Geodisy Project - Search Canadian research data data by location

Eugene Barsky, UBC

November 2019; Funded by CANARIE (RDM-059)

We aim for Canadian research data to be searched, filtered, and browsed using geographic locations

- Search results are driven by an interactive map
- Location is the primary search facet, linking resources from a similar area
- Relies less on textual searching, which is not ideal for spatial data

Find Data

Search FRDR to find research datasets originating from researchers affiliated with Canadian institutions. Data

Deposit Data

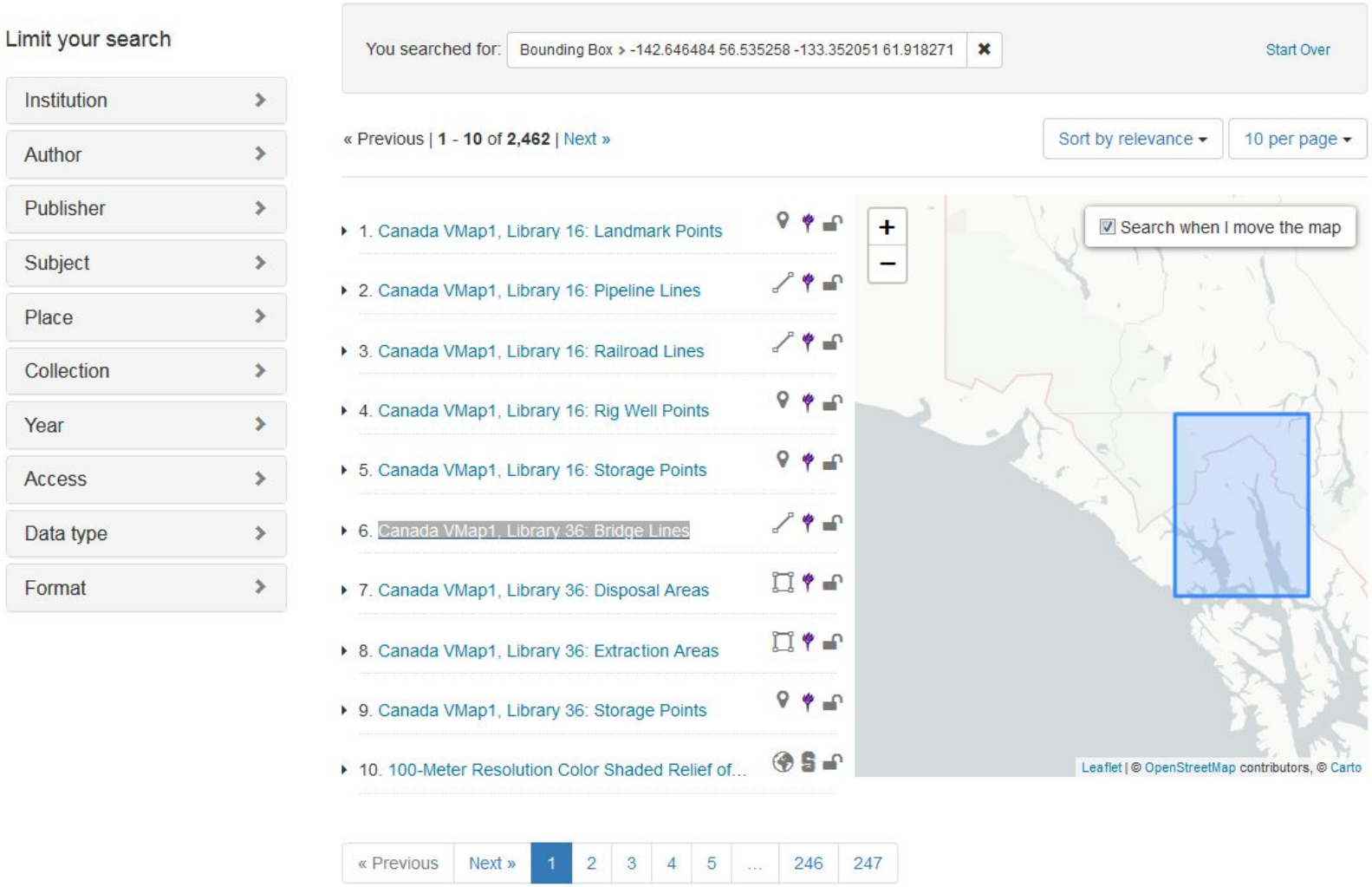
Any researcher affiliated with a Canadian institution can deposit data into FRDR. The platform can efficiently ingest datasets of

Why use it?

