Introduction to Geodisy: An open-source spatial discovery platform

Eugene Barsky, Mark Goodwin, and Paul Dante

May 2020

Geodisy is an open source discovery tool that allows users to find open data from Canadian researchers visually, spatially, and quickly

- Search results are driven by an interactive map
- Scholars Portal Dataverse is the current source for data content, expansion will occur this year
- Contains data that is geospatial or simply associated with a location
- All components are open source and code is freely available for others to adapt


FRDR
Feedback
Log In
Help
About
EN
FRDR
FEDERATED RESEARCH DATA REPOSITORY
DFDR
DÉPÔT FÉDÉRÉ DE DONNÉES DE RECHERCHE
Find and Share Canadian Research Data
Search
Advanced search
Map Search (beta)
Deposit Data
Find Data
Search FRDR to find research datasets originating from researchers affiliated with Canadian institutions. Data deposited to other repositories across Canada can also be found by searching in FRDR. View the growing list of collaborating repositories.
Deposit Data
Any researcher affiliated with a Canadian institution can deposit data into FRDR. The platform can efficiently ingest datasets of any size, and preservation processing is done automatically. Data professionals from the Portage Network and institutions across Canada work with researchers to

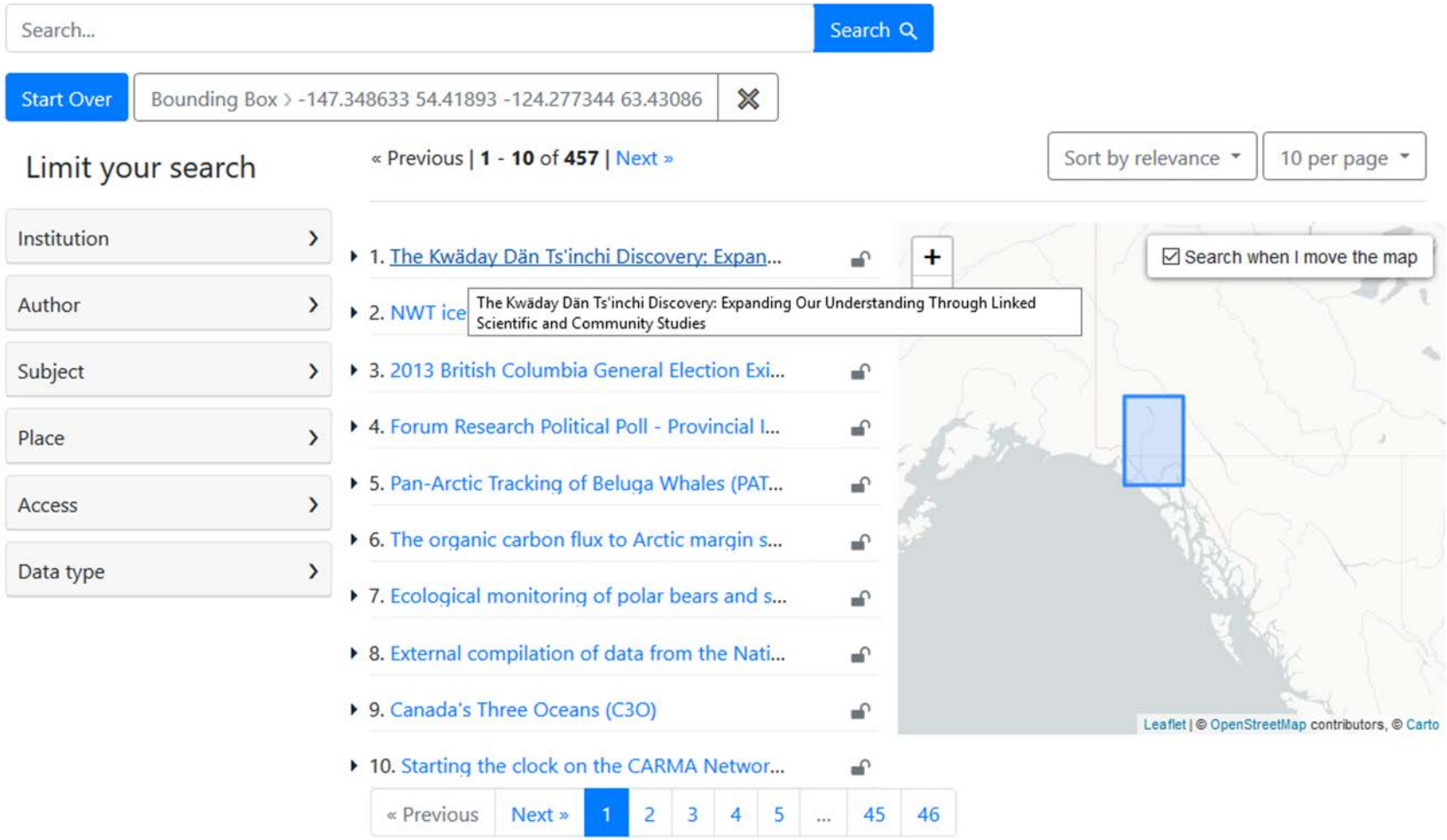
Why is it useful?

