GeoBTAA Metadata Handbook

None

Big Ten Academic Alliance Geospatial Information Network

2022

Table of contents

1. GeoBTAA Metadata Handbook

This handbook describes how to curate metadata records for the BTAA Geoportal.

1.1 Who is this for?

- Team Members in the Big Ten Academic Alliance Geospatial Information Network (BTAA-GIN)
- Development & Operations Staff in the BTAA-GIN
- Users & developers of open-source geospatial projects, such as OpenGeoMetadata and GeoBlacklight
- Users of the BTAA Geoportal

1.2 Contents:

1.2.1 GeoBTAA Metadata Profile

The GeoBTAA Metadata Profile combines the OpenGeoMetadata schema, local input guidelines, and custom elements.

1.2.2 Content Organization Model for the BTAA Geoportal

The Content Organization Model defines how records are organized and how they are related within the BTAA Geoportal.

1.2.3 Curation workflows

Step by step guides for selecting, submitting, harvesting, editing, publishing, and maintaining metadata records in the BTAA Geoportal

1.3 Version History

Changes for Version 4.4 (August 23, 2022)

- updated theme
- reorganized and expanded navigation menu
- new sections for Harvesting Guide and using GEOMG

Changes for Version 4.3 (August 15, 2022)

- migrate to MkDocs.org platform
- update bounding box entry guidelines
- add GEOMG page

Changes for Version 4.2 (March 24, 2022)

- New Entry and Usage Guidelines page
- Expands content organization model documentation
- Changes the name of the schema from 'Aardvark' to 'OpenGeoMetadata (version Aardvark)'
- Cleans up outdated links

Changes for Version 4.1 (Jan 2022)

- updates Status as optional; removes controlled vocabulary
- Clarifies relationship model

Changes for Version 4.0 (July 2021)

- Incorporation of GEOMG Metadata Editor
- Upgrade to Aardvark Metadata Schema for GeoBlacklight

Changes for version 3.3 (May 13, 2020)

- Added University of Nebraska
- Reorganized Metadata Elements to match editing template
- Updated the "Update the Collections" section to match new administrative process for tracking records

Changes for version 3.2 (Jan 8, 2020)

- Added Date Range element

Changes for version 3.1 (Dec 19, 2019)

- Added collection level records metadata schema

Changes for version 3 (Oct, 2019)

- GeoNetwork and Omeka deprecated
- all GeoBlacklight records are stored in a master spreadsheet in Google - Sheets
- records are transformed from CSV to GeoBlacklight JSON with a Python script
- additional metadata fields were added for administrative purposes
- IsPartOf field now holds a code pointing to the collection record
- Administrative groupings such as "State agencies geospatial data" are now subjects, not a Collection
- updated editing templates available
- all supplemental metadata can be stored as XML or HTML in project hosted folder
- updated links to collections database

1.4 Credits

Handbook prepared by:

- Karen Majewicz, Geospatial Product Manager
- Ziying (Gene) Cheng - Graduate Research Assistant

2. GeoBTAA Metadata Profile

2.1 Overview

The GeoBTAA Metadata Application Profile consists of the following components:

1. OpenGeoMetadata Elements

- The BTAA Geoportal uses the OpenGeoMetadata Schema for each resource.
- The current version of OpenGeoMetadata is called 'Aardvark'.
- This lightweight schema was designed specifically for the GeoBlacklight application and is geared towards discoverability.
- The GeoBTAA Metadata Profile aligns with all of the guidelines and recommendations in the official OpenGeoMetadata documentation.
- **The schema is documented on the OpenGeoMetadata website ↗.**

2. Local Input Guidelines for OpenGeoMetadata

- For a handful of elements, the GeoBTAA profile has more specific entry guidelines beyond what is documented in the OpenGeoMetadata schema.
- **See the Local Input Guidelines page for more detail**

3. Custom Elements

- The GeoBTAA profile includes custom fields for lifecycle tracking and administration
- These elements are generally added to the record by admin staff. When they appear on editing templates, they are grayed out.
- They all start with the namespace `b1g`
- **See the Custom Elements page for more detail**

Info

The GeoBTAA Metadata Template can be found at https://z.umn.edu/b1g-template

2.2 Local Input Guidelines

For the following elements, the GeoBTAA Metadata Profile has input guidelines beyond what is documented in the OpenGeoMetadata schema:

Title

Maps: The title for scanned maps is generally left as it was originally cataloged by a participating library. MARC subfields are omitted and can be inserted in the Description field.

Datasets: Harvested datasets often are lacking descriptive titles and may need to be augmented with place names. Dates may also be added to the end, but if the dataset is subject to updates, the data should be left off. Acronyms should be spelled out. The preferred format for dataset titles is: `Theme [place] {date}`. This punctuation allows for batch processing and splitting title elements.

Language

Although Language is optional in the OGM schema, a three-digit code is required for the BTAA Geoportal.

Creator

Spell all Acronyms out.

Publisher

Maps: Publisher values for maps are pulled from the original catalog record. Remove subfields for place names and dates.

Datasets: The BTAA Geoportal does not use the Publisher field for Datasets.

Provider

This is the name of the organization hosting the resources. If the organization is part of the BTAA library network, a university icon will display next to the resource's title. However, most Providers will not have an icon.

Bounding Box

On the Metadata Editing Template, provide Bounding Boxes in this format: **W,S,E,N**

This format will be programmatically converted to other formats when it is published to the Geoportal:

- The OpenGeoMetadata Bounding Box field (`dcat_bbox_s`) uses this order: `ENVELOPE(W,E,N,S)`
- The OpenGeoMetadata Geometry field (`locn_geometry`) uses a WKT format and the coordinate order will be converted to this layout: `POLYGON((W N, E N, E S, W S, W N))`
- The OpenGeoMetadata Centroid field (`dcat_centroid`) will be calculated to display longitude,latitude.

Example

Metadata CSV: **-120,10,-80,35**

converts to

`dcat_bbox_s: ENVELOPE(-120,-80,35,10)`

`locn_geometry: POLYGON((-120 35, -80 35, -80 10, -120 10, -120 35))`

`dcat_centroid: "22.5,-100.0"`

2.3 Custom Elements

This pages documents the custom metadata elements for the GeoBTAA Metadata Profile. These elements extend the official OpenGeoMetadata (Aardvark) schema.

b1g-id	Label	URI	Obligation
b1g-01	Code	`b1g_code_s`	Required
b1g-02	Status	`b1g_status_s`	Optional
b1g-03	Accrual Method	`b1g_dct_accrualMethod_s`	Required
b1g-04	Accrual Periodicity	`b1g_dct_accrualPeriodicity_s`	Optional
b1g-05	Date Accessioned	`b1g_dateAccessioned_s`	Required
b1g-06	Date Retired	`b1g_dateRetired_s`	Conditional
b1g-07	Child Record	`b1g_child_record_b`	Conditional
b1g-08	Mediator	`b1g_dct_mediator_sm`	Conditional
b1g-09	Access	`b1g_access_s`	Conditional
b1g-10	Image	`b1g_image_ss`	Optional
b1g-11	GeoNames	`b1g_geonames_sm`	Optional
b1g-12	Publication State	`b1g_publication_state_s`	Required
b1g-13	Language String	`b1g_language_sm`	Required
b1g-14	Creator ID	`b1g_creatorID_sm`	Optional

2.3.1 Code

Label	Code
URI	`b1g_code_s`
Profile ID	b1g-01
Obligation	Required
Multiplicity	1-1
Field type	string
Purpose	To group records based upon their source
Entry Guidelines	Codes are developed by the metadata coordinator and indicate the provider, the type of institution hosting the resources, and a numeric sequence number. For more details, see Code Naming Schema
Commentary	This administrative field is used to group and track records based upon where they are harvested. This is frequently an identical value to "Member Of". The value will differ for records that are retired (these are removed from "Member Of") and records that are part of a subcollection.
Controlled Vocabulary	yes-strict
Example value	12d-01
Element Set	B1G