- Data can be difficult to find! When searching for data about a particular place, keywords can be hit or miss. A text search might look something like this:

((British Columbia OR BC OR B.C.) N2 (north*)) OR (Alaska N2 south*) OR (Yukon N2 south*) OR (Glacier Bay ADJ2 (park or preserve)) OR (Tatshenshini-Alsek ADJ2 park) OR (Kluane ADJ2 (park OR reserve)) OR (Atlin ADJ2 (park OR recreation area) OR …

Geodisy will show you where, in addition to what (**expected** GeoBlacklight user interface - NYU Spatial Data Repository)

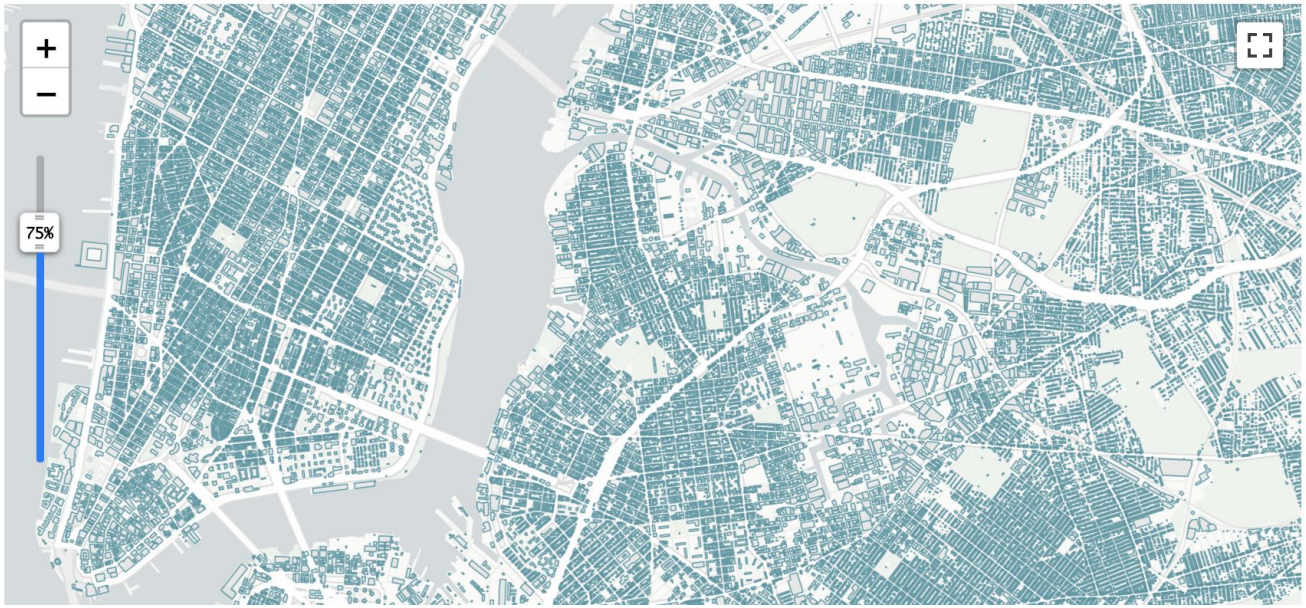


Example record page (**expected** GeoBlacklight user interface - NYU Spatial Data Repository)

« Previous | **6** of **3,128** | Next » Back to Search Start Over

2010 Buildings

Publisher: New York (City). Department of City Planning
Place(s): New York City, New York, United States
Subject(s): Buildings and Urban density
Format(s): Shapefile
Year(s): 2010
Held by: NYU
Preservation record: http://hdl.handle.net/2451/34043



Tools

- Email
- Web services
- Open in Carto

Download(s)

Shapefile

Other Formats

Derived Shapefile	Download
KMZ	Download
GeoJSON	Download

Why is it important?

