOFA

Bargaining 2023

We are still fighting for two important things: 1) complement (personnel numbers) and 2) a fair monetary package

- The Employer continues to reject any commitments to maintain let alone increase the complement of Librarians and Archivists. That means fewer people to help you.
- L&A Members deserve to be respected and valued. They are paid less at Western than other universities such as Brock, Waterloo, McMaster and Guelph
- The Bargaining Unit is small and the $ we are for is asking for is well within what the Employer can afford. Workloads and student-to-Librarian-and-Archivist ratios are at an all time high, which actually saves the Employer money.
- Western is in a very strong financial position and came through the pandemic in a much better state than many other institutions.
- The monetary offer currently on the table is less than recent settlements in the sector. It does not make up for wage suppression or inflation.

Librarians & Archivists

Western Libraries

Background

Tri-Agency Data Management Policy

- Policy includes 3 requirements:
  - Institutions: Institutional Strategy – “By March 1, 2023, research institutions subject to this requirement must post their RDM strategies and notify the agencies when they have been completed.”
  - Researchers: Data Management Plans – “By spring 2022, the agencies will identify the initial set of funding opportunities subject to the DMP requirement.”
  - Researchers: Data Deposit – “After reviewing the institutional strategies and in line with the readiness of the Canadian research community, the agencies will phase in the deposit requirement.”

Data deposit

- In cases where journal publications, pre-prints and other research outputs arise from agency-supported research, researchers may be required to deposit research data and code in an appropriate repository
  - The code is a crucial addition – focus here is not so much on reuse of the data but reproducibility of the results
- Goals:
  - Ensure proper use of public funds
  - Facilitate reproducibility of results
  - Facilitate data sharing

FAIR Principles for Open Data

- **Findable**
  - The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets
- **Accessible**
  - Once the user finds the required data, they need to know how they can be accessed, possibly including authentication and authorization.
- **Interoperable**
  - The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.
- **Reusable**
  - The ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.
- http://go-fair.org/

Concerns

- The concept of required data sharing is something that many researchers find problematic. Ethical, ownership and data sovereignty can all be legitimate barriers to data sharing, and the Tri-Agencies recognize this.
- Data deposit does not necessarily mean data sharing
- If you as a researcher think in good faith that there is a significant privacy, security or ownership issue that makes sharing inadvisable… **you are probably right**. Explain in your Data Management Plan and outline how you will handle preservation without sharing or explain why preservation is inadvisable. Funding agencies are aware that there are many cases in which data cannot or should not be shared.

Dataverse and Borealis

Western Libraries

What is Dataverse? What is Borealis?

- Dataverse is a research data platform and repository for sharing, discovering and preserving research data
  - An open source project of Harvard University's Institute for Quantitative Social Sciences. Formal development began in 2006, based on earlier projects dating back to 1987
- Many international collaborators and used world-wide for all kinds of data (not just social science!)
- Borealis is a Canadian implementation of Dataverse currently hosted by the Ontario Council of University Libraries and supported by the Digital Research Alliance of Canada as an approved repository
- Dataverse is the name for the software, for the hosted archive and for individual collections of data files – Western University's Dataverse runs on the Dataverse installation at Scholar's Portal, and contains several researcher dataverses

63 Installations
Dataverse installations around the world
NORTH AMERICA
SOUTH AMERICA
ASIA
AFRICA
OCEANIA
Atlantic Ocean
Pacific Ocean
Indian Ocean
Leaflet | © OpenStreetMap contributors © CARTO

Some benefits of Borealis for researchers

- Web-based tool with no installation required
- Meet grant & journal requirements for sharing research data, compliant with Tri-Agency requirements
- Share research data publicly, privately, within a single institution, or within a single research team
- Store data for long-term access
- Create a persistent identifier (DOI) for data
- Track versions of data over time
- Automatically generate citations for datasets

The Borealis project at Scholar's Portal

- Scholars Portal is a collaborative project of the Ontario Council of University Libraries to provide shared infrastructure, collections and technology
- The SP Dataverse project began in 2012. Housed on University of Toronto servers to provide fully Canadian-hosted data archiving
- Local development includes French internationalization project with University of Montreal, integration with DataCite Canada, etc. Hosts dataverses for institutions across Canada
- Recently rebranded Borealis: a nationally coordinated, standardized data archiving service using Dataverse

Key software features

- Supports uploading of any file type, including shapefiles, images, text, tabular data and code
- Certain types of tabular data files (Stata, SPSS, Excel, csv) are ingested into the data explorer and can be viewed online using the Data Explorer
- Geospatial shapefiles can be further explored and manipulated through integration with WorldMap
- Data can be made available openly, shared by request only, or only shared within a specific research team or user group
- Securely backed up, versioning control allows you to access old versions
- Generates a standard data citation (DOI) for use in research publications
- Author retains control and can choose from several default licenses

Key limitations

- Current file size limit is 3.5 GB (although changes regularly and exceptions are possible). For larger datasets (into the terabytes) consider Canada's Federated Research Data Repository (FRDR)
- Not suitable for confidential data – while you can restrict access to files, it is not approved for data with ethical concerns around sharing and they **really** do **not** want your confidential data
- File ingest doesn't deal nicely with Excel data that doesn't fit its expectations for a tabular file
- While data can be updated, removing a dataset or Dataverse is strongly discouraged and can only be done by platform administrators. Make sure you know what you are doing before you hit "publish"