- 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 (Tatshenshini-Alsek ADJ2 park) OR (Glacier Bay ADJ2 (park or preserve)) OR (Kluane ADJ2 (park OR reserve)) OR (Atlin ADJ2 (park OR recreation area) OR …

Emoji image from https://emojipedia.org/apple/

Geodisy interface



Emoji image from https://emojipedia.org/apple/

Why is it important?

Geodisy allows for more open and effective research to better society

“FAIR Guiding Principles for scientific data management and stewardship” - published in Scientific Data, 2016. www.go-fair.org/fair-principles

FAIR: Findable, Accessible, Interoperable, Reusable

Geodisy supports FAIR Principles:

- Cleaning, enhancing, and standardizing of metadata
- Increasing discovery points for research data and metadata
- Adaptability of Geodisy for other groups
- Championing open data and open-source software

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

- **Dataverse:** Research data repository (Scholars Portal for phase 1)

  +

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

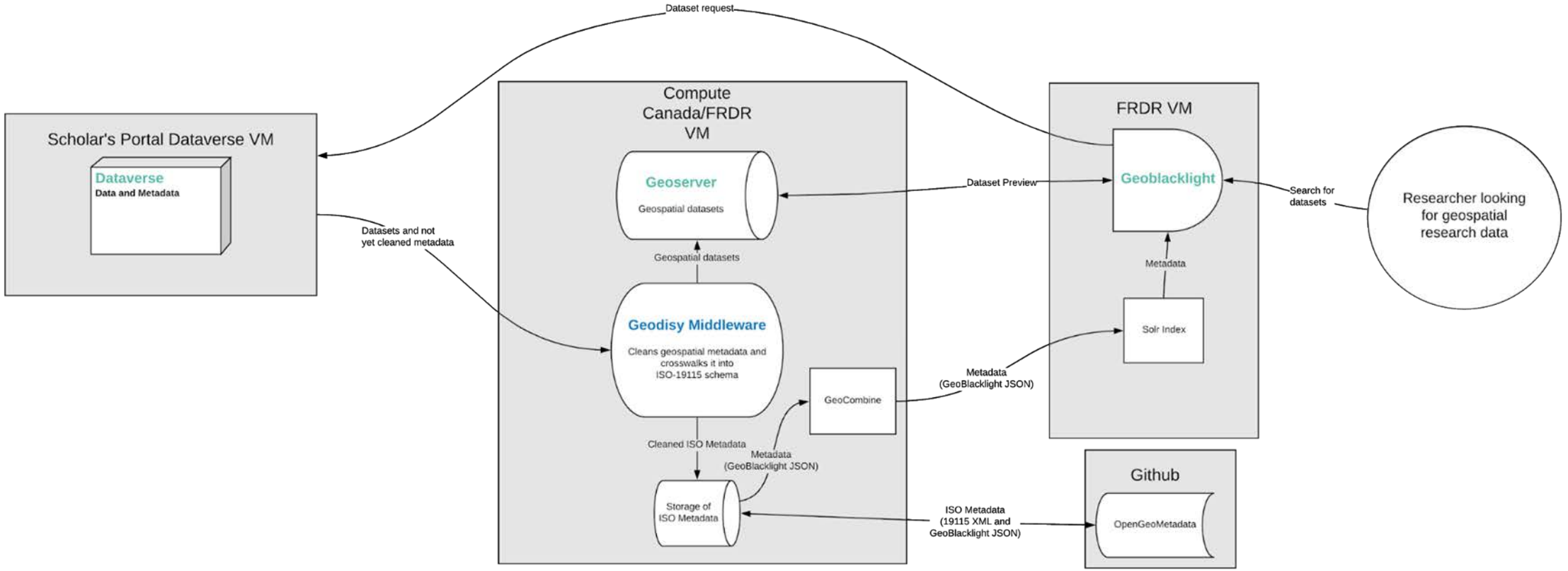
  +

- **GeoBlacklight:** Geospatial discovery layer

Project pipeline summary

1. Metadata is harvested from all **Scholars Portal** Dataverse with a geospatial component, and geospatial data files are harvested when possible (2,500 datasets from 500 Dataverses)
2. Bounding boxes are generated for each eligible dataset
3. Metadata is crosswalked to an international standard (ISO 19115) and GeoBlacklight JSON
4. Geospatial data files are deposited into **GeoServer**
5. All metadata is made publicly available on OpenGeoMetadata (Github)
6. Metadata and geospatial data files are harvested by **GeoBlacklight** for discovery