2.3.2 Status

Label	**Status**
URI	`b1g_status_s`
Profile ID	b1g-02
Obligation	Optional
Multiplicity	0-1
Field type	string
Purpose	To indicate if a record is currently active, retired, or unknown. It can also be used to indicate if individual data layers from website has been indexed in the Geoportal.
Entry Guidelines	Plain text string is acceptable
Commentary	This is a legacy admin field that was previously used to track published vs retired items. The current needs are still TBD.
Controlled Vocabulary	no
Example value	Active
Element Set	B1G

2.3.3 Accrual Method

Label	**Accrual Method**
URI	`b1g_dct_accrualMethod_s`
Profile ID	b1g-03
Obligation	Required
Multiplicity	1-1
Field type	string
Purpose	To describe how the record was obtained
Entry Guidelines	Some values, such as "ArcGIS Hub" should be entered consistently. Others may be more descriptive, such as "Manually entered from text file."
Commentary	This allows us to find all of the ArcGIS records in one search. It also can help track records that have been harvested via different methods within the same collection.
Controlled Vocabulary	no
Example value	ArcGIS Hub
Element Set	B1G/ Dublin Core

2.3.4 Accrual Periodicity

Label	**Accrual Periodicity**
URI	`b1g_dct_accrualPeriodicity_s`
Profile ID	b1g-04
Obligation	Optional
Multiplicity	0-1
Field type	string
Purpose	To indicate how often a collection is reaccessioned
Entry Guidelines	Enter one of the following values: Annually, Semiannually, Quarterly, Monthly, As Needed
Commentary	This field is primarily for collection level records.
Controlled Vocabulary	yes-not strict
Example value	As Needed
Element Set	B1G/ Dublin Core

2.3.5 Date Accessioned

Label	**Date Accessioned**
URI	`b1g_dateAccessioned_s`
Profile ID	b1g-05
Obligation	Required
Multiplicity	1-1
Field type	string
Purpose	To store the date a record was harvested
Entry Guidelines	Enter the date a record was harvested OR when it was added to the geoportal using the format yyyy-mm-dd
Commentary	This field allows us to track how many records are ingested into the portal in a given time period and to which collections.
Controlled Vocabulary	no
Example value	2021-01-01
Element Set	B1G

2.3.6 Date Retired

Label	**Date Retired**
URI	`b1g_dateRetired_s`
Profile ID	b1g-06
Obligation	Conditional
Multiplicity	0-1
Field type	string
Purpose	To store the date the record was removed from the geoportal public interface
Entry Guidelines	Enter the date a record was removed from the geoportal
Commentary	This field allows us to track how many records have been removed from the geoportal interface by time period and collection.
Controlled Vocabulary	no
Example value	2021-01-02
Element Set	B1G

2.3.7 Child Record

Label	**Child Record**
URI	`b1g_child_record_b`
Profile ID	b1g-07
Obligation	Optional
Multiplicity	0-1
Field type	string boolean
Purpose	To apply an algorithm to the record that causes it to appear lower in search results
Entry Guidelines	Only one of two values are allowed: true or false
Commentary	This is used to lower a record's placement in search results. This can be useful for a large collection with many similar metadata values that might clutter a user's experience.
Controlled Vocabulary	string boolean
Example value	true
Element Set	B1G

2.3.8 Mediator

Label	**Mediator**
URI	`b1g_dct_mediator_sm`
Profile ID	b1g-08
Obligation	Conditional
Multiplicity	0-0 or 1-*
Field type	string
Purpose	To indicate the universities that have licensed access to a record
Entry Guidelines	The value for this field should be one of the names for each institution that have been coded in the GeoBlacklight application.
Commentary	This populates a facet on the search page so that users can filter to only databases that they are able log into based upon their institutional affiliation.
Controlled Vocabulary	yes
Example value	University of Wisconsin-Madison
Element Set	B1G/ Dublin Core

2.3.9 Access

Label	**Access**
URI	`b1g_access_s`
Profile ID	b1g-09
Obligation	Conditional
Multiplicity	0-0 or 1-1
Field type	string JSON
Purpose	To supply the links for restricted records
Entry Guidelines	The field value is an array of key/value pairs, with keys representing an insitution code and values the URL for the library catalog record. See the Access Template for entry.
Commentary	This field is challenging to construct manually, is it is a JSON string of institutional codes and URLs. The codes are used instead of the actual names in order to make the length of the field more manageable and to avoid spaces.
Controlled Vocabulary	no
Example value	{\"03\":\"https://purl.lib.uiowa.edu/PolicyMap\",\"04\":\"https://www.lib.umd.edu/dbfinder/id/UMD09180\",\"05\":\"https://primo.lib.umn.edu/permalink/f/1q7ssba/UMN_ALMA51581932400001701\",\"06\":\"http://catalog.lib.msu.edu/record=b10238077~S39a\",\"07\":\"https://search.lib.umich.edu/databases/record/39117\",\"09\":\"https://libraries.psu.edu/databases/psu01762\",\"10\":\"https://digital.library.wisc.edu/1711.web/policymap\",\"11\":\"https://library.ohio-state.edu/record=b7869979~S7\"}
Element Set	B1G

2.3.10 Image

Label	**Image**
URI	`b1g_image_ss`
Profile ID	b1g-10
Obligation	Optional
Multiplicity	0-0 or 0-1
Field type	stored string (URL)
Purpose	To show a thumbnail on the search results page
Entry Guidelines	Enter an image file using a secure link (https). Acceptable file types are JPEG or PNG
Commentary	This link is used to harvest an image into the Geoportal server for storage and display. Once it has been harvested, it will remain in storage, even if the orginal link to the image stops working.
Controlled Vocabulary	no
Example value	https://gis.allencountyohio.com/GIS/Image/countyseal.jpg
Element Set	B1G

2.3.11 GeoNames

Label	**GeoNames**
URI	`b1g_geonames_sm`
Profile ID	b1g-11
Obligation	Optional
Multiplicity	0-*
Field type	stored string (URI)
Purpose	To indicate a URI for a place name from the GeoNames database
Entry Guidelines	Enter a value in the format "http://sws.geonames.org/`URI`"
Commentary	This URI provides a linked data value for one or more place names. It is optional as there is currently no functionality tied to it in the GeoBlacklight application
Controlled Vocabulary	yes
Example value	https://sws.geonames.org/2988507
Element Set	B1G

2.3.12 Publication State

Label	**Publication State**
URI	`b1g_publication_state_s`
Profile ID	b1g-12
Obligation	Required
Multiplicity	1-1
Field type	string
Purpose	To communicate to Solr if the item is public or hidden
Entry Guidelines	Use the dropdown or batch editing functions to change the state
Commentary	When items are first added to GEOMG, they are set as Draft by default. When they are ready to be published, they can be manually changed to Published. If the record is retired or needs to be hidden, it can be changed to Unpublished
Controlled Vocabulary	yes
Example value	Draft
Element Set	B1G

2.3.13 Language string

Label	**Language string**
URI	`b1g_language_sm`
Profile ID	b1g-13
Obligation	Required
Multiplicity	1-*
Field type	string
Purpose	To display the spelled out string (in English) of a language code to users
Entry Guidelines	This field is automatically generated from the Language field in the main form
Commentary	The OGM schema specified using a 3-digit code to indicate lanuage. In order to display this to users, it needs to be translated into a human-readable string.
Controlled Vocabulary	yes
Example value	French
Element Set	B1G

2.3.14 Creator ID

Label	**Creator ID**
URI	`b1g_creatorID_sm`
Profile ID	b1g-14
Obligation	Optional
Multiplicity	0-*
Field type	string
Purpose	To track the URI of a creator value
Entry Guidelines	This field is entered as a URI representing an authority record
Commentary	These best practices recommend consulting one or two name registries when deciding how to standardize names of creators: the Faceted Application of Subject Terminology (FAST) or the Library of Congress Name Authority File (LCNAF). FAST is a controlled vocabulary based on the Library of Congress Subject Headings (LCSH) that is well-suited to the faceted navigation of the Geoportal. The LCNAF is an authoritative list of names, events, geographic locations and organizations used by libraries and other organizations to collocate authorized creator names to make searching and browsing easier.
Controlled Vocabulary	yes
Example value	fst02013467
Element Set	B1G

2.4 Supplemental Metadata

All other forms of metadata, such as ISO 19139, FGDC Content Standard for Digital Geospatial Metadata, attribute table definitions, or custom schemas are treated as **Supplemental Metadata**.