Geodisy is open to all users and will benefit any research area that has use for location-based discovery, including climate change, community development, public health, conservation, journalism, and many more

Geospatial discovery is possible using metadata

- Metadata = information that describes a resource
- Geospatial data = machine readable using a GIS
- Quasi-geospatial data = data with a location component that isn't true geospatial data*
- Bounding boxes = rectangles representing the spatial extent of a data set

*to generate bounding boxes from quasi-geospatial data we are using geonames.org

Geodisy[1] (re-)uses 3 main open-source software components

- **Dataverse[2]:** Research data repository

  +

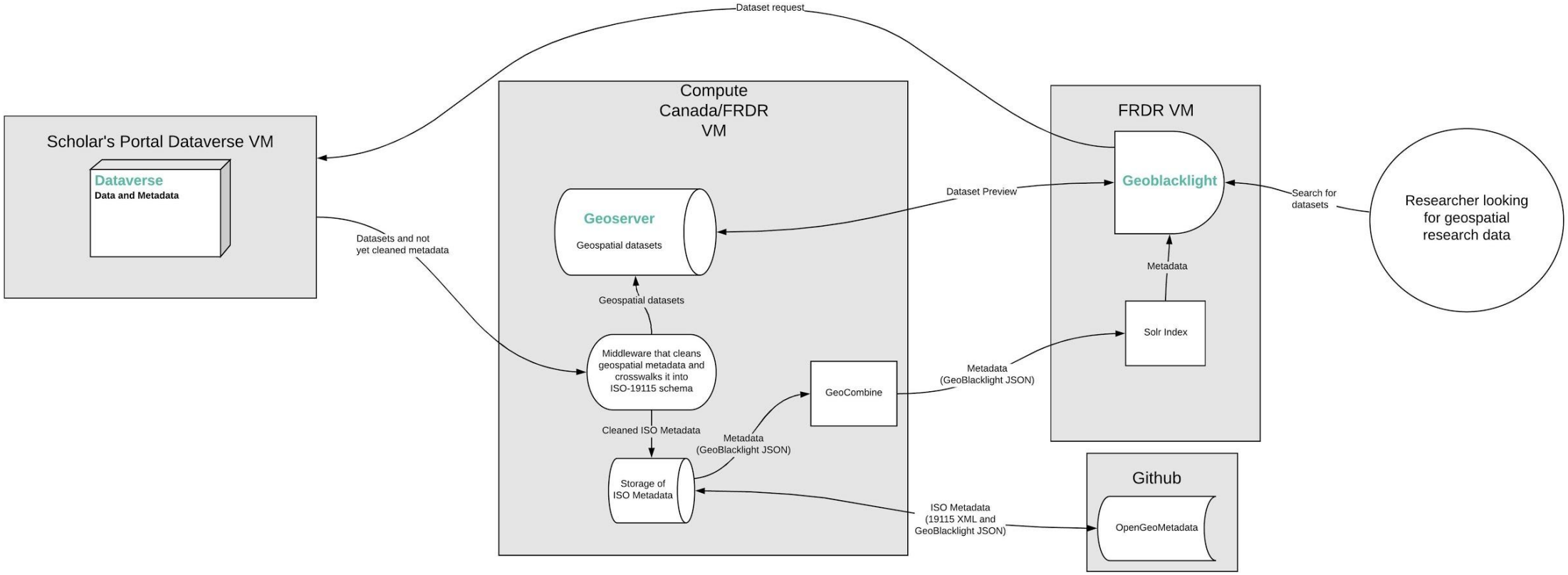
- **GeoServer:** Server for publishing and distributing geospatial data