Geodisy Architecture

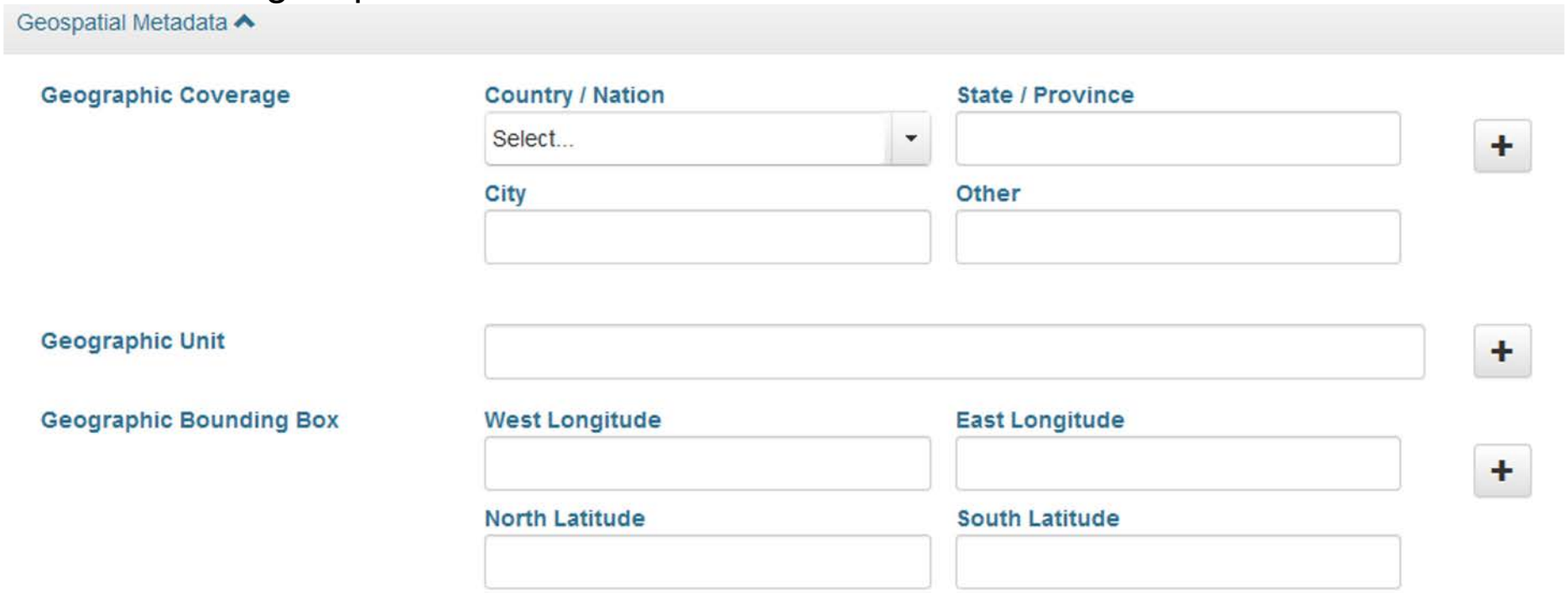


Bounding boxes in Geodisy

- Without bounding boxes, Geodisy cannot function - GeoBlacklight requires bounding boxes for discovery
- All datasets that go through the Geodisy pipeline 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
- Each geospatial file or set of coordinates (even if there are multiple in a single dataset) will get its own GeoBlacklight record
- Datasets that do not include coordinates, geospatial filetypes, or geographic coverage metadata are not eligible for inclusion

Dataverse metadata

- Dataverse 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
- Geodisy first attempts to analyze geospatial files using GDAL. If unsuccessful, it utilizes the contents of the geospatial block