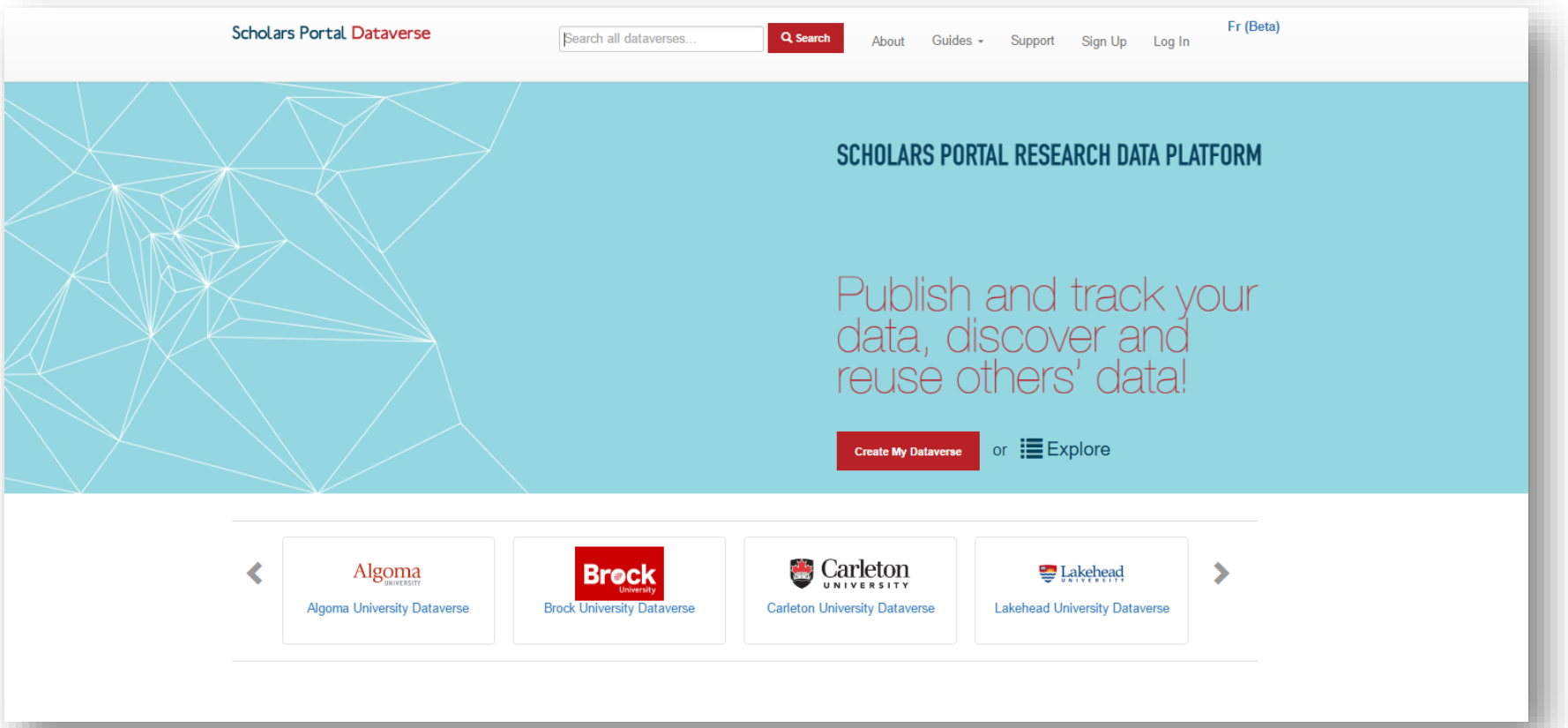
Publishing control

- Choose your access settings: modify access to your data files - ranging from fully open to completely restricted.
- Customize the terms of use, specifying how your data can be used, using one of our default licenses or a custom one.
- Modify access to your data files at any time.
- Once your dataset has been published, your metadata is publicly available. Therefore, your data is always discoverable.
  - If you don’t want your data to be discoverable but need to share your data (e.g. within your research group or to meet a journal requirement) keep the file unpublished. Data will be available at a private url but not discoverable.