  +

- **GeoBlacklight:** Geospatial discovery layer

[1] Geodisy source code and documentation is available in github - https://github.com/ubc-library/geodisy.
[2] For the initial step, for March 2020, Geodisy is funded to work with Canadian Dataverses only

Geodisy Architecture



Project pipeline (in steps):

1. Software will query datasets from the Scholars Portal Dataverses (and later from UAlberta, Dal, UNB, UManitoba, etc.) to determine which have geospatial information
2. Software will harvest metadata from relevant quasi-geospatial datasets
3. Software will harvest both metadata and data files from geospatial datasets
4. Software will enrich the metadata with bounding boxes using GDAL (or from metadata or Geonames, if needed) and other information
5. Software will transform metadata to more universal standards (ISO 19115 and GeoBlacklightJSON)
6. Software will deposit geospatial data into Geoserver
7. Software will deposit geospatial and quasi-geospatial metadata into OpenGeoMetadata
8. Metadata will be harvested by GeoBlacklight for discovery

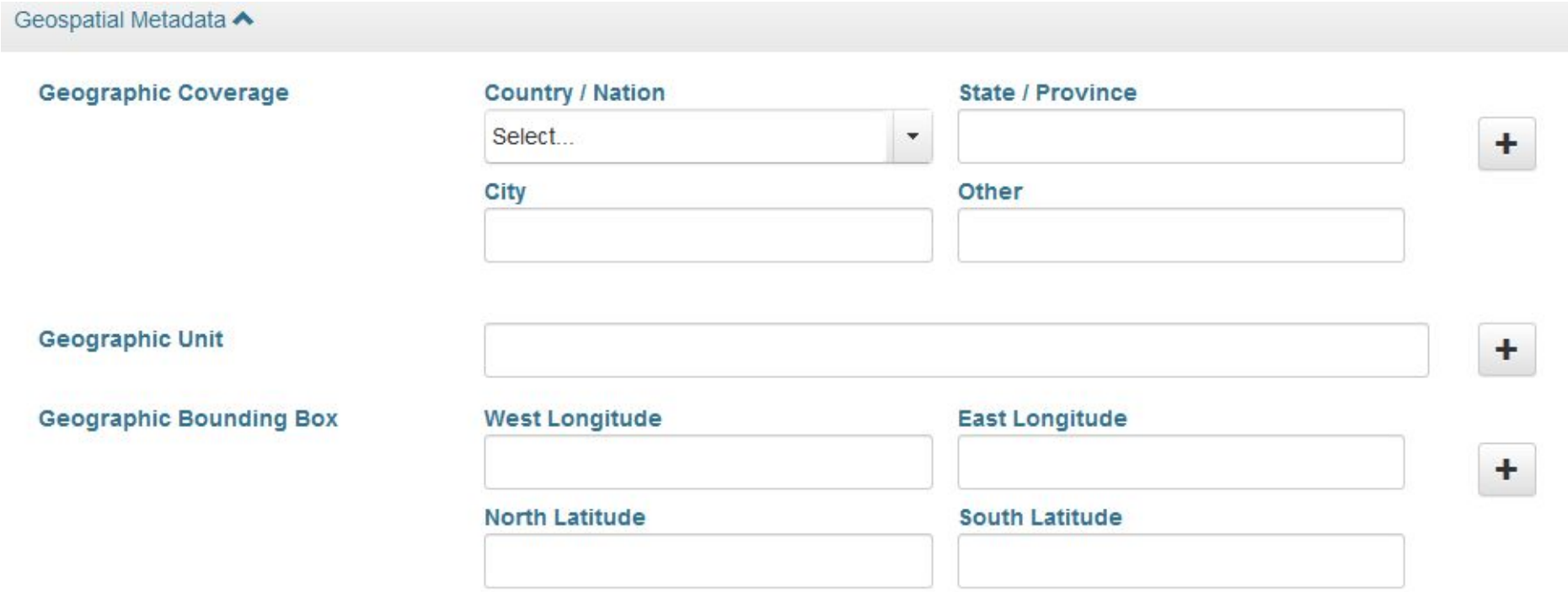
Bounding boxes in Geodisy

- Without bounding boxes, Geodisy cannot function
  - GeoBlacklight uses bounding boxes for discovery