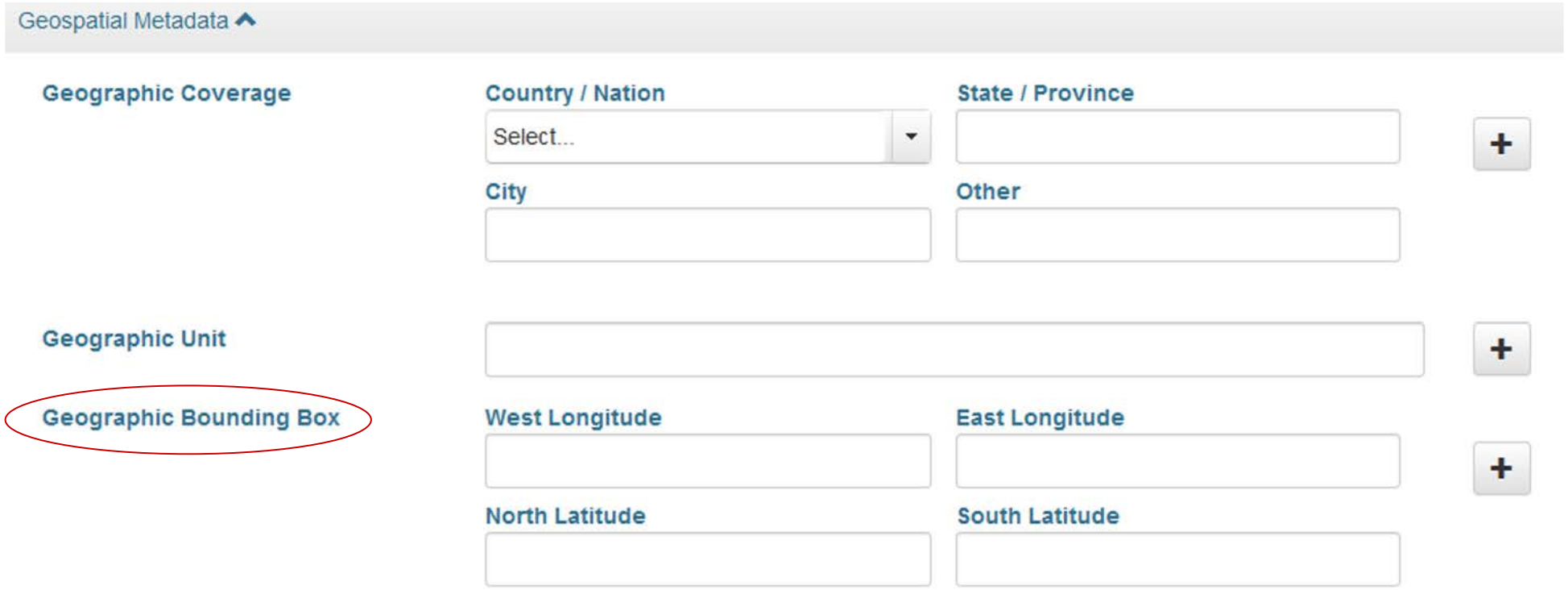


Automated system for generating bounding boxes (1)

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