Sharing your data

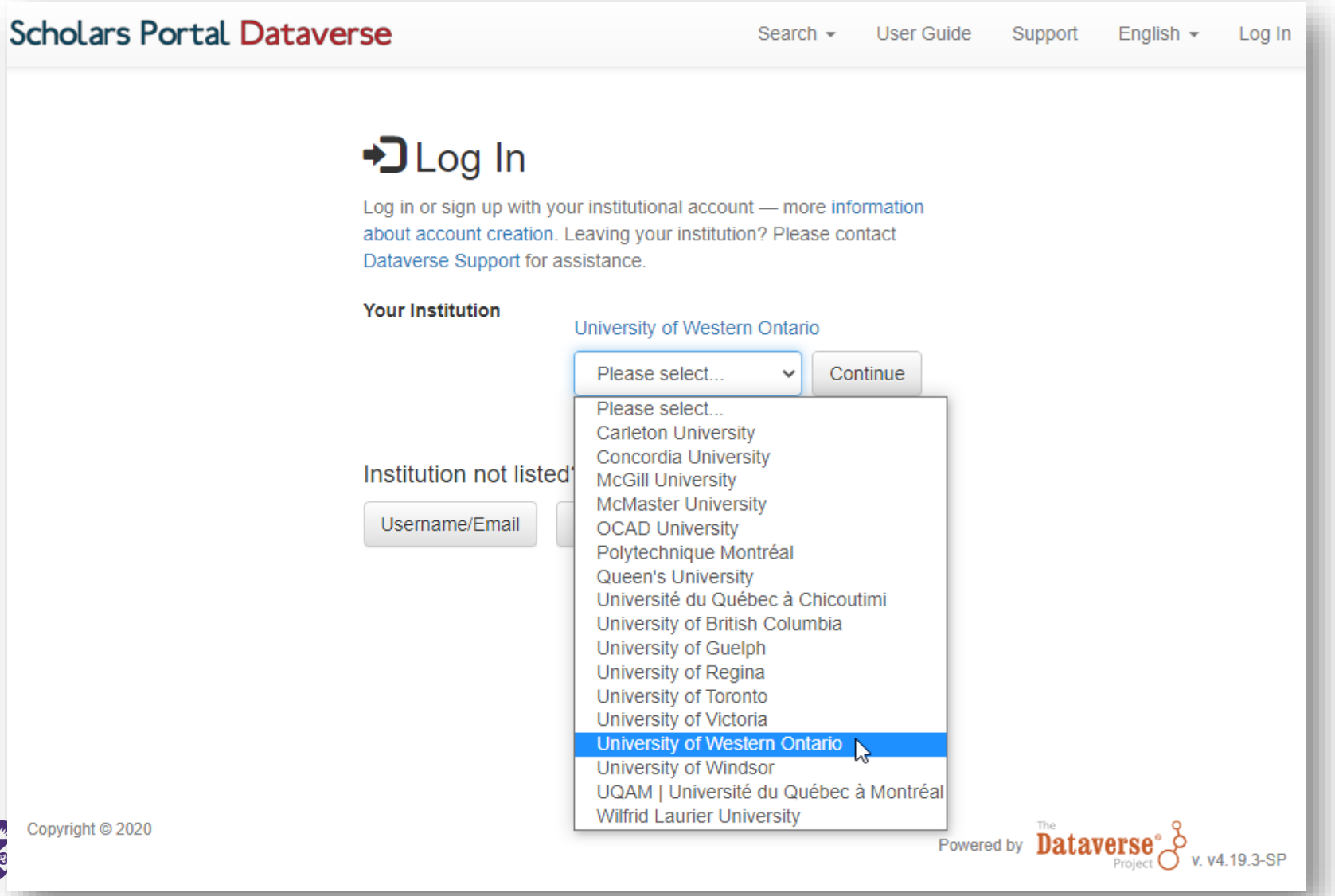
Western Libraries

Getting Started



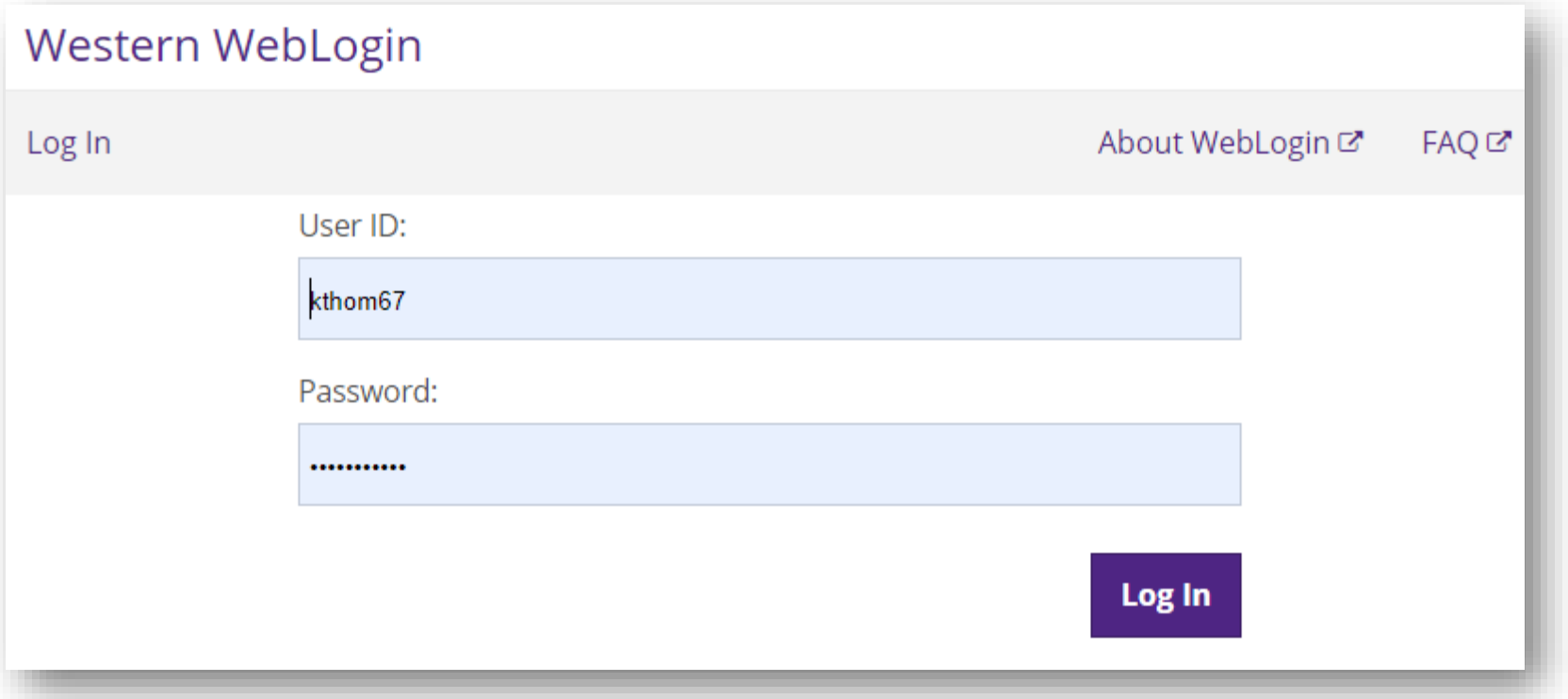
Western Libraries

Log in



Western

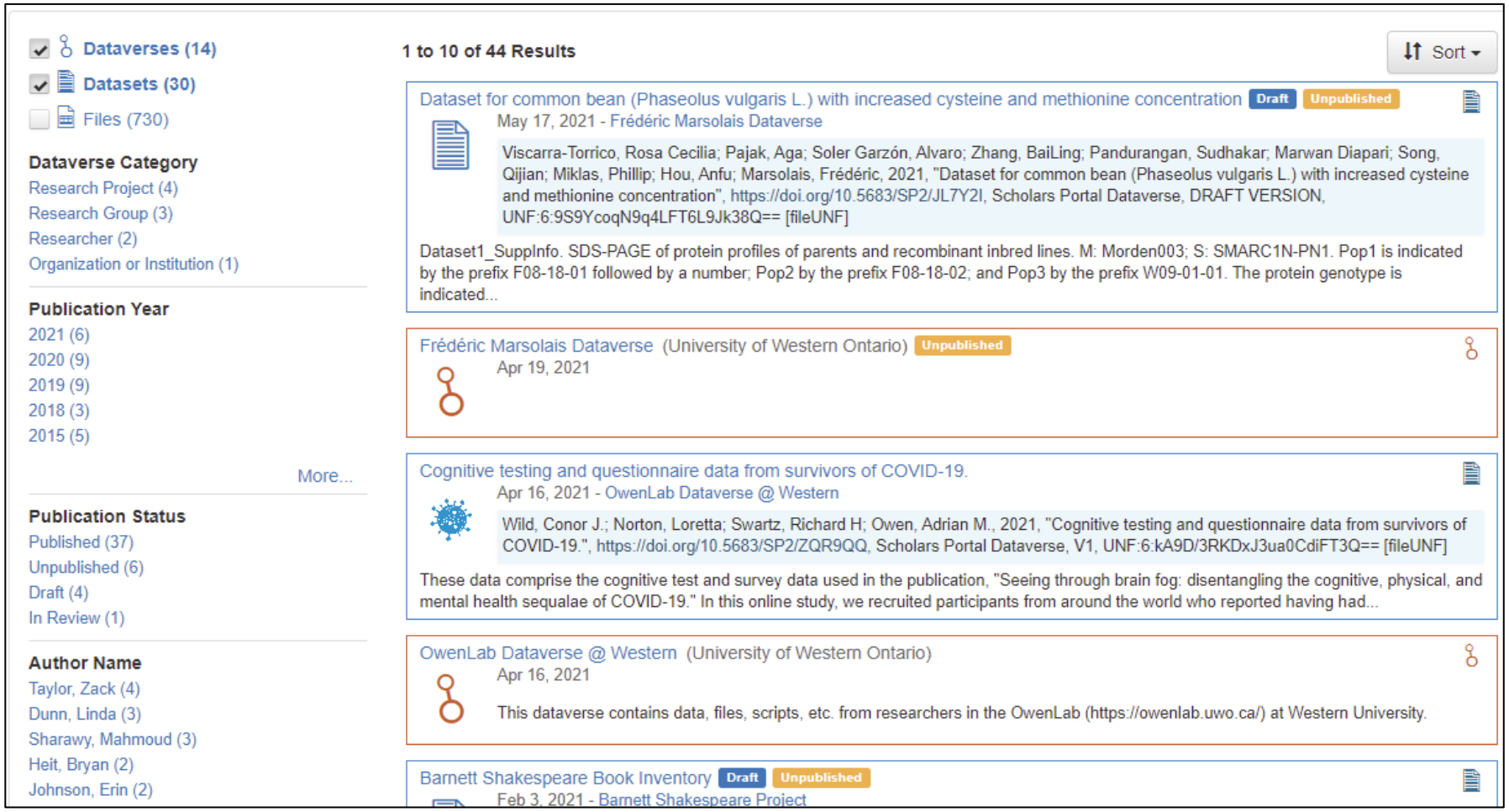
Using your Western credentials



Western Libraries

Western data repository

- https://borealisdata.ca/dataverse/westernu



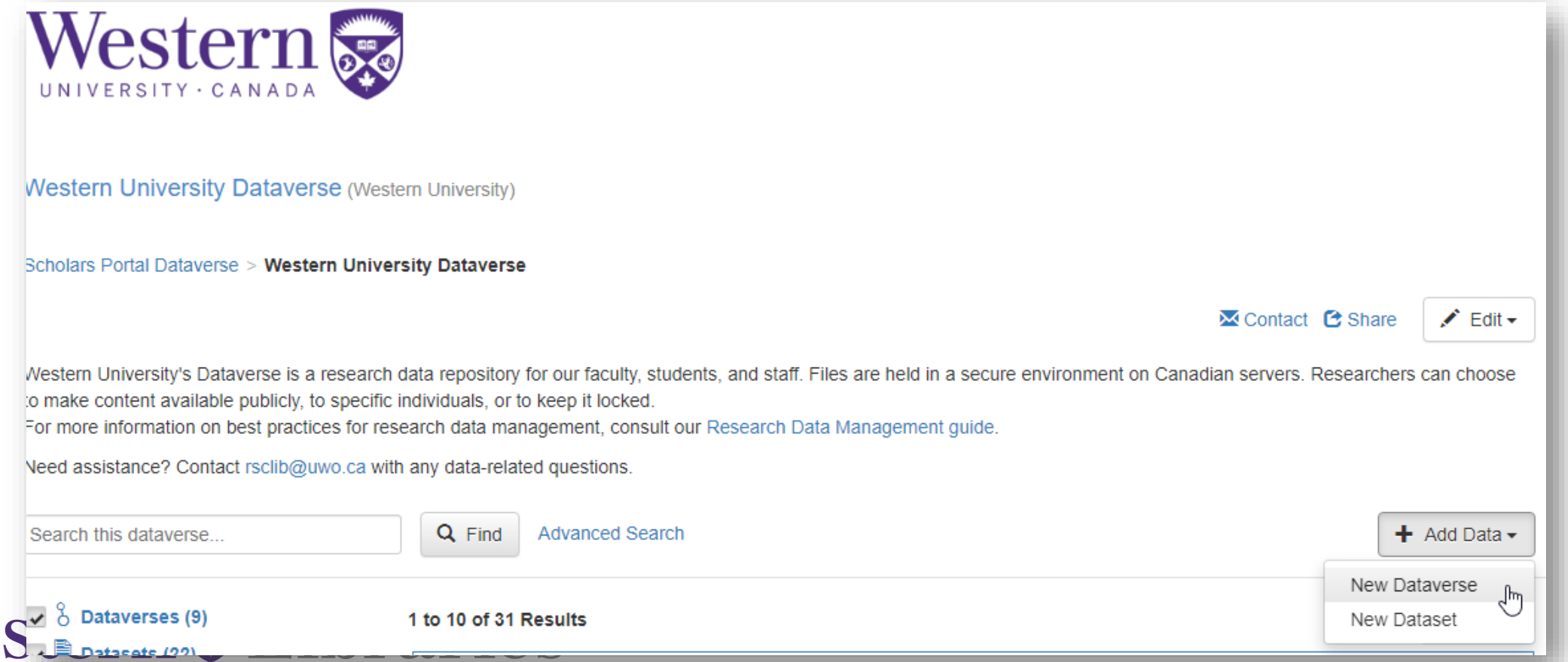
Western Libraries

In-line file types