- All datasets going through the Geodisy pipeline will either have bounding box coordinates entered into the metadata by the depositor or a bounding box that is programmatically generated by Geodisy using dataset files or other metadata

Dataverse metadata

- Uses different blocks of metadata: citation (basic description), geospatial, social science, astronomy, and life sciences
- The citation block includes several required fields for basic description
- For bounding boxes, Geodisy first attempts to analyze geospatial files using GDAL. If unsuccessful, it utilizes the contents of the geospatial block

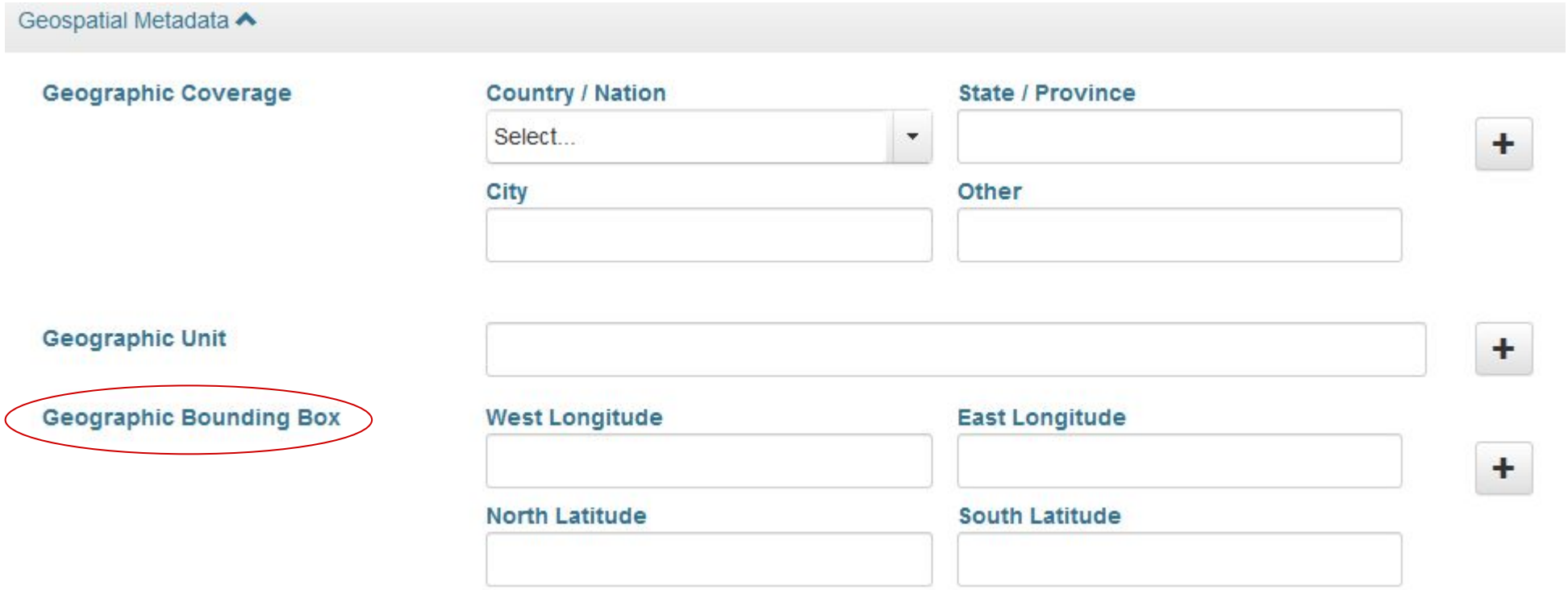


Automated system for generating bounding boxes (1)

- If the dataset contains geospatial type files, GDAL is used to generate coordinates