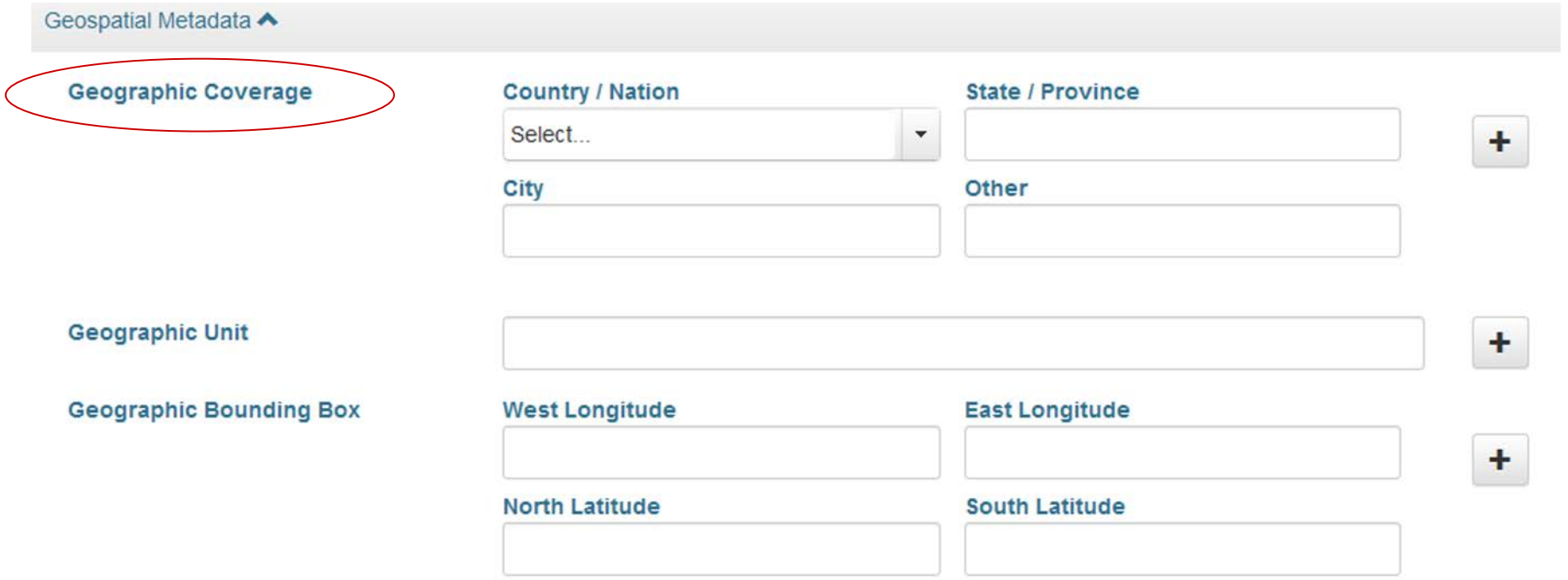
Automated system for generating bounding boxes (2)