- Supplemental Metadata is not usually edited directly for inclusion in the project.
- If this metadata is available as XML or HTML, it can be added as a hosted link for the Metadata preview tool in GeoBlacklight.
- XML or HTML files can be parsed to extract metadata that forms the basis for the item’s GeoBlacklight schema record.
- The file formats that can be viewed within the geoportal application include:

- ISO 19139 XML
- FGDC XML
- MODS XML
- HTML (any standard)

2.5 Editing Template

The GeoBTAA Metadata Template is a set of spreadsheets that are formatted for upload to our metadata editor, GEOMG.

Users will need to make a copy of the spreadsheet to use for editing. In some cases, the Metadata Coordinator can provide a customized version of the sheets for specific collections.

The Template contains the following tabs:

- Map Template
- Dataset Template
- Website Record Template
- Values: All of the controlled vocabulary values for the associated fields.

Info

The GeoBTAA Metadata Template can be found at https://z.umn.edu/b1g-template

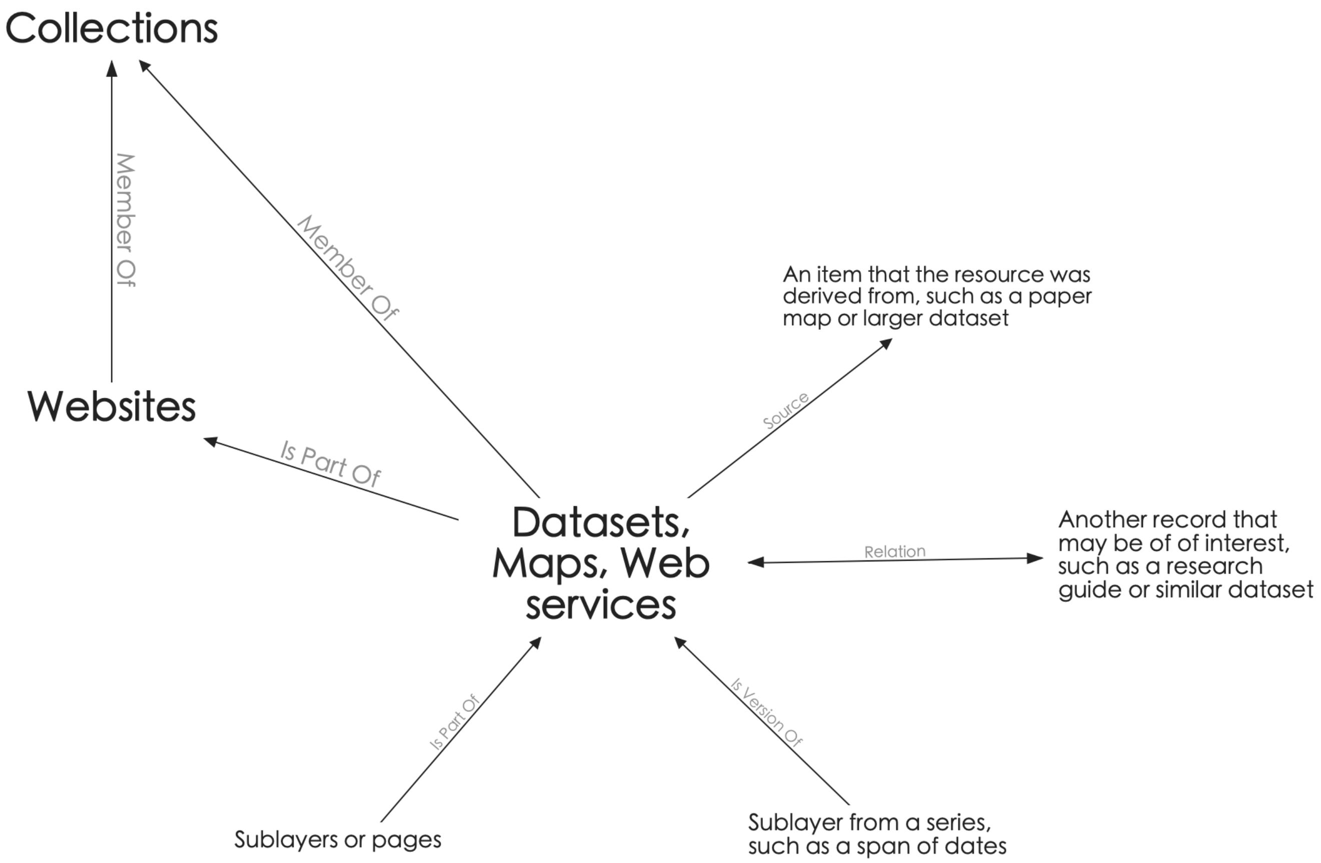
3. Content Organization Model

3.1 Content Organization Model

GeoBlacklight organizes records with a network model rather than with a hierarchical model. It is a flat system whereby every database entry is a "Layer" and uses the same metadata fields. Unlike many digital library applications, it does not have different types of records for entities such as "communities," "collections," or "groups." As a result, it does not present a breadcrumb navigation structure, and all records appear in the same catalog directory with the URL of https:geo.btaa.org/catalog/`ID`.

Instead of a hierarchy, GeoBlacklight relates records via metadata fields. These fields include `Member Of`, `Is Part Of`, `Is Version Of`, `Source`, and a general `Relation`. This flexibility allows records to be presented in several different ways. For example, records can have multiple parent/child/grandchild/sibling relationships. In addition, they can be nested (i.e., a collection can belong to another collection). They can also connect data layers about similar topics or represent different years in a series.

The following diagram illustrates how the BTAA Geoportal organizes records. The connecting arrow lines indicate the name of the relationship. The labels reflect each record's Resource Class (**Collections**, **Websites**, **Datasets**, **Maps**, **Web services**).


Collections
Member Of
Member Of
Websites
Is Part Of
An item that the resource was derived from, such as a paper map or larger dataset
Source
Datasets, Maps, Web services
Relation
Another record that may be of interest, such as a research guide or similar dataset
Is Part Of
Is Version Of
Sublayers or pages
Sublayer from a series, such as a span of dates

3.2 Resource Classes

3.2.1 Collections

The BTAA Geoportal interprets the Resource Class, **Collections**, as top-level, custom groupings. These reflect our curation activities and priorities.

Other records are linked to Collections using the `Member Of` field. The ID of the parent record is added to the child record only. View all of the current **Collections** in the geoportal at this link: https://geo.btaa.org/?f%5Bgbl_resourceClass_sm%5D%5B%5D=Collections

3.2.2 Websites

The BTAA Geoportal uses the Resource Class, **Websites**, to create parent records for data portals, digital libraries, dashboards, and interactive maps. These often start off as standalone records. Once the items in a website have been indexed, they will have child records.

Individual **Datasets**, **Maps**, or **Web service**s are linked to the **Website** they came from using the `Is Part Of` field. The ID of the parent record is added to the child record only.

View all of the current **Websites** in the geoportal at this link: https://geo.btaa.org/?f%5Bgbl_resourceClass_sm%5D%5B%5D=Websites

3.2.3 Datasets, Maps, and Web services

The items in this Resource Class represent individual data layers, scanned map files, and/or geospatial web services. (Some items may have multiple Resource Classes attached to the same record.)

This item class is likely to have the most relationships specified in the metadata. A typical **Datasets** record might have the following:

1. `Member Of` a **Collections** record
2. `Is Part Of` a **Websites** record
3. If the data was digitized from a paper map in the geoportal, it can be linked to the **Maps** record via the `Source` relation
4. a general `Relation` to a research guide or similar dataset

3.2.4 Multipart Items

Many items in the geoportal are multipart. There may be individual pages from an atlas, sublayers from a larger project, or datasets broken up into more than one download. In these cases, the `Is Part Of` field is used.