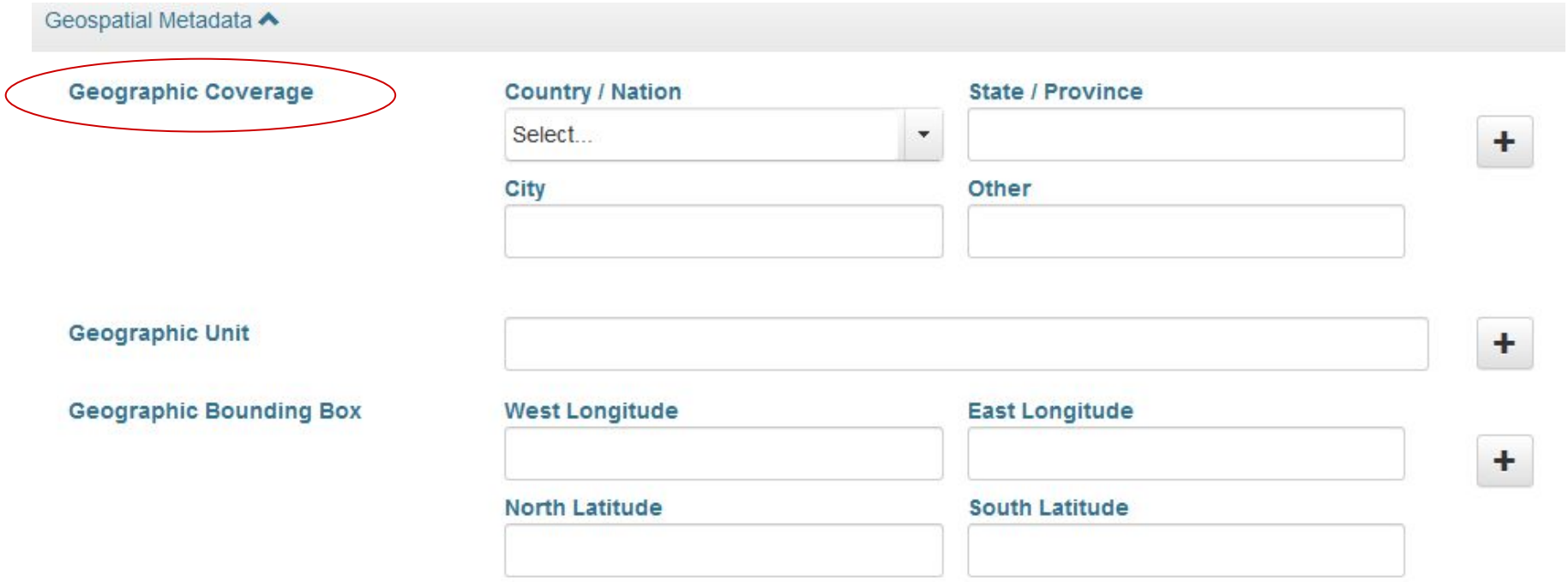
Automated system for generating bounding boxes (2)

- If Dataverse dataset includes geographic bounding coordinates and no usable geospatial files, it is used for Geodisy's bounding box



Automated system for generating bounding boxes (3)

- If coordinates are not included/invalid and there are no geospatial filetypes, geographic coverage metadata is sent to Geonames for coordinates. However, it must contain one of the following combinations:
  - Country/Nation
  - Country/Nation AND State/Province
  - Country/Nation AND State/Province AND City