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



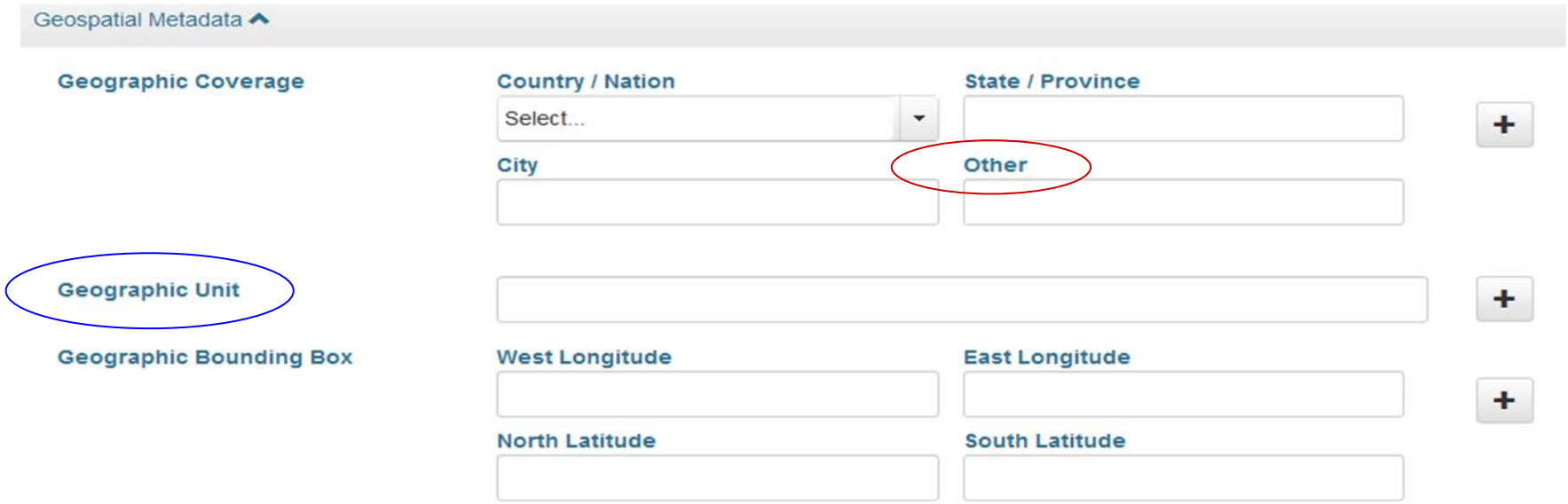
Automated system for generating bounding boxes (3)