- Allows users to view certain file types directly in the browser without needing to download the file
- File types include:
  - Text and document files: txt, html, pdf
  - Audio files: mp3, mpeg, wav
  - Image files: gif, jpeg, png
  - Video files: mp4, quicktime
  - Statistical formats: Stata, R

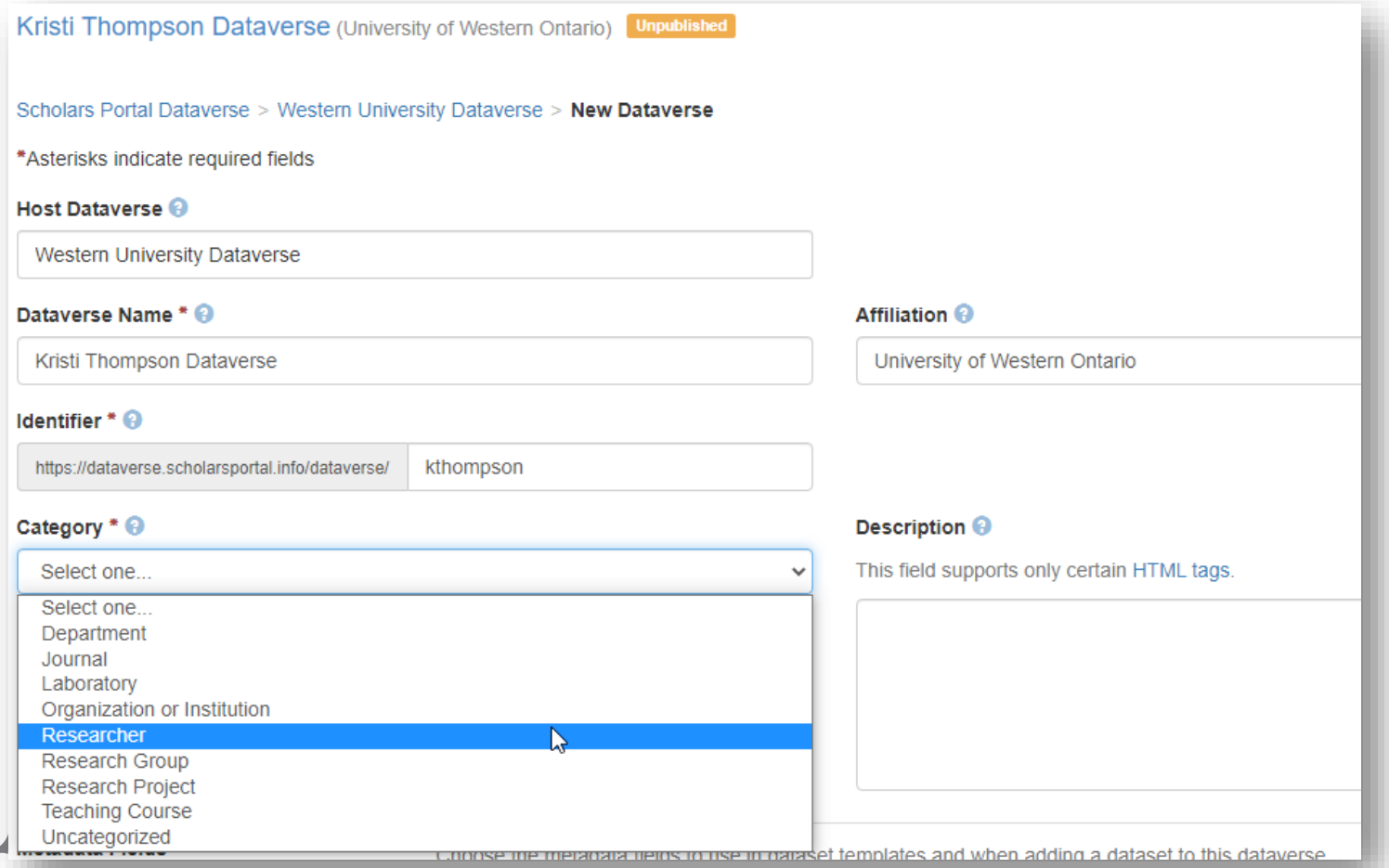
Adding a new data collection

- Create a new Dataverse to hold your data – don't just try to dump it in the root Western Dataverse