As a result, these items may have multiple `Is Part Of` relationships- (1) the parent for the multipart items and (2) the original website.

4. Curation Workflows

4.1 Identifying Resources

Identifying Resources to Submit to the BTAA Geoportal

The BTAA Geoportal holds metadata records that point to geospatial data, maps, aerial imagery, and websites hosted online by external organizations. It is the role of the Team members to seek out new content for the geoportal. Regional or national collections will be selected by the Collections Steering Group.

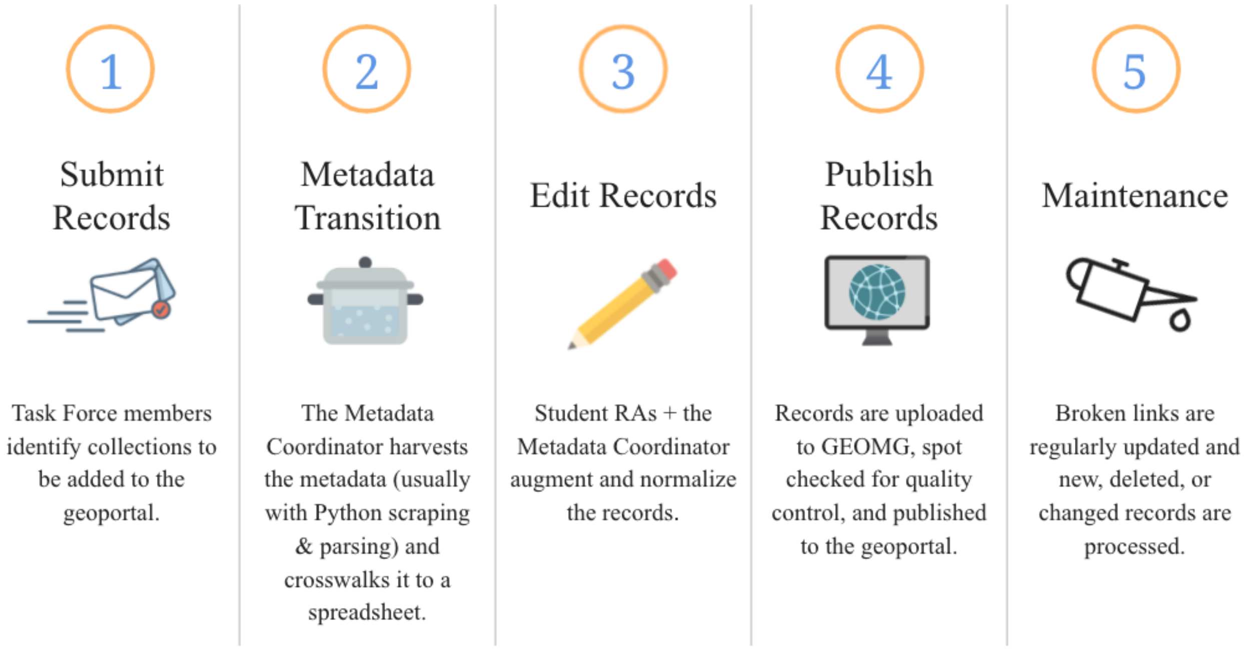
Review the Collection Strategy and the Collection Development Policy for more details.

Places to find public domain collections

- State GIS clearinghouses
- State agencies (especially DNRs and DOTs)
- County or city GIS departments
- Library digital collections
- Research institutes
- Nonprofit organizations

4.2 Resource Lifecycle

⚠ This guide is a work in progress (August 2022)



1 Submit Records

Task Force members identify collections to be added to the geoportal.

2 Metadata Transition

The Metadata Coordinator harvests the metadata (usually with Python scraping & parsing) and crosswalks it to a spreadsheet.

3 Edit Records

Student RAs + the Metadata Coordinator augment and normalize the records.

4 Publish Records

Records are uploaded to GEOMG, spot checked for quality control, and published to the geoportal.

5 Maintenance

Broken links are regularly updated and new, deleted, or changed records are processed.

4.2.1 1. Submit Records

1. Optional: Contact the organization

- Use this template to inform the organization that we plan to include their resources in our geoportal.
- If metadata for the resources are not readily available, the organization may be able to send them to you.

2. Send an email to the Metadata Coordinator

Things to include:

- a link to the website
- Title and Description of the collection
- (If known) information about how to harvest the metadata or construct access links

3. The submission will be added to our collections processing queue

Metadata processing tasks are tracked on our public GitHub project dashboard.

4.2.2 2. Metadata Transition

This stage involves batch processing of the records, including harvesting, transformations, crosswalking information. This stage is carried out by the Metadata Coordinator, who may contact Team members for assistance.

ℹ See our Harvest Guide for more information on formats and techniques for harvesting metadata

Regardless of the method used for acquiring the metadata, it is always transformed into a spreadsheet for editing. These spreadsheets are uploaded to GEOMG Metadata Editor.

Because of the variety of platforms and standards, this process can take many forms. The Metadata Coordinator will contact Team members if they need to supply metadata directly.

4.2.3 3. Edit Records

Once the metadata is in spreadsheet form, it is ready to be normalized and augmented. UMN Staff will add template information and use spreadsheet functions or scripts to programmatically complete the metadata records.

- The GBL Metadata Template is for creating GeoBlacklight metadata.
- Refer to the documentation for the Aardvark fields and the B1G profile fields for guidance on values and formats.

4.2.4 4. Publish Records

Once the editing spreadsheets are completed, UMN Staff uploads the records to `GEOMG`, a metadata management tool. GEOMG validates records and performs any needed field transformations. Once the records are satisfactory, they are published and available in the BTAA Geoportal.

Read more on the GEOMG documentation page.

4.2.5 5. Maintenance

General Maintenance

All project team members are encouraged to review the geoportal records assigned to their institutions periodically to check for issues. Use the feedback form at the top of each page in the geoportal to report errors or suggestions. This submission will include the URL of the last page you were on, and it will be sent to the Metadata Coordinator.

Broken Links

The geoportal will be programmatically checked for broken links on a monthly basis. Systematic errors will be fixed by UMN Staff. Some records from this report may be referred back to Team Members for investigating broken links.

Subsequent Accessions

- Portals that utilize the DCAT metadata standard will be re-accessioned monthly.
- Other GIS data portals will be periodically re-accessioned by the Metadata Coordinator at least once per year.
- Team members may be asked to review this work and provide input on decisions for problematic records.

Retired Records

When an external resource has been moved, deleted, or versioned to a new access link, the original record is retired from the BTAA Geoportal. This is done by converting the Publication State of the record from 'Published' to 'Unpublished'. The record is not deleted from the database and can still be accessed via a direct link. However, it will not show up in any search queries.

4.3 Harvest Guide

⚠ **This guide is a work in progress (August 2022)**

4.3.1 Harvesting Options

The BTAA Geoportal holds metadata records that point to geospatial data, maps, aerial imagery, web services, and websites hosted online by external organizations. This metadata is acquired by harvesting from the organization. Here are the most common harvesting methods:

API Harvesting or HTML Parsing

Most data portals have APIs or HTML structures that can be programmatically parsed to obtain metadata for each record.

- DCAT enabled portals: ArcGIS Open Data Portals (HUB), Socrata portals, and some others share metadata in the DCAT standard.
- CKAN / DKAN portals: This application uses a custom metadata schema for their API.
- HTML Parsing: If a data portal or website does not have an API, we may be able to parse the HTML pages to obtain the metadata needed to create GeoBlacklight schema records. This is done using custom View our harvesting scripts for HTML parsing here.

Individual Geospatial Metadata Standard files

Geospatial metadata standards are expressed in the XML format, which can be parsed to extract metadata needed to create GeoBlacklight schema records. The following file types are accepted for metadata extraction and can serve as Supplemental Metadata:

- **ISO 19139 XML and FGDC XML files**: They are parsed to extract metadata values for GeoBlacklight metadata using the project created Python scripts found in BTAA-Geospatial-Data-Project/parse-xml
- **ArcGIS 1.0 Metadata XML files**: These records are transformed to ISO 19139 using XSLT. They are then treated the same as the ISO as described above.