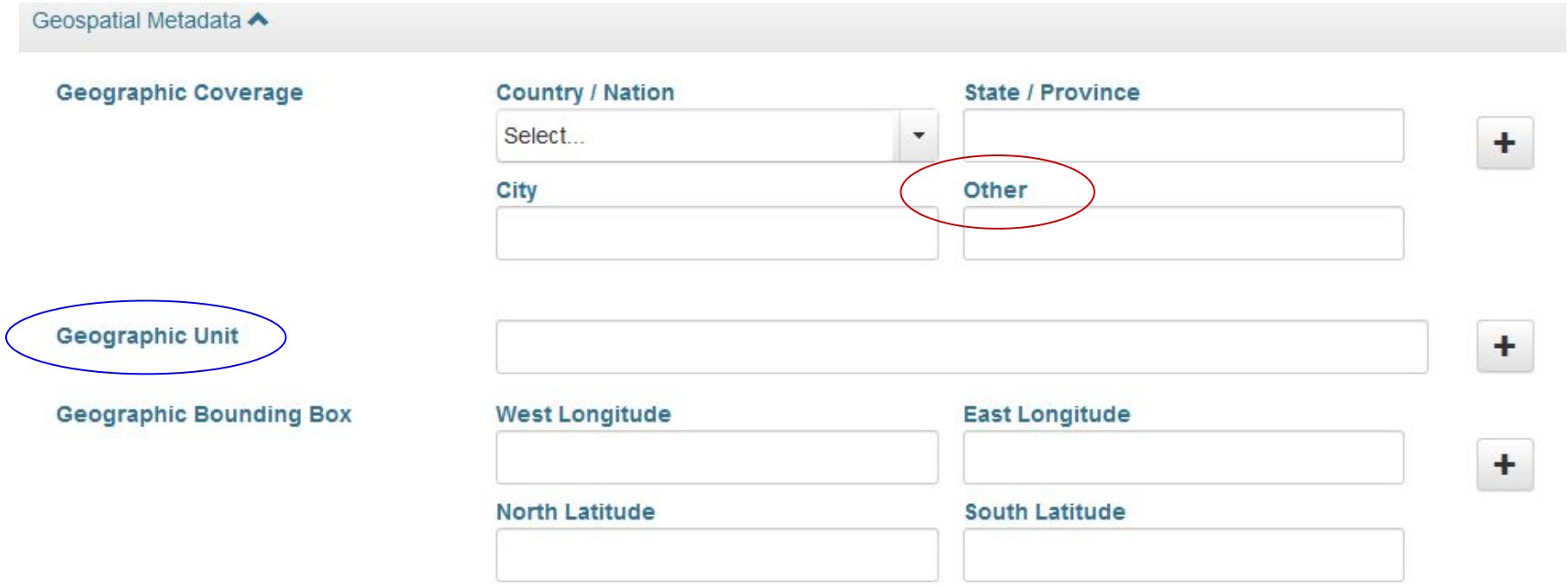


Automated system for generating bounding boxes (4)

- Datasets that do not include coordinates, geospatial filetypes, or geographic coverage metadata are ignored

Automated system for generating bounding boxes (5)

- Regardless of geographic coverage metadata, if bounding boxes cannot be generated by GDAL, datasets will be logged for manual review in some cases:
  - If the “other” geographic coverage field contains text
  - If the geographic coverage fields do not provide enough information
  - If the geographic coverage information does not find a valid match in Geonames
  - If Geonames finds a match but that match has no bounding box coordinates