Creating your Dataverse

- Basic metadata

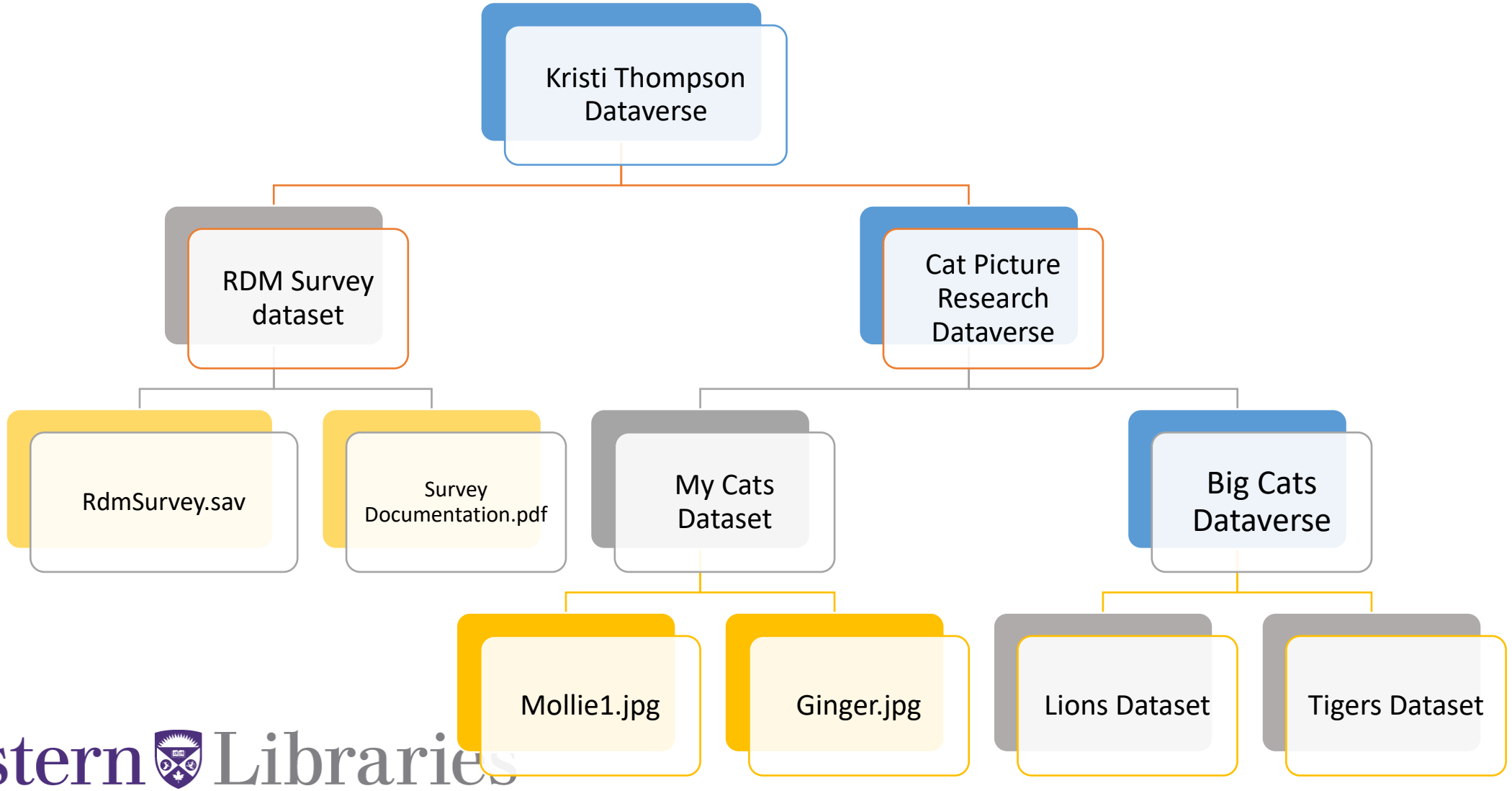


Western L

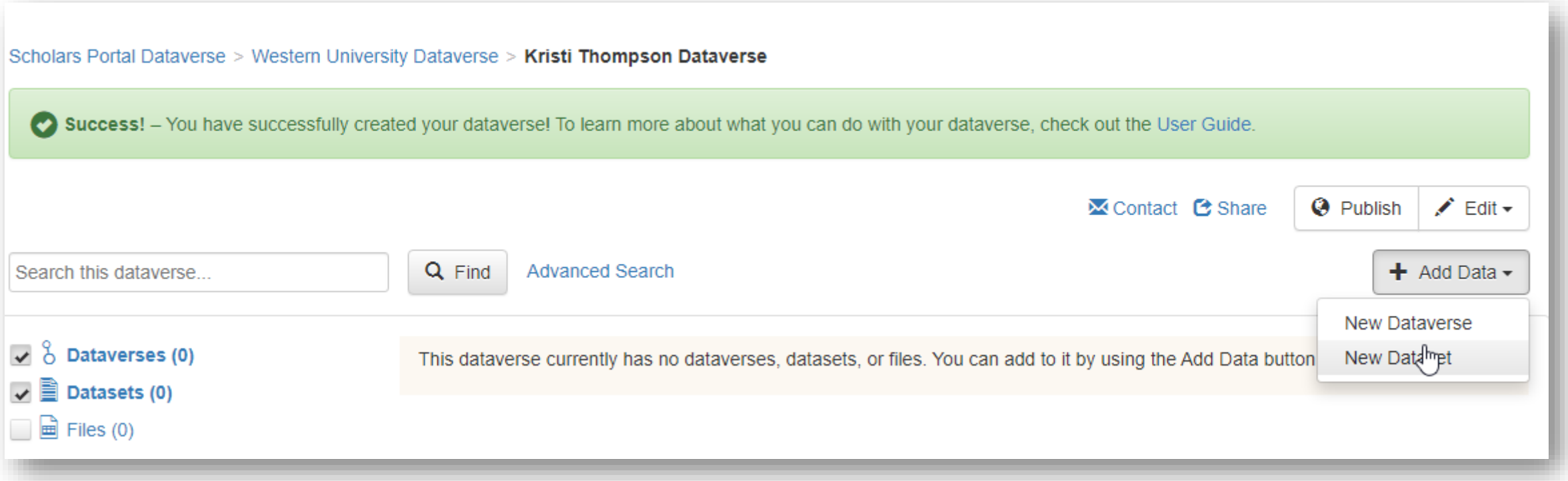
Organizing with sub-Dataverses

- I might decide to keep all my datasets in one “Kristi Thompson” researcher Dataverse
- Each dataset you create can have multiple files associated with it
  - In theory you could create one dataset as a holder for a bunch of unrelated files, but that would be weird…
- If you have multiple unrelated research projects you might create a Dataverse for each of them
- If you have a very complicated research life you could even create several levels of Dataverse, treating it like a folder hierarchy
- Whatever you do, make up your mind before publishing. After a dataset or Dataverse has been published, it can only be moved by an employee of our host at Scholars Portal