Downloading Data

Some metadata is only available as part of a zipped download of the datasets. In this case, UMN staff will use scripts to batch download the records, unzip them, and process their metadata locally.

OAI-PMH

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a framework that can be used to harvest metadata records from enabled repositories. The records are usually available as a simple Dublin Core XML format. If the protocol is not set up to include extra fields, such as the map image's download link or bounding box, this method may not be sufficient on its own.

Spreadsheets

Administrators for university repositories for scanned maps or data can often export metadata into a spreadsheet, especially for Dublin Core. This method is preferred, because the University IT professionals and librarians can control which fields to export, and because transformations by the Metadata Coordinator are not necessary. The B1G Map Template shows all of the fields needed for GeoBlacklight.

MARC files The best way to transfer this type of metadata is to send a single file containing multiple records in the .MRC or MARC XML format. The Metadata Coordinator will use MarcEdit or XML parsing to transform the records to GeoBlacklight using the project-designated MARC to GBL crosswalk, but Team Members can specify preferences.

4.3.2 DCAT Harvesting

DCAT stands for Data Catalog, which is a standard schema and vocabulary to organize datasets and data services. This is typically accessed by appending "/ data.json" to the end of the portal's base URL. View our harvesting scripts for DCAT enabled portals here.

We normally harvest the DCAT portals once a month to compare the JSON files with latest reaccession, add the new records into our geoportal, and also retire the deleted records as well. By saying records, we only want to harvest data types that are either **Datasets** or **Imagery**.

GitHub Repository:

This repository stores up-to-date **harvest script** as well as the **basic portal information**. It also stores the **historic harvest data** and **harvest report** (without manual edits). Remember to pull down the latest repo to your local machine before regular reaccession. Here're some import files or folders:

- arcPortals.csv includes the details about each portal that will be used for metadata construction every time.
- harvest.py is written in Python and mainly does the following jobs:

a. Request the JSON by portal and store them to local folder jsons

b. For each portal, compare the new JSON with the latest one from last reaccession, and output the new records as well as the deleted records.

- For new records, create the metadata by following the dcat metadata format and write into one CSV report called `allNewItems_yyyymmdd.csv` and store in the folder reports.
- For deleted records, we only need to know their IDs rather the metadata, so that write all retired records' ID into one CSV called `allDeletedItems_yyyymmdd.csv` and store in reports as well.

c. Test the download links from `allNewItems_yyyymmdd.csv` to check if valid. If it is valid, calculate and add the file size for this new record. Test again if it is not. Leave a message in the Title field for manual test later.

d. Populate spatial coverage based on bounding box for each record in `allNewItems_yyyymmdd.csv`.

e. Check duplicated records by same tile and bounding box, or same ID. If duplicates exist, leave a message in the Title field for manual delete later.

- retired_items_GEOMG.ipynb is to retire the deleted records on GEOMG, change them to unpublished and set the retire date by Python script.

There are still some manual work on harvest report, so that we need to upload them on google drive after creating a new folder for each harvest. Open the `allNewItems_yyyymmdd.csv` with google sheets and do following things: 1. Check whether there are message added in **Title** field. 2. Check whether the auto-populated Title has the right format of `AlternativeTitle [titleSource] {year if it exists}` 3. Choose a proper Theme value for each record.

4.3.3 CKAN Harvesting

CKAN (Comprehensive Knowledge Archive Network) is an open-source data management system (DMS) for powering data hubs and data portals. Different from DCAT, we cannot access the JSON data by adding the `/data.json` after the portal's URL.

Instead, we need to request the **package URL** for each CKAN portal to get a JSON-formatted list of **resource names**. By comparing with the latest reaccession, we find the new resource names, and deleted resource names. Then we use the resource name as an identifier to query this resource's metadata.

Since datasets are not frequently updated on CKAN portals, we only perform reaccession once a quarter. CKAN uses the similar metadata template as DCAT does.

GitHub Repository

This repository stores the **harvest script**, **basic portal information**, **resource name list** by portal for each harvest, and the **harvest report** (without manual edits).

- CKANportals.csv stores the basic portal information.
- harvest.ipynb is the harvest script written in Jupyter Notebook. It will request the resource name list for each portal, store in resource folder and compare and get both new and missing resource name. For new resource name, request the resource by name to create metadata into CSV `allNewItems_yyyymmdd.csv` and store in reports folder. Normally we get many new resource names, however, after filter out data types and keep **datasets** and **imagery**, only a few left.

4.4 GEOMG Metadata Toolkit

⚠ This guide is a work in progress (August 2022)

4.4.1 About

What is it?

GEOMG is a custom tool that functions as a backend metadata editor and manager for the GeoBlacklight application.

Who uses it?

BTAA-GIN Operations technical staff at the University of Minnesota

Who developed it?

The BTAA Geoportal Lead Developer, Eric Larson, created GEOMG, with direction from the BTAA-GIN. It is based upon the Kithe framework.

Can other GeoBlacklight projects adopt it?

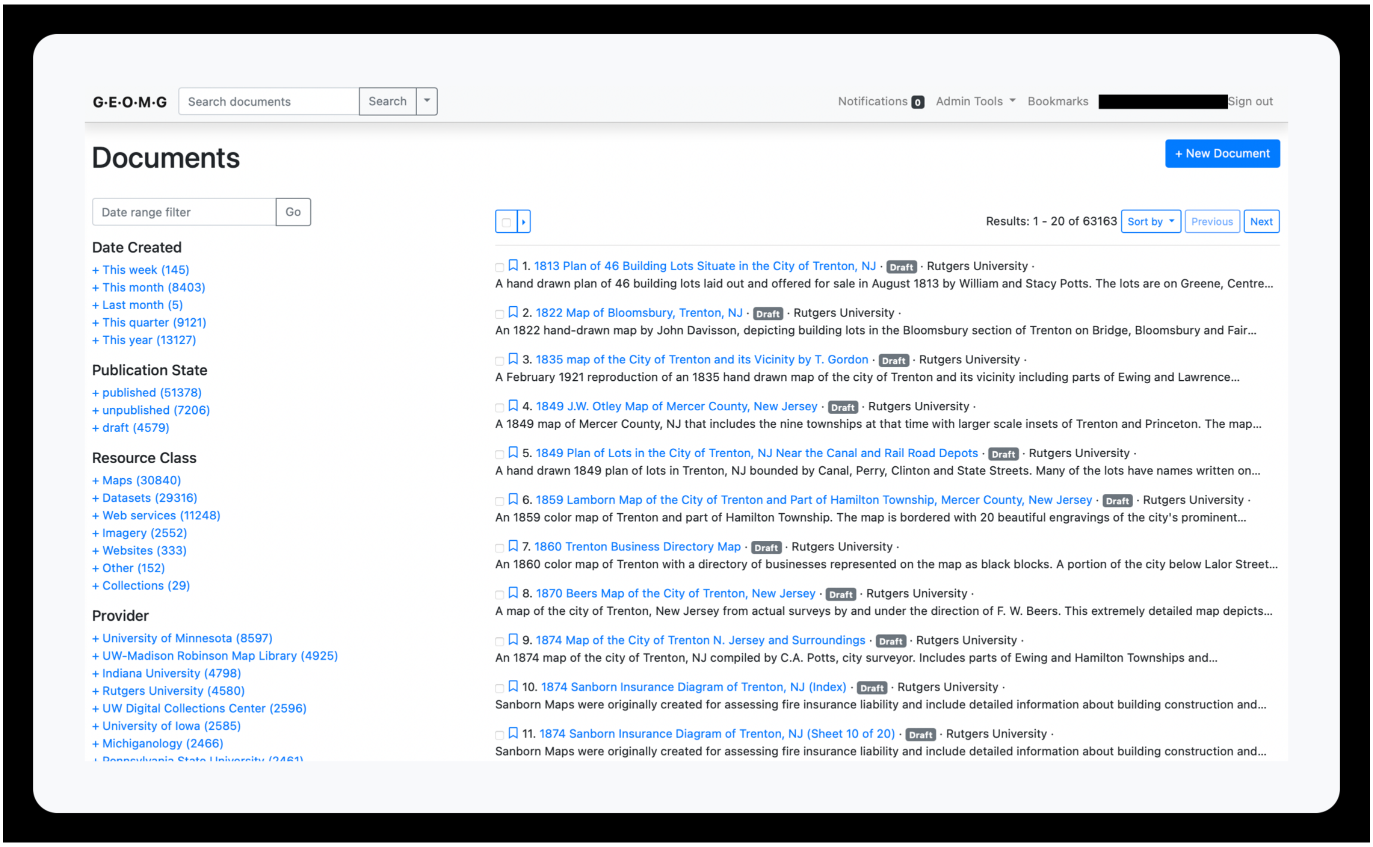
Not yet. We are currently working on offering this tool as a plugin for GeoBlacklight. Our tentative plan for release is early 2023. In the meantime, this presentation describes the motivation for building the tool and a few screencasts showing how it works:

4.4.2 Layout

Dashboard (Home page)

The Dashboard shows a list of all records in the index. These can be selected or filtered. The search functionality mimics the GeoBlacklight interface:

- a search bar at the top: lets a user enter text searches
- a list of facets on the left: lets a user filter records
- a Date Range filter above the facets: lets a user select items by Date Created (when they were first added to GEOMG).



G·E·O·M·G | Search documents | Search | Notifications 0 | Admin Tools | Bookmarks | [redacted] Sign out

Documents

+ New Document

Date range filter | Go

Date Created
- + This week (145)
- + This month (8403)
- + Last month (5)
- + This quarter (9121)
- + This year (13127)

Publication State
- + published (51378)
- + unpublished (7206)
- + draft (4579)

Resource Class
- + Maps (30840)
- + Datasets (29316)
- + Web services (11248)
- + Imagery (2552)
- + Websites (333)
- + Other (152)
- + Collections (29)

Provider
- + University of Minnesota (8597)
- + UW-Madison Robinson Map Library (4925)
- + Indiana University (4798)
- + Rutgers University (4580)
- + UW Digital Collections Center (2596)
- + University of Iowa (2585)
- + Michigiangology (2466)

Results: 1 - 20 of 63163 | Sort by | Previous | Next

□ 1. 1813 Plan of 46 Building Lots Situate in the City of Trenton, NJ · Draft · Rutgers University ·
A hand drawn plan of 46 building lots laid out and offered for sale in August 1813 by William and Stacy Potts. The lots are on Greene, Centre...

□ 2. 1822 Map of Bloomsbury, Trenton, NJ · Draft · Rutgers University ·
An 1822 hand-drawn map by John Davisson, depicting building lots in the Bloomsbury section of Trenton on Bridge, Bloomsbury and Fair...

□ 3. 1835 map of the City of Trenton and its Vicinity by T. Gordon · Draft · Rutgers University ·
A February 1921 reproduction of an 1835 hand drawn map of the city of Trenton and its vicinity including parts of Ewing and Lawrence...

□ 4. 1849 J.W. Otley Map of Mercer County, New Jersey · Draft · Rutgers University ·
A 1849 map of Mercer County, NJ that includes the nine townships at that time with larger scale insets of Trenton and Princeton. The map...

□ 5. 1849 Plan of Lots in the City of Trenton, NJ Near the Canal and Rail Road Depots · Draft · Rutgers University ·
A hand drawn 1849 plan of lots in Trenton, NJ bounded by Canal, Perry, Clinton and State Streets. Many of the lots have names written on...

□ 6. 1859 Lamborn Map of the City of Trenton and Part of Hamilton Township, Mercer County, New Jersey · Draft · Rutgers University ·
An 1859 color map of Trenton and part of Hamilton Township. The map is bordered with 20 beautiful engravings of the city's prominent...

□ 7. 1860 Trenton Business Directory Map · Draft · Rutgers University ·
An 1860 color map of Trenton with a directory of businesses represented on the map as black blocks. A portion of the city below Lalor Street...

□ 8. 1870 Beers Map of the City of Trenton, New Jersey · Draft · Rutgers University ·
A map of the city of Trenton, New Jersey from actual surveys by and under the direction of F. W. Beers. This extremely detailed map depicts...

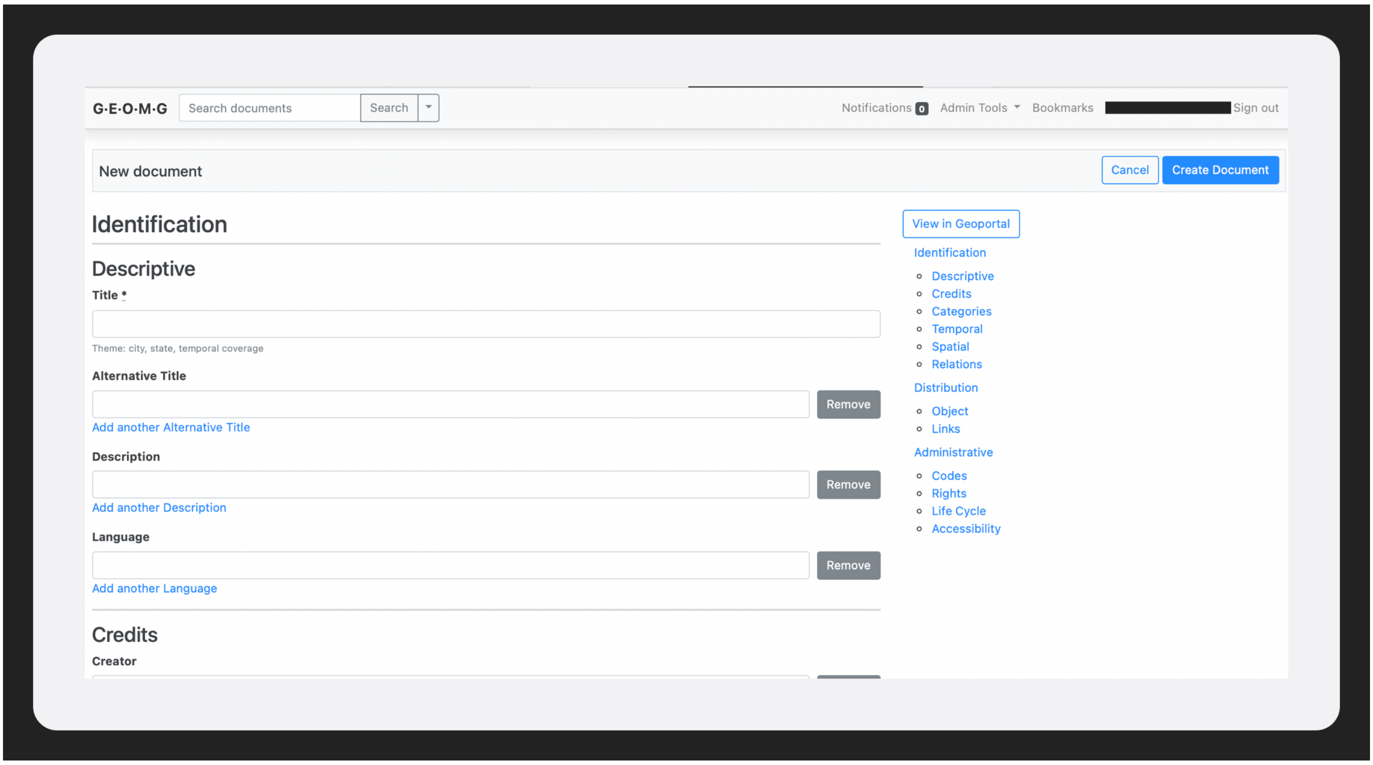
□ 9. 1874 Map of the City of Trenton N. Jersey and Surroundings · Draft · Rutgers University ·
An 1874 map of the city of Trenton, NJ compiled by C.A. Potts, city surveyor. Includes parts of Ewing and Hamilton Townships and...

□ 10. 1874 Sanborn Insurance Diagram of Trenton, NJ (Index) · Draft · Rutgers University ·
Sanborn Maps were originally created for assessing fire insurance liability and include detailed information about building construction and...