- 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

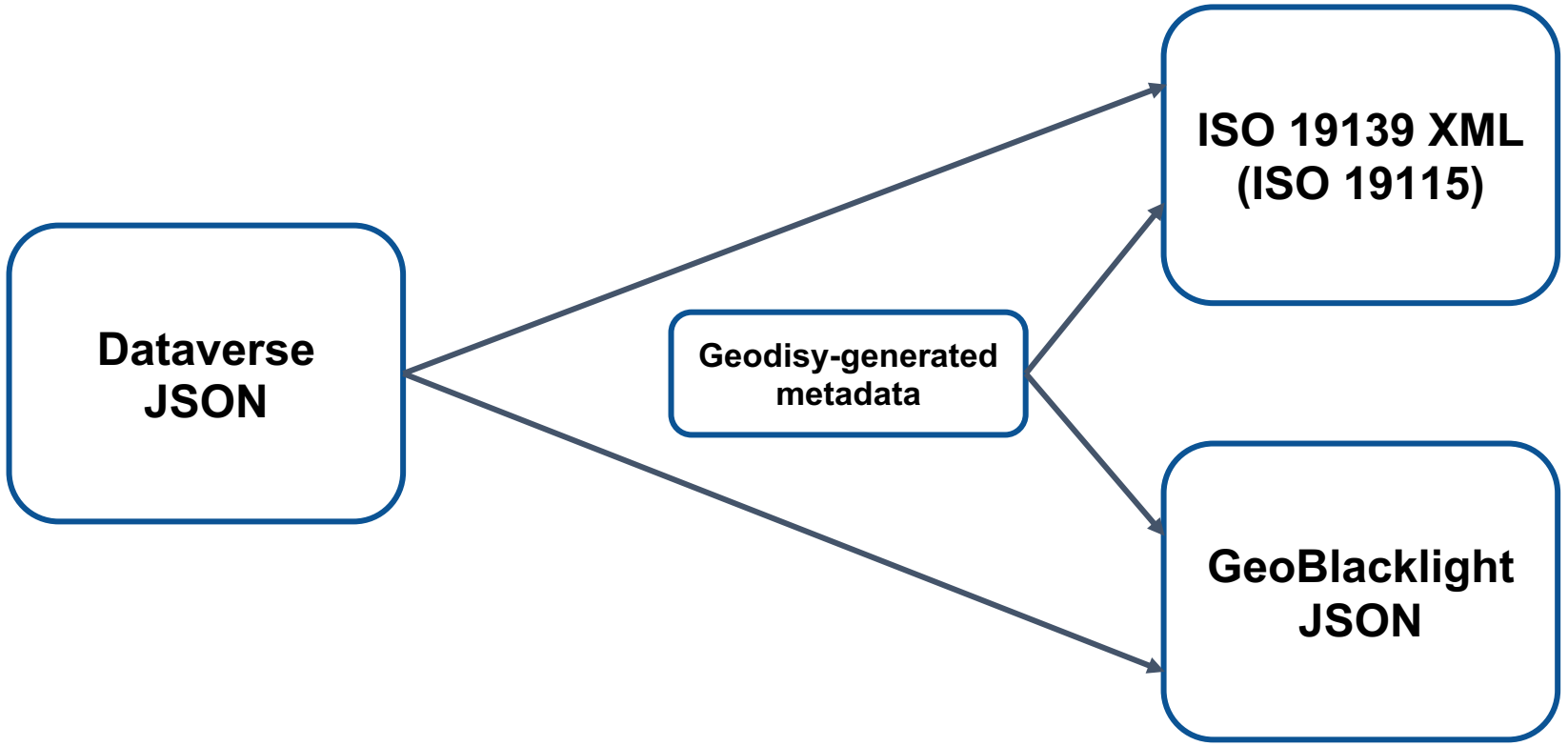


Automated system for generating bounding boxes (4)

- Regardless of geographic coverage metadata, if bounding boxes cannot be generated by GDAL datasets, they will be logged for manual review:
  - 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