Researcher Dataverse organization



Adding data



Western Libraries

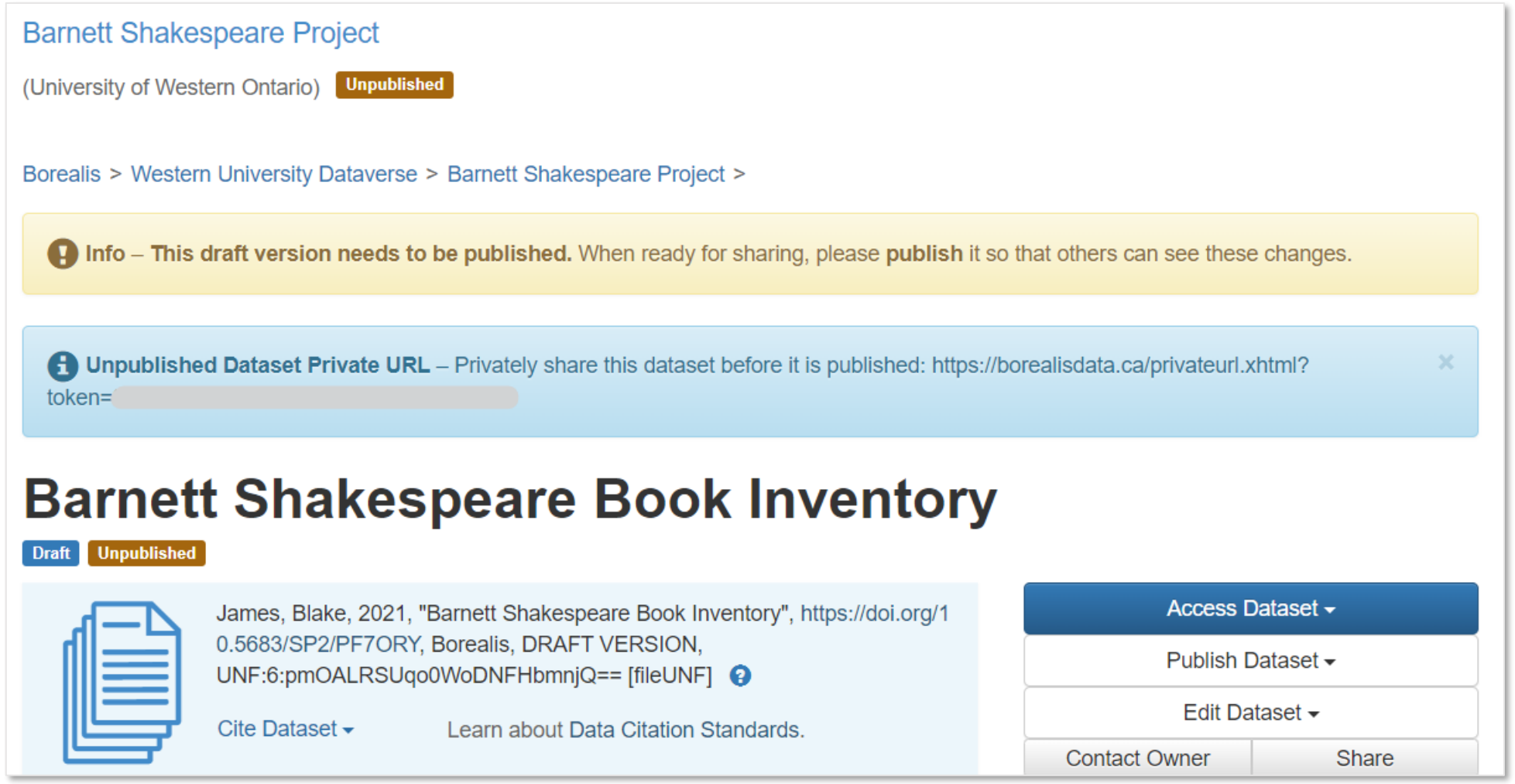
Data ingest

- If you upload tabular data (e.g., SPSS, STATA, RData, CSV, and Excel), the Dataverse software will ingest your tabular files and convert them to a tab-delimited file
- Ingested tabular data can be exported to / downloaded as a variety of formats including Excel, SPSS, Stata
- The pre-ingested tabular file you uploaded is saved within the repository and is available for download. However, the download method may not be obvious to users so consider adding a note if you want people to do this.

Ingest Oddities

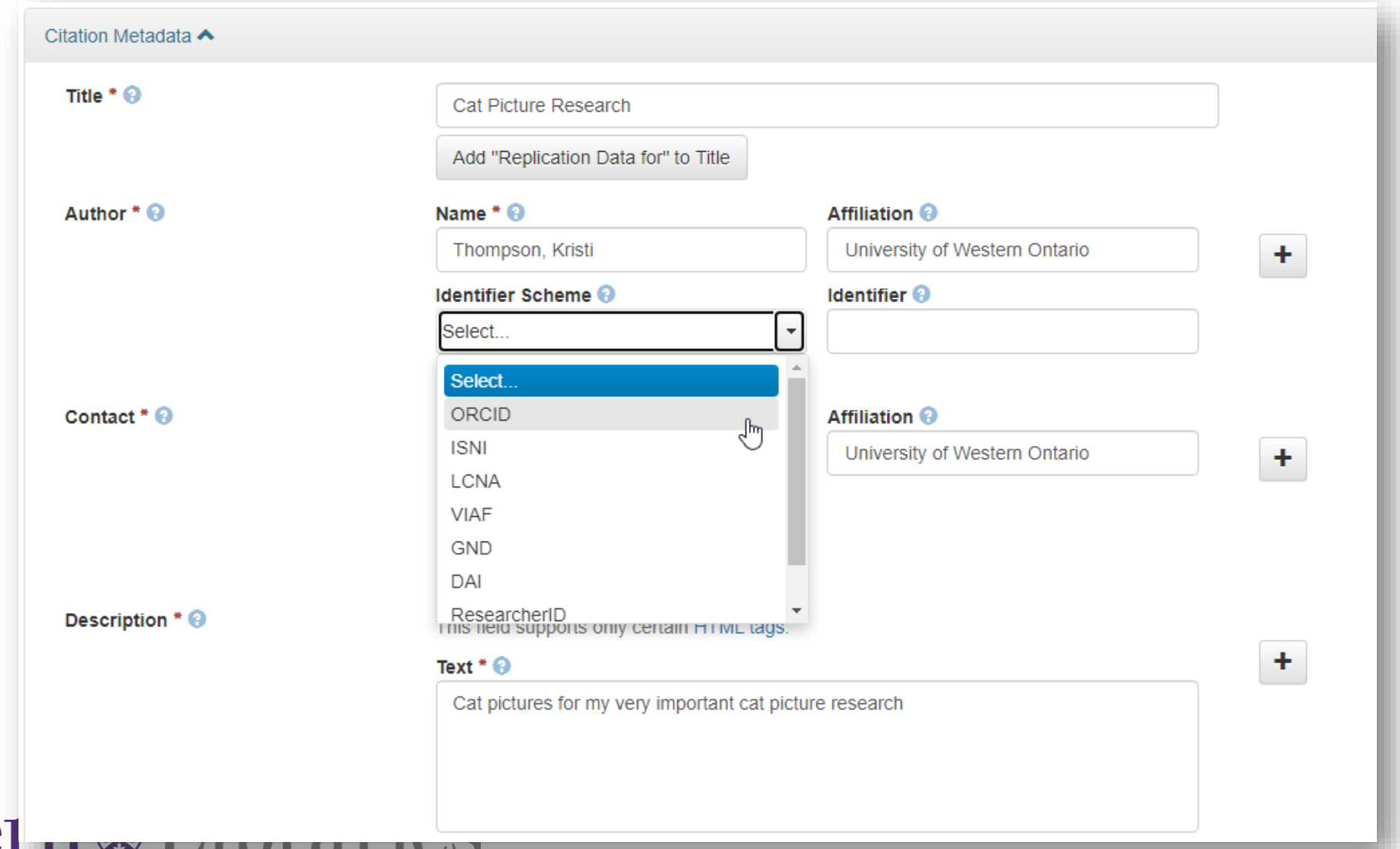
- ZIP (compressed) files can be uploaded to your dataset using the same method as any other file format. However, when the upload is complete, each file within the ZIP file will be listed, instead of a single ZIP file. The Dataverse software unzips the first layer of compression during the upload process (if you have a zipped zip file, it will stay zipped)
- The ingest process will only process the first tab in a spreadsheet file. All other tabs will be ignored and will not be included in the tabular file that's created. If you want all the tabs in a spreadsheet file to be converted to the TAB format, you'll need to upload each tab as its own file.