□ 11. 1874 Sanborn Insurance Diagram of Trenton, NJ (Sheet 10 of 20) · Draft · Rutgers University ·
Sanborn Maps were originally created for assessing fire insurance liability and include detailed information about building construction and...

Form view

This page is where new records can be manually created and existing records can be edited. Click on the button "View in Geoportal" to open a new tab with the record in the Geoportal. Note: the record is still viewable in the Geoportal via this button, even if it is a Draft or Unpublished.



G·E·O·M·G | Search documents | Search | Notifications 0 | Admin Tools | Bookmarks | [redacted] Sign out

New document | Cancel | Create Document

Identification

Descriptive

Title *

Theme: city, state, temporal coverage

Alternative Title | Remove

Add another Alternative Title

Description | Remove

Add another Description

Language | Remove

Add another Language

Credits

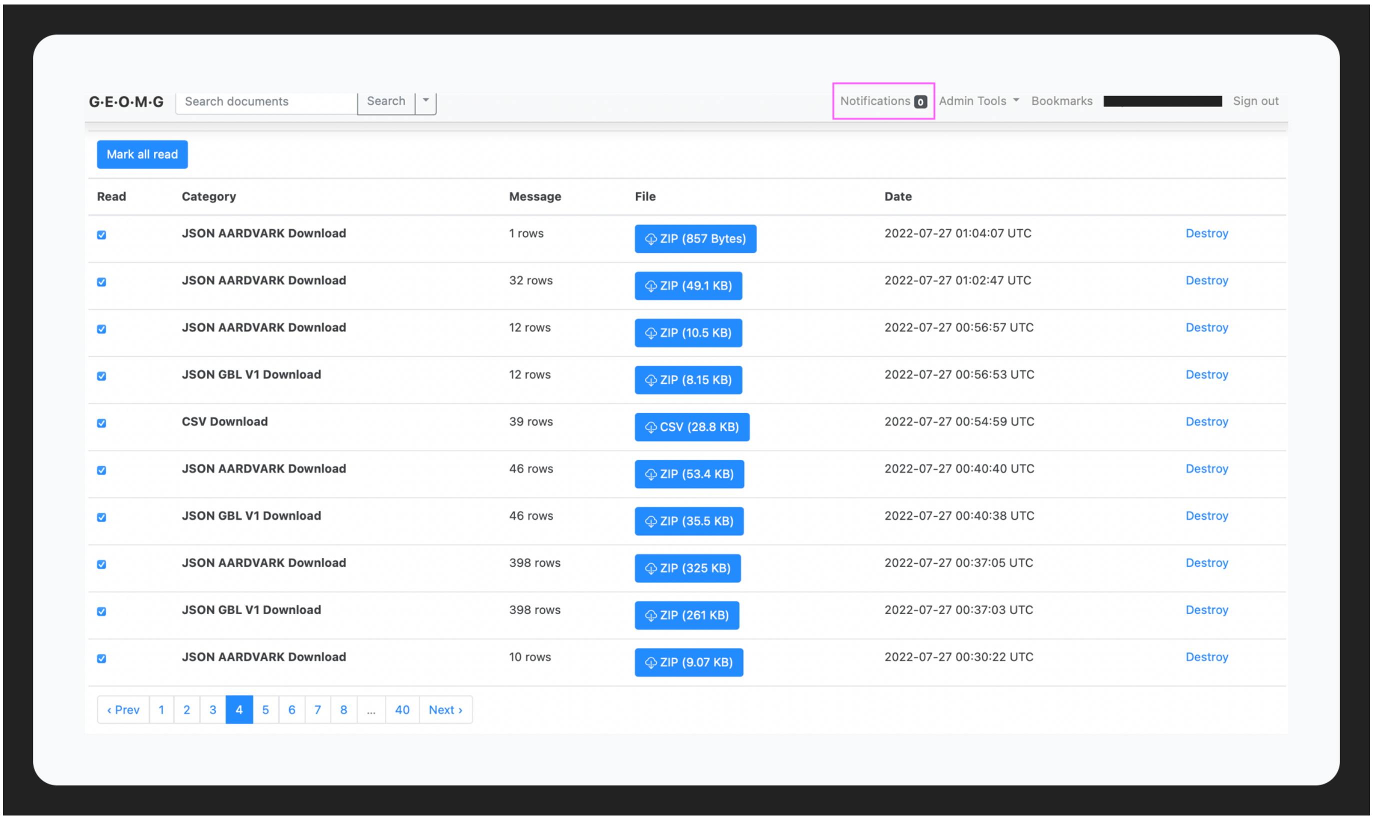
Creator

View in Geoportal

- Identification
  - Descriptive
  - Credits
  - Categories
  - Temporal
  - Spatial
  - Relations
- Distribution
  - Object
  - Links
- Administrative
  - Codes
  - Rights
  - Life Cycle
  - Accessibility

Notifications

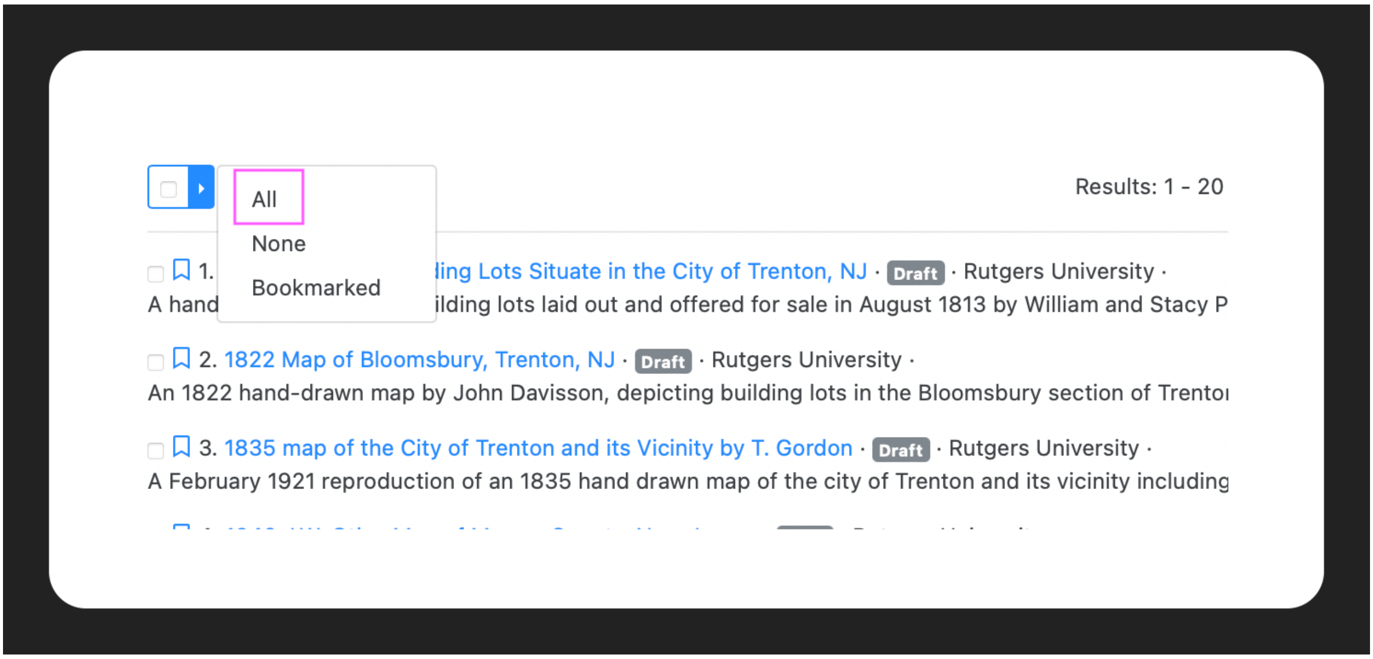
Notifications is where the exported files can be found.