Geodisy metadata treatment



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

This is the metadata GeoBlacklight uses for functionality

FRDR Geodisy demonstration

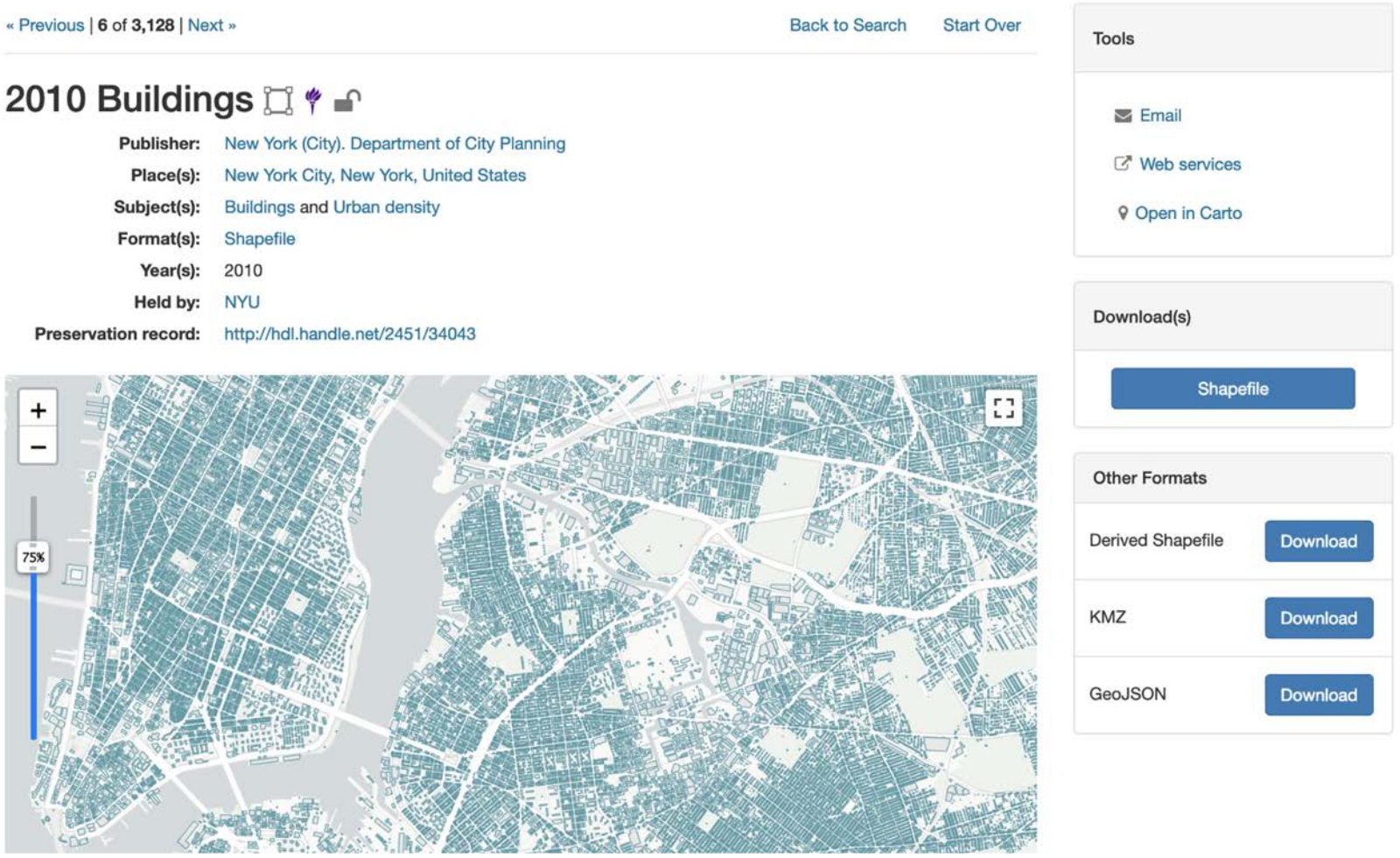
www.frdr-dfdr.ca

Next steps

- Funded by NDRIO for 2020-2021
- Integration with FRDR team
- Adding multiple new repositories: working with and improving the FRDR metadata feed
- Adding geospatial data preview
- French translation of interface

Next steps

- Geospatial preview (NYU Spatial Data Repository GeoBlacklight)



Project Team

- Eugene Barsky – Principal Investigator
- Mark Goodwin – Metadata Coordinator
- Paul Dante – Software Engineer

- Github: https://github.com/ubc-library/geodisy
- #geodisy on social media

Project Partners

- Paul Lesack – University of British Columbia
- Evan Thornberry – University of British Columbia
- Edith Domingue – University of British Columbia
- Tang Lee – University of British Columbia
- Venkat Mahadevan – University of British Columbia
- Jason Brodeur – McMaster University
- Amber Leahey – Scholars Portal

- Lee Wilson – Portage Network
- Kelly Stathis – Portage Network
- Neha Milan – Portage Network
- Alex Garnett – Simon Fraser University
- Todd Trann – University of Saskatchewan
- Joel Farthing – University of Saskatchewan
- Marcel Fortin – University of Toronto