Sharing unpublished data

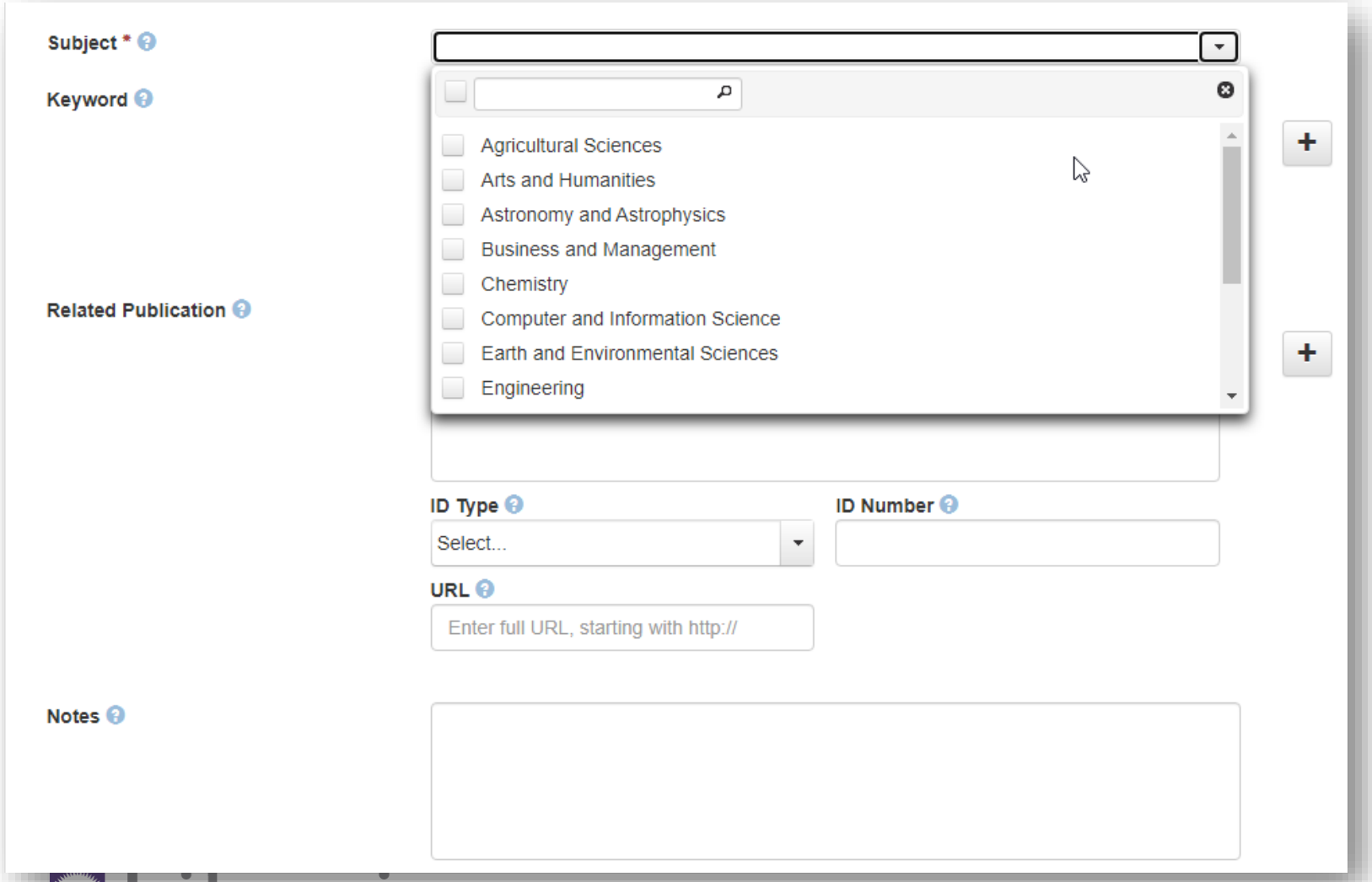


Western Libraries

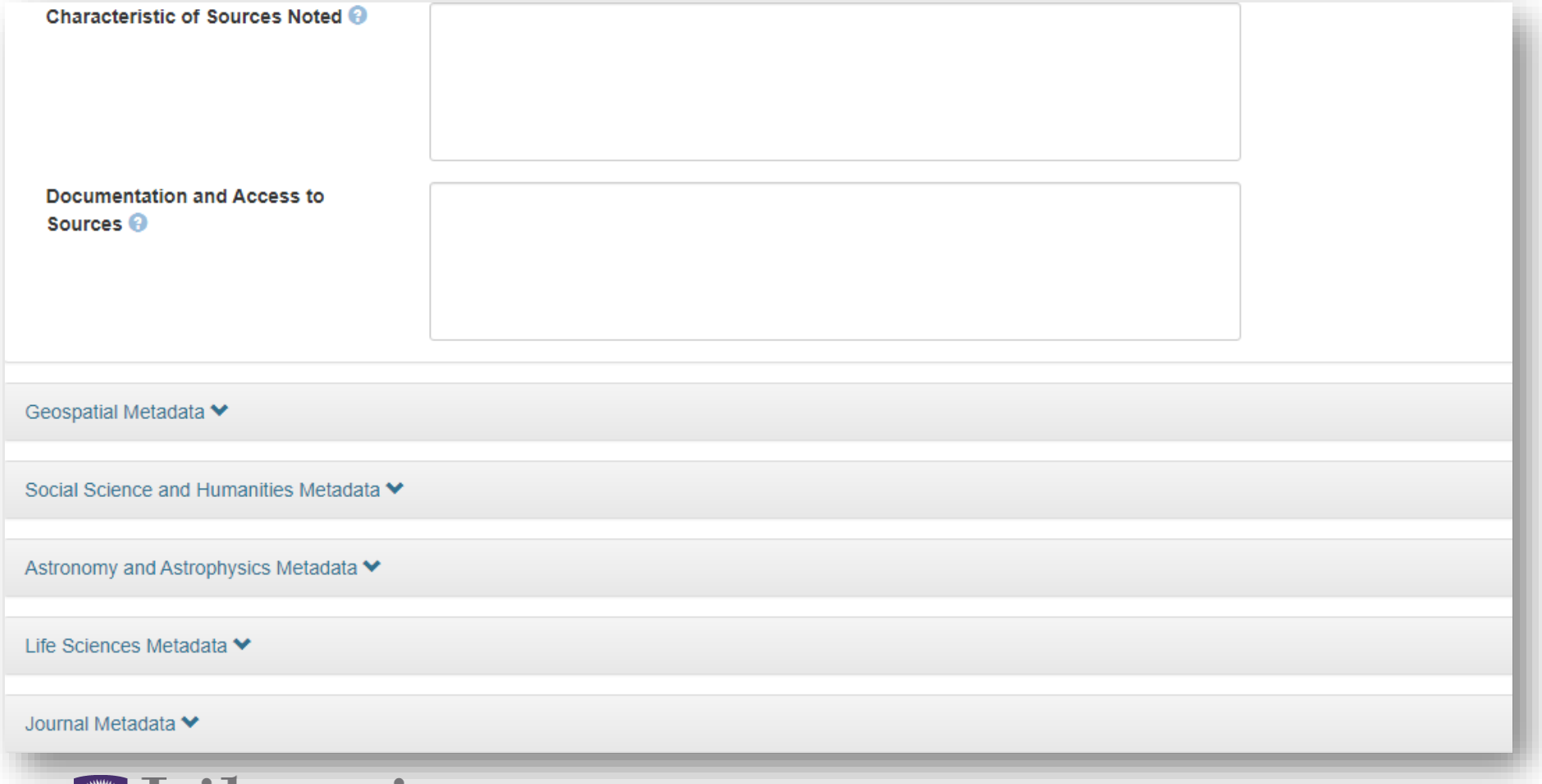
Metadata!