Read	Category	Message	File	Date	
☑	JSON AARDVARK Download	1 rows	ZIP (857 Bytes)	2022-07-27 01:04:07 UTC	Destroy
☑	JSON AARDVARK Download	32 rows	ZIP (49.1 KB)	2022-07-27 01:02:47 UTC	Destroy
☑	JSON AARDVARK Download	12 rows	ZIP (10.5 KB)	2022-07-27 00:56:57 UTC	Destroy
☑	JSON GBL V1 Download	12 rows	ZIP (8.15 KB)	2022-07-27 00:56:53 UTC	Destroy
☑	CSV Download	39 rows	CSV (28.8 KB)	2022-07-27 00:54:59 UTC	Destroy
☑	JSON AARDVARK Download	46 rows	ZIP (53.4 KB)	2022-07-27 00:40:40 UTC	Destroy
☑	JSON GBL V1 Download	46 rows	ZIP (35.5 KB)	2022-07-27 00:40:38 UTC	Destroy
☑	JSON AARDVARK Download	398 rows	ZIP (325 KB)	2022-07-27 00:37:05 UTC	Destroy
☑	JSON GBL V1 Download	398 rows	ZIP (261 KB)	2022-07-27 00:37:03 UTC	Destroy
☑	JSON AARDVARK Download	10 rows	ZIP (9.07 KB)	2022-07-27 00:30:22 UTC	Destroy

4.4.3 Selecting records

Use the check boxes to select individual records or click the right arrow to select all records on the page.

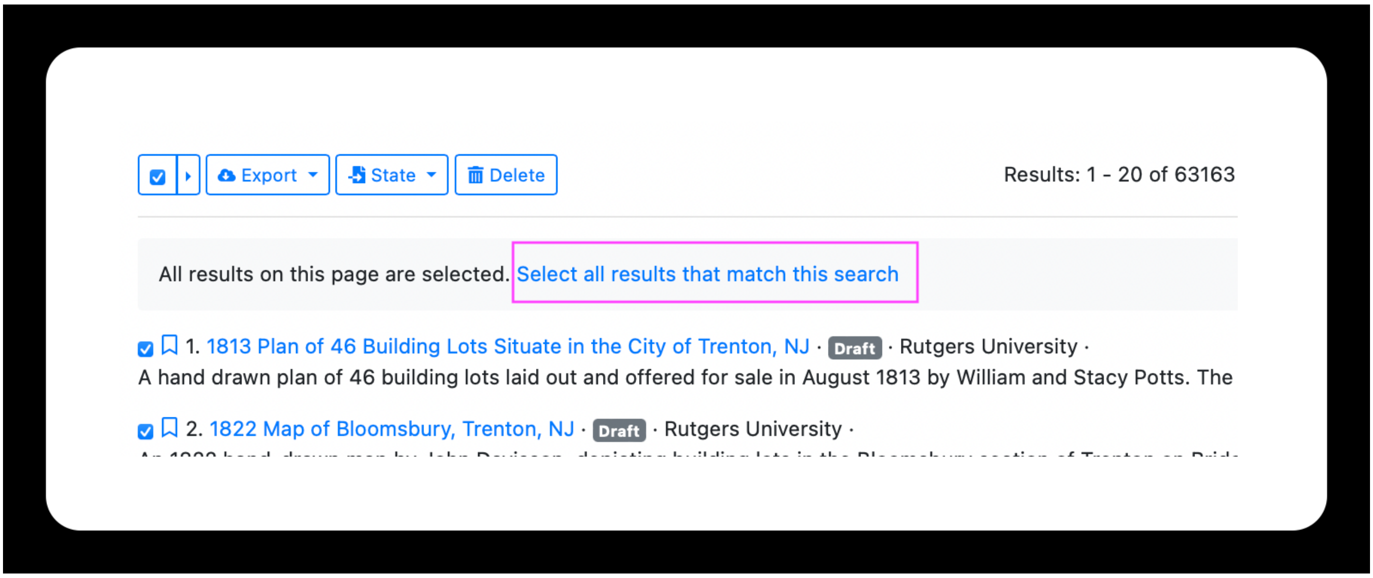


□ 1. …ing Lots Situate in the City of Trenton, NJ · Draft · Rutgers University ·
A hand… …ilding lots laid out and offered for sale in August 1813 by William and Stacy P

□ 2. 1822 Map of Bloomsbury, Trenton, NJ · Draft · Rutgers University ·
An 1822 hand-drawn map by John Davisson, depicting building lots in the Bloomsbury section of Trento

□ 3. 1835 map of the City of Trenton and its Vicinity by T. Gordon · Draft · Rutgers University ·
A February 1921 reproduction of an 1835 hand drawn map of the city of Trenton and its vicinity including

To select all records in the list, first select all records on the page. Then, click the text, "Select all results that match this search."



Results: 1 - 20 of 63163

All results on this page are selected. Select all results that match this search

☑ 1. 1813 Plan of 46 Building Lots Situate in the City of Trenton, NJ · Draft · Rutgers University ·
A hand drawn plan of 46 building lots laid out and offered for sale in August 1813 by William and Stacy Potts. The

☑ 2. 1822 Map of Bloomsbury, Trenton, NJ · Draft · Rutgers University ·

4.4.4 Adding new records

There are two ways to add new records to GEOMG: with the Form view or with a CSV upload.

Form view

A user can create records one by one using the form view.

1. Start on the main dashboard
2. click the button labeled **+New Document**
3. Manually type in values. Some fields need to be entered before the record can be saved including:

- Title
- Resource Class
- ID
- Access Rights

Batch Uploading with a CSV

1. Save edited template to your desktop as a CSV file Use the B1G Template

2. Upload the spreadsheet to GEOMG

- Go to Admin Tools and select Imports. Click **New Import**.
- Give a name to the upload and enter details about the source and description. These details are helpful later in tracking imports.
- Select the CSV file for upload
- For Type, choose **BTAA CSV**.
- Click the **Create Import** button
- Review the *Field Mappings* page. If the CSV was formatted with column headers spelled the same way as the template, the fields should automatically map to the correct elements. Otherwise, manually choose the crosswalk mapping.
- Scroll to the bottom and click the button **Create Mapping**
- On the *Import* page, check that the number in the CSV Row Count matches your CSV.
- Click the button **Run Import**
- The import may take a few minutes. During the process, you can view the Import Results tab. Items in the queue will show up in the first sub-tab ("Failed"), but will transfer to the second tab upon import ("Success").
- When complete, review any items that did not import in the Failed tab. See Troubleshooting (*coming soon*) for help.

3. Spot check records for errors and consistency

- The newly uploaded records will be listed as Draft under the Publication State on the main dashboard
- Select 'Draft' under Publication State and select an item. This will open it in editing view.
- Click the button **View in Geoportal**
- Inspect the record and test the links. (note: Metadata and Web Service links will not open while the item is still in Draft)
- Repeat this process for about 3 records.

4. Convert records from 'Draft' to 'Published'

- If the records are satisfactory, return to the Dashboard view and select all Draft items in the upload.
- Select All and then select the text "Select all results that match this search"
- Click the State button. From the dropdown, select Published.
- On the *Bulk Action* page, click the button **Run Bulk Action**
- Review 3-5 of the published records and test all the links.

4.4.5 Secondary tables

There are two metadata fields, `Multiple Download Links` and `Institutional Access Links` that use secondary tables. This occurs when the field needs parts to the value: a label + a link.

- When using the Form view, these values can be entered directly.
- For CSV uploads, these values use a separate sheet than for the main import template.
- Multiple Download Links:

- on the Form view, scroll down to the Multiple Download Links inside the editor
- to enter manually, click the New Download URL button
- to upload multiple links, click the Import CSV button

- Institutional Access Links

- on the Form view, click the text "Institutional Access Links" on the bottom of the right hand navigation OR go to Admin Tools - Access Links
- to enter manually, click the New Access URL button
- to upload multiple links, click the Import CSV button

CSV field headers for secondary tables

Multiple Downloads   Institutional Access Links

```
| friendlier_id        | label            | value      |
|---------------------|------------------|------------|
| ID of target record |  any string      | the link   |
```

```
| friendlier_id        | institution_code | access_URL |
|---------------------|------------------|------------|
| ID of target record |  2 digit code    | the link   |
```