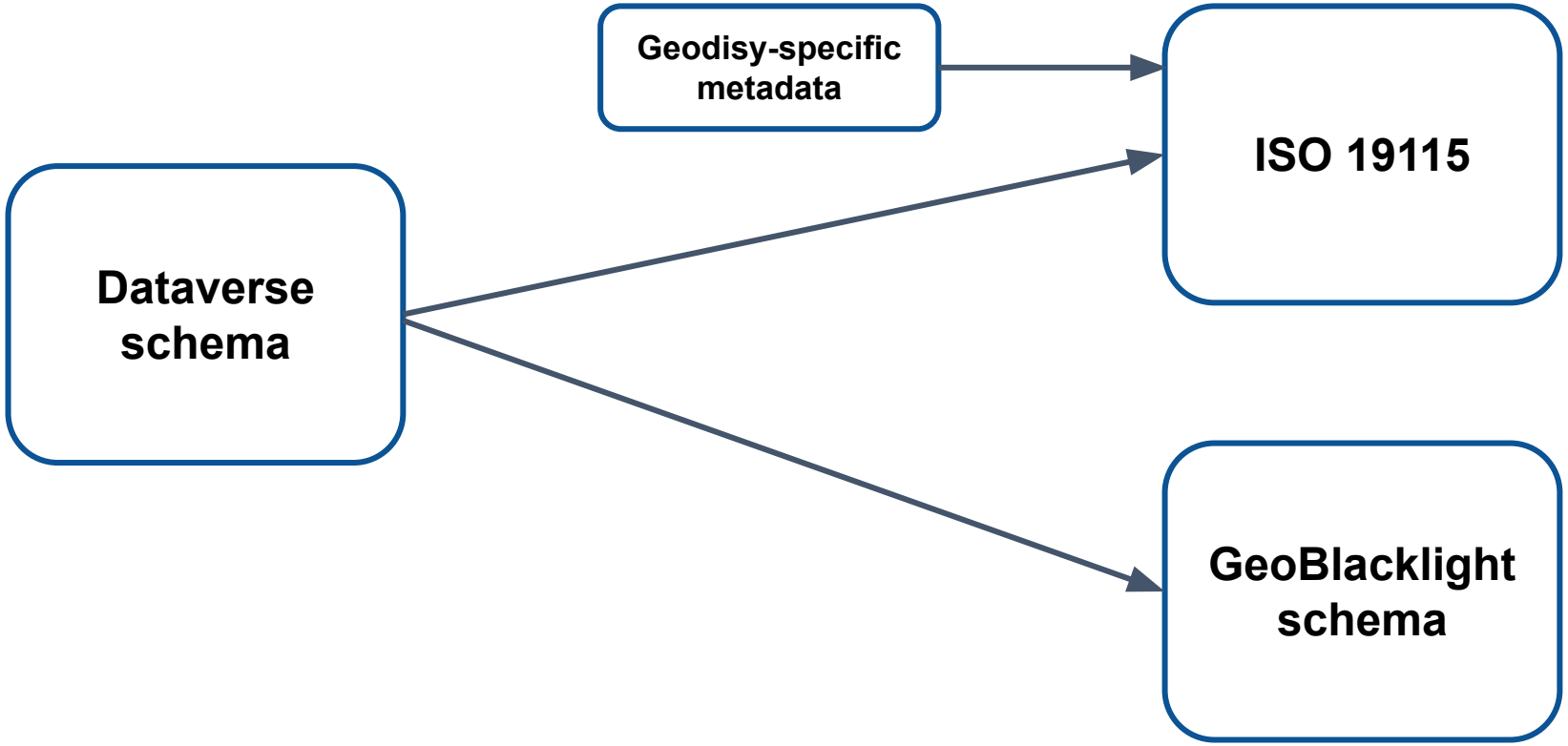


Metadata standards and crosswalks

- Metadata standard: a standardized series of metadata fields that describe a certain type of resource
  - e.g. Dublin Core for general resources, DDI for social science resources, Darwin Core for biology resources, etc

- Crosswalk: when the elements/fields from one metadata standard are mapped onto the elements/fields of another standard

Geodisy metadata treatment



Each dataset's ISO metadata will be available for download on GeoBlacklight

This is the metadata GeoBlacklight uses for functionality

Core Project Team (UBC)

- Eugene Barsky – Principal Investigator
- Paul Dante – Software Developer
- Edith Domingue – ARC Client Services Manager
- Mark Goodwin – Geospatial Metadata Coordinator
- Tang Lee – Project Manager
- Paul Lesack – Co-Principal Investigator
- Evan Thornberry – Co-Principal Investigator

Project Partners

- Jason Brodeur – McMaster University
- Marcel Fortin – University of Toronto
- Alex Garnett – SFU
- Amber Leahey – Scholars Portal
- Jason Hlady – University of Saskatchewan
- Venkat Mahadevan – UBC ARC
- Todd Trann – University of Saskatchewan
- Lee Wilson – Portage Network

Launching in spring 2020

Keep up to date:

#Geodisy on social media

researchdata.library.ubc.ca/find/geodisy

github.com/ubc-library/geodisy

Email: geodisy.info@ubc.ca