Citation Metadata
Title *
Cat Picture Research
Add "Replication Data for" to Title
Author *
Name *
Affiliation
Thompson, Kristi
University of Western Ontario
Identifier Scheme
Identifier
Select...
Select...
ORCID
ISNI
LCNA
VIAF
GND
DAI
ResearcherID
Contact *
Affiliation
University of Western Ontario
Description *
Text *
Cat pictures for my very important cat picture research

More metadata!



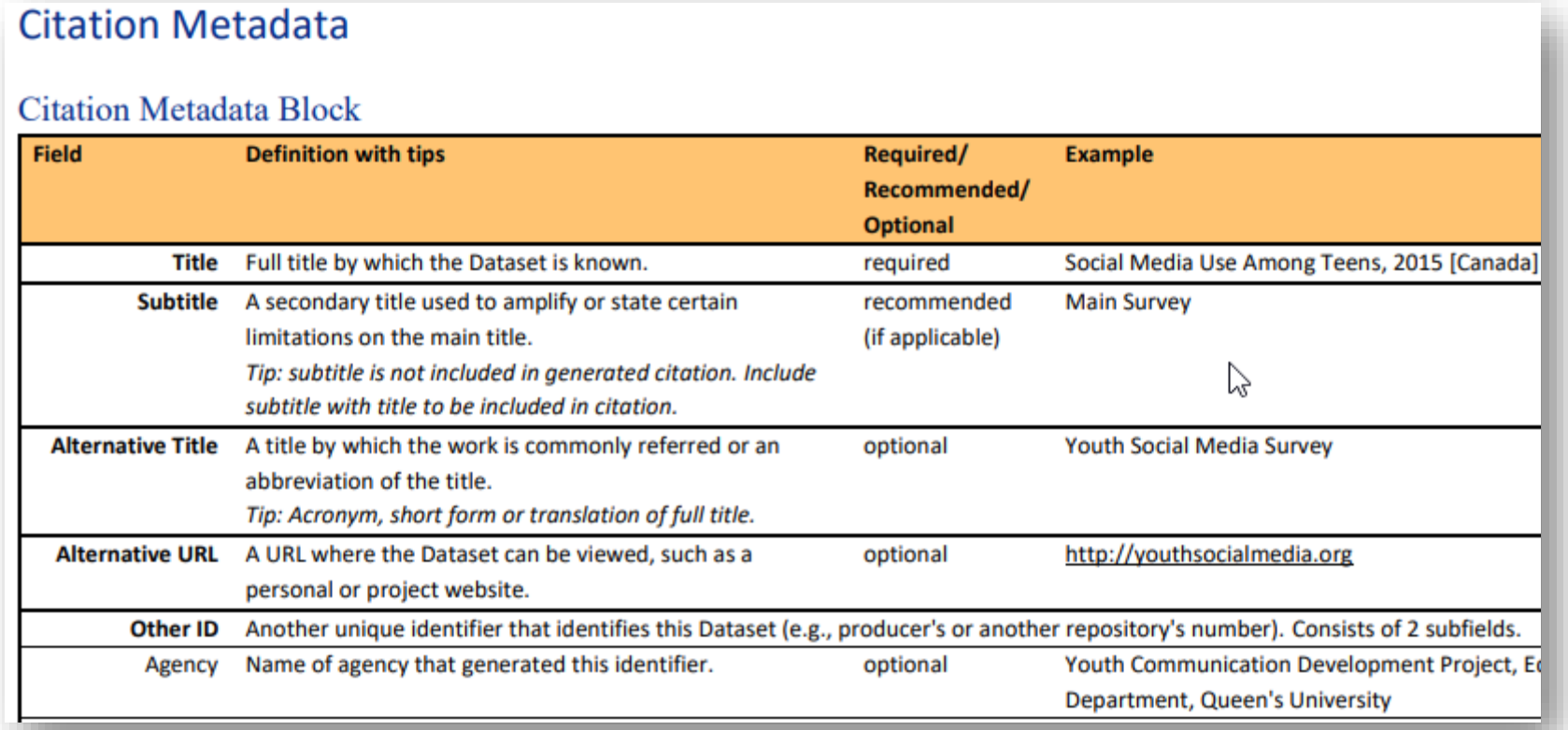
Optional metadata – only accessible after you save your dataset and then go back to “edit”



Western Libraries

Do you really love metadata?

- [Canadian Dataverse Metadata Best Practices Guide]



Citation Metadata

Citation Metadata Block

Field	Definition with tips	Required/ Recommended/ Optional	Example
Title	Full title by which the Dataset is known.	required	Social Media Use Among Teens, 2015 [Canada]
Subtitle	A secondary title used to amplify or state certain limitations on the main title. *Tip: subtitle is not included in generated citation. Include subtitle with title to be included in citation.*	recommended (if applicable)	Main Survey
Alternative Title	A title by which the work is commonly referred or an abbreviation of the title. *Tip: Acronym, short form or translation of full title.*	optional	Youth Social Media Survey
Alternative URL	A URL where the Dataset can be viewed, such as a personal or project website.	optional	http://youthsocialmedia.org
Other ID	Another unique identifier that identifies this Dataset (e.g., producer's or another repository's number). Consists of 2 subfields.		
Agency	Name of agency that generated this identifier.	optional	Youth Communication Development Project, E Department, Queen's University

Western Libraries

Applying a license

- Default to CC0 (open data) OR:
  - Templates for Create Commons Licences
  - Custom terms of use ‘Data Usage Agreement’
- Control access to your data at the dataset and file level:
  - Assign permissions for collaborators, curators, file downloaders (access)
  - Granular file-level permissions
  - Restricted files and custom terms of access

Publishing your Dataverse and datasets

- After creating your Dataverse you can customize the appearance with header images, custom text and links, and create a guestbook
- A new Western Dataverse user can create unpublished Dataverses and datasets within Western's Dataverse and submit them for publication, but the administrator of the Western Dataverse (that's me!) needs to approve them before they are made public
- Remember – once it's published, you can't delete it!

Western Libraries

Why default to CCO (and why can't I delete it?)

- Dataverse supports the ideal of "FAIR" data
- **F**indable
  - Open archive, documented, searchable
- **A**ccessible
  - Persistent identifier, always remains available – at least metadata even if data is withdrawn
- **I**nteroperable
  - Machine readable, standard formats
- **R**eusable
  - Well-documented, understandable, license allows for re-use

FAIR Data:
“As open as possible, as closed as necessary.”

Links and resources:

- Western University's Data Archive on Borealis
- Western Research Data Management Guide
- Borealis User Guide
- Dataverse North Metadata Best Practices Guide

For more information or to book an appointment to discuss your data, contact Western Libraries Research and Scholarly Communications Team, rsclib@uwo.ca

- Research Data Librarian Kristi Thompson
- Data Librarian Liz Hill

Western Libraries